

Encyclopedia of Earth Sciences Series

Encyclopedia of Mathematical Geosciences

Edited by

B. S. Daya Sagar · Qiuming Cheng
Jennifer McKinley · Frits Agterberg

 Springer

Encyclopedia of Earth Sciences Series

Series Editors

Charles W. Finkl, Department of Geosciences, Florida Atlantic University, Boca Raton, FL,
USA

Rhodes W. Fairbridge, New York, NY, USA

The *Encyclopedia of Earth Sciences Series* provides comprehensive and authoritative coverage of all the main areas in the Earth Sciences. Each volume comprises a focused and carefully chosen collection of contributions from leading names in the subject, with copious illustrations and reference lists. These books represent one of the world's leading resources for the Earth Sciences community. Previous volumes are being updated and new works published so that the volumes will continue to be essential reading for all professional earth scientists, geologists, geophysicists, climatologists, and oceanographers as well as for teachers and students. Most volumes are also available online. **Accepted** for inclusion in Scopus.

B. S. Daya Sagar • Qiuming Cheng •
Jennifer McKinley • Frits Agterberg
Editors

Encyclopedia of Mathematical Geosciences

With 555 Figures and 97 Tables

 Springer

Editors

B. S. Daya Sagar
Systems Science & Informatics Unit
Indian Statistical Institute – Bangalore Centre
Bangalore, India

Qiuming Cheng
Institute of Earth Sciences
China University of Geosciences
Beijing, China

Jennifer McKinley
Geography
School of Natural and Built Environment
Queen's University Belfast
Belfast, UK

Frits Agterberg
Canada Geological Survey
Ottawa, ON, Canada

ISSN 1388-4360 ISSN 1871-756X (electronic)
Encyclopedia of Earth Sciences Series
ISBN 978-3-030-85039-5 ISBN 978-3-030-85040-1 (eBook)
<https://doi.org/10.1007/978-3-030-85040-1>

© Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

We dedicate the Encyclopedia of Mathematical Geosciences to two scientists: John Cedric Griffiths and William Christian Krumbein, who inspired the four of us and were directly involved in the creation of the field of mathematical geoscience. These two scientists symbolize our intention for the Encyclopedia of Mathematical Geosciences to connect the basic and applied divisions of the mathematical geosciences.



William Christian Krumbein (1902–1979) was a founding officer of the Association. Born at Beaver Falls, Pennsylvania, in January 1902, Krumbein attended the University of Chicago, receiving the degree of bachelor of philosophy in business administration in 1926, an MS in geology in 1930, and a PhD in geology in 1932. He taught at the University of Chicago from 1933 to 1942, advancing from instructor to associate professor. During World War II, from 1942 until 1945, he served in Washington, D.C. with the Beach Erosion Board of the U.S. Army Corps of Engineers. Following a short stint with Gulf Research and Development Company, immediately after the end of the war, he joined Northwestern University in 1946, serving there until mandatory retirement in 1970. He was named the William Deering Professor of Geological Sciences in 1960. Krumbein died on August 18, 1979, a few months after Syracuse University had awarded him a DSc (honoris causa). At his memorial service, former Northwestern colleague Larry Sloss said of Krumbein “that by constitutionally rejecting conventional wisdom, he continually pursued innovative methods whereby the natural phenomena of geology could be

expressed with mathematical rigor.” IAMG instituted the William Christian Krumbein Medal. This award was fittingly named for William C. Krumbein. He is considered one of the fathers of the subject.



John Cedric Griffiths (1912–1992), formerly Professor of Petrography at the Pennsylvania State University, was the first recipient of the IAMG’s Krumbein Medal in 1976. Like Krumbein, Griffiths remains well known as a pioneer of introducing quantitative methods across a wide spectrum of geological and economic problems. Born in Llanelli, Carmarthenshire, Wales, he graduated in petrology from the Universities of Wales and London prior to working for a petroleum company in Trinidad for 7 years. In June 1987, he retired after 40 years of teaching in the Department of Mineralogy and Petrology at the Pennsylvania State faculty. He directed more than 50 MSc and PhD students. From his classes, students learned the communality of scientific problems across different disciplines including industrial engineering, computer science, and psychology. During 1967–1968, he was the first Distinguished Visiting Lecturer at Kansas University in a program instigated by Professor Daniel Merriam. In addition to his well-known 1967 textbook Scientific Method in Analysis of Sediments, the scientific literature has been greatly enriched by his many articles. His work on unit regional values received international acclaim. Professor Griffith’s scientific distinction, coupled with his wit and lively often provocative oral presentations, have stimulated everyone lucky enough to have experienced them.

Foreword

It is a great honor for me to be invited to contribute this foreword to what is in so many ways an outstanding and impressive work. Dr. Sagar and his team are to be congratulated on bringing together such a complete and comprehensive reference work that will be invaluable to students, researchers, and educators for years to come.

Many people who are not already aware of the progress that has been made in the mathematical geosciences in the past few decades may wonder about the existence of such a field at the intersection between mathematics and the geosciences. The natural world often appears random and unpredictable, and thus very unlike the pure logic and beautiful symmetries of mathematics. To be sure, there are examples of something approaching purity and simplicity in the natural world: one thinks of the sine-generated curves of river meanders, the almost-perfect circles of craters, the elegant geometry of some periglacial features such as frost-wedge polygons, or the statistics of stream networks. It was features such as these that attracted my interest many years ago when I found myself enrolled as a PhD student working under a karst geomorphologist (Derek Ford) after a conventional background in undergraduate physics, where the idea that the world can be reduced to a few simple principles has dominated science for centuries.

At the opposite extreme, the geosciences have provided a rich set of applications for the mathematics of disorder and chaos. Many of Mandelbrot's initial examples of his developing mathematics of fractals were provided by the geosciences, including the geometry of shorelines. Moreover, some simple discoveries in the geosciences, such as the scaling laws and the laws of stream number, can be attributed to nothing more than the asymptotic result of randomness.

In a way, history has brought us full circle in the increasing use of artificial intelligence and machine learning in the geosciences. Thoroughly grounded in data science and in the fourth-paradigm mantra "let the data speak for themselves," these methods abandon any semblance of theory or generalizability in the search for algorithms that support searches and provide predictions based in complexity. While Occam's Razor and the traditional scientific search for simplicity may have largely failed to provide understanding and explanation in the geosciences, these methods are at least capable of revealing similarities and supporting an early, hypothesis-generating phase of a research project.

Before we go too far into data science, however, it is worth remembering the famous saying of Korzybski (1933): "The map is not the territory," that is, what the map or the data are trying to capture is no more than a representation of reality, and will be at best only an approximation, generalization, abstraction, or interpretation. The interpretations drawn from data, whether through statistical analysis or machine learning, will always be subject to the uncertainties that must be present in the data, along with other uncertainties that result from the use of imperfect models, and there will always be a danger that conclusions tell us more about those uncertainties than about reality.

In recent years, many branches of science have been the subject of inquiries into ethics. This may have begun with concerns about reproducibility and replicability in experimental psychology (NASEM 2019), and it grew further during the civil unrest in the USA in 2020 with the call to improve diversity, equity, and inclusion. Ethical issues arise in many areas of the geosciences, and more broadly in the social and environmental sciences. Individual privacy and surveillance often feature strongly in debates over ethics, but many other issues can be regarded as problems

of ethics (AAG 2022). They include making inappropriate or biased inferences from data, ignoring the presence of uncertainty in findings, or unwarranted generalization across space or in time. While none of the entries in the encyclopedia address ethical issues exclusively, and there are none on replicability or diversity, many of them provide a context for further discussion of ethics in specific areas, and perhaps future editions of the encyclopedia might include more on social issues such as these.

Anyone reviewing the encyclopedia and its many entries will wonder what it is that is different about the mathematics that is useful in the geosciences, and similarly about the statistical and computational methods, and the methods of data science. They might wonder what new developments in each of these areas have originated in the mathematical geosciences and perhaps spread into other fields. There are several examples of this in the encyclopedia: the article on Geographically Weighted Regression, which is entirely driven by the need for a relaxed form of replicability in order to deal with the essential spatial heterogeneity of the Earth's surface; and the entire field of geostatistics, which arose because of the necessity of dealing with geospatial data that exhibit spatial dependence, or regionalized variables in the terminology of Matheron. In inferential statistics, it is clear that the simple model, developed by Fisher and others in order to make conclusions about populations from samples drawn randomly and independently from that population, requires special methods to deal with samples in the geosciences that may be far from independent, and often represent an entire population rather than a sample from it.

This is an exciting time for the mathematical geosciences: new data streams are coming online, at finer and finer temporal and spatial resolution; new methods are emerging in machine learning and advanced analytics; computational power is no longer as expensive and constrained as it has been; the commercial sector is expanding; and new graduate students and faculty are bringing new ideas and enthusiasm. I hope this encyclopedia proves to be as useful and defining as it clearly has the potential to be, and I congratulate Dr. Sagar and his team once more on a magnificent achievement.

University of California
Santa Barbara

Michael F. Goodchild

References

- American Association of Geographers (2022) A white paper on locational information and the public interest. <https://www.aag.org/wp-content/uploads/1900/09/2022-White-Paper-on-Locational-Information-and-the-Public-Interest.pdf>. Accessed 16 Nov 2022
- Korzybski A (1933) Science and sanity. An introduction to non-Aristotelian systems and general semantics. The International Non-Aristotelian Library Pub. Co., pp 747–761
- National Academies of Sciences, Engineering, and Medicine (2019) Reproducibility and replicability in science. National Academies Press, Washington, DC

Preface

The past six decades have proved monumental in the discovery and accumulation of facts and detailed data on the solid earth, oceans, atmosphere, and space science. Now an established field, Mathematical Geosciences, covers the original approaches, while bringing new applications of well-established mathematical and statistical techniques to address the various challenges encountered. Pioneering academics and scientists who developed the inceptive fabric of mathematical geosciences, in chronological order, to name a few, including Andrey Kolmogorov (Probability Theory), Ronald Fisher (Mathematical Statistics), Danie Krige (Ore Reserve Valuation), William Christian Krumbein (Sedimentary Geology), John C Griffiths (Econometrics and Sedimentary Petrology), John Tukey (Exploratory Data Analysis), Georges Matheron (Geostatistics), Benoit Mandelbrot (Fractal Geometry), Frits Agterberg (Geomathematics), Geoffrey Watson (Directional Statistics), Daniel Merriam (Computational Geosciences), Jean Serra (Mathematical Morphology), Dietrich Stoyan (Stochastic Geometry), John Aitchison (Compositional Data Analysis), and William Newman (Complexity Science), are among many others who emerged as mathematical geoscientists during the last three decades. Some of the pioneering associations, besides the International Association for Mathematical Geosciences (IAMG), that promote mathematical geosciences include Society for Industrial and Applied Mathematics (SIAM), IEEE's Geoscience and Remote Sensing Society (GRSS), American Geophysical Union (AGU), European Geoscience Union (EGU), and the Indian Geophysical Union (IGU). With this backdrop, we have worked toward releasing the *Encyclopedia of Mathematical Geosciences* – as a sequel to *Dictionary of Mathematical Geosciences* and *Handbook of Mathematical Geosciences* published by Springer during the latter part of the last decade – receiving requisite attention from mathematicians and geoscientists alike, thus facilitating dialogue among the stakeholders to address outstanding issues in “Mathematics in Geosciences” in the years to come.

We bring scientists, engineers, and mathematicians together contributing chapters related to the geosciences, and we opine that this encyclopedia is essential for the next generation of geoscience researchers and educators. It provides more than 300 entries in a lucid manner concerned with relevant conventional and modern mathematical, statistical, and computational techniques for application in the geosciences and space science. Such a resource should be highly useful for the next generation of Mathematical Geoscientists. Chapters in this encyclopedia are of three categories. On much broader topics of relevance to the Mathematical Geosciences, experienced pioneers have contributed about 30 Category-A type chapters and about 265 Category-B type chapters on entries of relevance to the mathematical geosciences, while Category-C type chapters which are brief biographies of fifty pioneering mathematical geoscientists. Most existing papers and books published in the Mathematical Geosciences are introduced for mathematically inclined readers prepared to go through the texts from beginning to end. The main goal of this *Encyclopedia of Mathematical Geosciences* is to make important keyword-specific chapters accessible to readers interested in specific topics but lacking the time to peruse the more comprehensive publications in their areas of interest. This encyclopedia humbly provides comprehensive references to over 300 topics in the field of Mathematical Geosciences. Needless to say, keywords that are mathematically important but without established geoscientific relevance and keywords that are relevant to geosciences without direct

connection to mathematics were omitted. Each chapter was prepared in such a way that a layperson with high-school qualifications should be able to understand its content without too much difficulty. As much as possible, each chapter was prepared in a descriptive style with minimal mathematical jargon but directing the mathematically oriented readers to key references for further study. To our best, we have avoided redundancy in this two-volume encyclopedia, and we suggest that the reader refers to other encyclopedias in the Earth Science Series whenever needed.

A host of challenges of geoscientific relevance commands the attention of applied mathematicians. Addressing some of those challenges also offers opportunities as well as the need to advance the foundations and techniques of applied mathematics in order to meet the challenges. Further scope entails several other entries – from dynamical systems, ergodic theory, renormalization, and multiscaling – which possess high potential to address challenges encountered in the geosciences. Some challenges relevant to fluid and solid media – such as the Earth’s interior and exterior, analysis, simulation, and prediction of terrestrial processes, ranging from climate change to earthquakes, space sciences, etc. – command the involvement of dynamical systems and ergodic theory, renormalization, and topological data analysis.

While we like to receive feedback on what has been given emphasis within the gambit of “Mathematical Geosciences,” we also would like to explore what more needs to be done to broaden the portfolio of our scientific discipline. The encyclopedia will require revisions and updates at least once per decade. Its four editors reflect diverse disciplines and they have contributed significantly, from the EOMG’s inception stage to its release. The encyclopedia is available not only as a two-volume printed version but also as an online reference work at <http://springerlink.com> which will allow us to make regular updates as new entries and techniques emerge.

B. S. Daya Sagar
Qiuming Cheng
Jennifer McKinley
Frits Agterberg

Acknowledgments

Half-a-Decade-long work in bringing the *Encyclopedia of Mathematical Geosciences* to its concluding stage was a herculean task. During the last 5 years (2018–2023), countless hours were spent on the listing of keyword-specific entries, inviting potential authors to contribute keyword-specific chapters, reviewing the submitted chapters, finalizing and forwarding them for subsequent work at the production department, and organizing the corrected galley proofs, followed by the preparation of preliminary content, and front and back matter. There was huge coordination involved throughout among authors, reviewers, sectional editors, editors, the staff of the Springer Nature Publishers, and the production department. We gratefully acknowledge the help and support of the authors ranging from multiple decades of experience to the new generation who have graciously contributed excellent chapters, investing their time, knowledge, and expertise. The basic idea of producing this encyclopedia came from Daya Sagar who subsequently by far has made the largest contribution to making this project a success. The other editors are grateful to him for his initiatives and guidance. This kind of project is successful mostly because of the support, cooperation, and guidance received from highly experienced senior colleagues; viz., Frits Agterberg, Ricardo Olea, Gabor Korvin, Eric Grunsky, John Schuenemeyer, and younger colleagues Jaya Sreevalsan-Nair and Lim Sin Liang, among others. We (Sagar, Cheng, and McKinley) owe a lot to Frits for his guidance and sharp memory, and we feel fortunate to have him and his 24×7 working style that has helped us past several hiccups during the past 5 years. Frits: We apologize for requiring your time very often even while you were hospitalized and ill for three months during the first quarter of 2021. The timely permissions granted/arranged for a few copyrighted media of information by Teja Rao, Laurent Mandelbrot, Karen Kafadar, Ramana Sonty, and Britt-Louise Kinnefors are gratefully acknowledged.

The considerable support received from Annett Buettner, Sylvia Blago, and Johanna Klute from Springer Nature Publishers is simply remarkable, and the efficiency, patience, relentless reminders, and strong-willed determination that they have shown throughout every phase of this monumental project are highly professional. The lion's share of the success should be theirs, besides the contributions of the authors.

B. S. Daya Sagar
Qiuming Cheng
Jennifer McKinley
Frits Agterberg

Contents

Accuracy and Precision	1
P. A. Dowd	
Additive Logistic Normal Distribution	4
Gianna Serafina Monti, Glòria Mateu-Figueras and Karel Hron	
Additive Logistic Skew-Normal Distribution	10
Glòria Mateu-Figueras and Karel Hron	
Agterberg, Frits	15
Qiuming Cheng	
Aitchison, John	17
John Bacon-Shone	
Algebraic Reconstruction Techniques	17
Utkarsh Gupta and Uma Ranjan	
Allometric Power Laws	24
Gabor Korvin	
Argand Diagram	28
James Irving and Eric P. Verrecchia	
Artificial Intelligence in the Earth Sciences	31
Norman MacLeod	
Artificial Neural Network	43
Yuanyuan Tian, Mi Shu and Qingren Jia	
Autocorrelation	47
Alejandro C. Frery	
Automatic and Programmed Gain Control	49
Gabor Korvin	
Backprojection Tomography	55
Utkarsh Gupta and Uma Ranjan	
Bayes's Theorem	61
Konstantinos Modis	
Bayesian Inversion in Geoscience	65
Dario Grana, Klaus Mosegaard and Henning Omre	
Bayesian Maximum Entropy	71
Junyu He and George Christakos	
Best Linear Unbiased Estimation	79
Peter K. Kitanidis	

Big Data	85
Xiaogang Ma	
Binary Mathematical Morphology	92
Aditya Challa, Sravan Danda and B. S. Daya Sagar	
Binary Partition Tree	94
Sravan Danda, Aditya Challa and B. S. Daya Sagar	
Bonham-Carter, Graeme	95
Eric Grunsky	
Bootstrap	96
Oktay Erten and Clayton V. Deutsch	
Burrough, Peter Alan	100
Andrew U. Frank	
Cartogram	103
Michael T. Gastner	
Chaos in Geosciences	107
Frits Agterberg and Qiuming Cheng	
Chayes, Felix	113
Richard J. Howarth	
Cheng, Qiuming	114
Frits Agterberg	
Circular Error Probability	115
Frits Agterberg	
Cloud Computing and Cloud Service	123
Liping Di and Ziheng Sun	
Cluster Analysis and Classification	127
Antonella Bucciante and Caterina Gozzi	
Compositional Data	133
Vera Pawlowsky-Glahn and Juan José Egozcue	
Computational Geoscience	143
Eric Grunsky	
Computer-Aided or Computer-Assisted Instruction	165
Madhurima Panja and Uttam Kumar	
Concentration-Area Plot	169
Behnam Sadeghi	
Constrained Optimization	175
Deeksha Aggarwal and Uttam Kumar	
Convex Analysis	180
Indu Solomon and Uttam Kumar	
Copula in Earth Sciences	184
Sandra De Iaco and Donato Posa	
Correlation and Scaling	193
Frits Agterberg	

Correlation Coefficient	201
Guocheng Pan	
Coupled Modeling	208
Mohammad Ali Aghighi and Hamid Roshan	
Cracknell, Arthur Phillip	213
Kasturi Devi Kanniah	
Cressie, Noel A.C.	214
Jay M. Ver Hoef	
Crystallographic Preferred Orientation	215
Helmut Schaeben	
Cumulative Probability Plot	222
Pravan Danda, Aditya Challa and B. S. Daya Sagar	
Dangermond, Jack	225
Lowell Kent Smith	
Data Acquisition	226
Rajendra Mohan Panda and B. S. Daya Sagar	
Data Dictionary	230
Madhurima Panja, Tanujit Chakraborty and Uttam Kumar	
Data Life Cycle	235
Fang Huang	
Data Mining	238
Tao Wen	
Data Visualization	241
Jaya Sreevalsan-Nair	
Database Management System	247
Rajendra Mohan Panda and B. S. Daya Sagar	
David, Michel	250
Denis Marcotte	
Davis, John Clements	251
Ricardo A. Olea	
De Wijs Model	252
Shuyun Xie	
Decision Tree	257
Rajendra Mohan Panda and B. S. Daya Sagar	
Deep CNN-Based AI for Geosciences	262
Mehala Balamurali	
Deep Learning in Geoscience	267
Fuchang Gao	
Dendrogram	271
Rogério G. Negri	
Differential Equations	274
Sergio Zlotnik and Jaume Soler Villanueva	

Digital Elevation Model	277
Pradeep Srivastava and Sanjay Singh	
Digital Filters	290
Gabor Korvin	
Digital Geological Mapping	293
Chengbin Wang and Xiaogang Ma	
Digital Twins of the Earth	295
Stefano Nativi and Max Craglia	
Dimensionless Measures	299
Ekta Baranwal	
Discrete Prolate Spheroidal Sequence	303
Dionissios T. Hristopulos	
Discriminant Analysis	307
D. Arun Kumar	
Discriminant Function Analysis	311
Norman MacLeod	
Doveton, John Holroyd	317
John C. Davis	
Earth Surface Processes	319
Vikrant Jain, Ramendra Sahoo and R. N. Singh	
Earth System Science	325
Gang Liu and Hongfei Zhang	
Earthquake Complexity	328
William I. Newman	
Eigenvalues and Eigenvectors	336
Jaya Sreevalsan-Nair	
Electrofacies	339
John H. Doveton	
Ensemble Kalman Filtering	342
J. Jaime Gómez-Hernández	
Entropy	346
Julian M. Ortiz and Jorge F. Silva	
Euler Poles of Tectonic Plates	350
Helmut Schaeben, Uwe Kroner and Tobias Stephan	
Expectation-Maximization Algorithm	355
Jaya Sreevalsan-Nair	
Exploration Geochemistry	358
David R. Cohen	
Exploratory Data Analysis	364
Emmanuel John M. Carranza	
FAIR Data Principles	369
Abdullah Alowairdhi and Xiaogang Ma	

Fast Fourier Transform	372
Frits Agterberg	
Fast Wavelet Transform	376
Indu Solomon and Uttam Kumar	
Favorability Analysis	380
Guocheng Pan	
Fisher, Ronald A.	387
Frits Agterberg	
Flow in Porous Media	388
Daniele Pedretti	
Forward and Inverse Stratigraphic Models	393
Cedric M. Griffiths	
Fractal Geometry and the Downscaling of Sun-Induced Chlorophyll Fluorescence Imagery	404
Juan Quiros-Vargas, Bastian Siegmann, Alexander Damm, Ran Wang, John Gamon, Vera Krieger, B. S. Daya Sagar, Onno Muller and Uwe Rascher	
Fractal Geometry in Geosciences	407
Qiuming Cheng and Frits Agterberg	
Fractal Landscape	430
Ramendra Sahoo	
Frequency Distribution	435
Ricardo A. Olea	
Frequency-Wavenumber Analysis	437
Frits Agterberg	
Full Normal Plot	445
Gianna Serafina Monti and Glòria Mateu-Figueras	
Fuzzy C-Means Clustering	447
Jaya Sreevalsan-Nair	
Fuzzy Inference Systems for Mineral Exploration	449
Bijal Chudasama, Sanchari Thakur and Alok Porwal	
Fuzzy Set Theory in Geosciences	454
Behnam Sadeghi, Alok Porwal, Amin Beiranvand Pour and Majid Rahimzadegan	
Genetic Algorithms	465
D. Arun Kumar	
Geocomputing	468
Alice-Agnes Gabriel, Marcus Mohr and Bernhard S. A. Schuberth	
Geographical Information Science	473
Parvatham Venkatachalam	
Geographically Weighted Regression	485
Xiang Que and Shaoqiang Su	
Geohydrology	489
Faramarz Doulati Ardejani, Behshad Jodeiri Shokri, Soroush Maghsoudy, Majid Shahhosseiny, Fojan Shafaei, Farzin Amirkhani Shiraz and Andisheh Alimoradi	

Geoinformatics	494
Jeffrey M. Yarus, Jordan M. Yarus and Roger H. French	
Geologic Time Scale Estimation	502
Felix M. Gradstein and Frits Agterberg	
Geomathematics	512
Frits Agterberg	
Geomatics	519
Rifaat Abdalla	
Geomechanics	523
Pejman Tahmasebi	
Geoscience Signal Extraction	526
Frits Agterberg	
Geosciences Digital Ecosystems	534
Stefano Nativi and Paolo Mazzetti	
Geostatistical Seismic Inversion	539
Leonardo Azevedo and Amílcar Soares	
Geostatistics	543
H. S. Pandalai and A. Subramanyam	
GeoSyntax	568
Cedric M. Griffiths	
Geotechnics	573
Pejman Tahmasebi	
Geothermal Energy	575
Katsuaki Koike and Shohei Albert Tomita	
Global and Regional Climatic Modeling	582
Yuriy Kostyuchenko, Igor Artemenko, Mohamed Abioui and Mohammed Benssaou	
Goodchild, Michael F.	586
B. S. Daya Sagar	
Gradstein, Felix	587
Frits Agterberg	
Grain Size Analysis	588
Michael Klichowicz and Holger Lieberwirth	
Graph Mathematical Morphology	594
Rahisha Thottolil and Uttam Kumar	
Grayscale Mathematical Morphology	600
Shravan Danda, Aditya Challa and B. S. Daya Sagar	
Griffiths, John Cedric	601
Donald A. Singer	
Harbaugh, John W.	603
Johannes Wendebourg	
Harff, Jan	604
Martin Meschede	

High-Order Spatial Stochastic Models	605
Roussos Dimitrakopoulos and Lingqing Yao	
Hilbert Space	613
Daisy Arroyo	
Horton, Robert Elmer	618
Keith Beven	
Howarth, Richard J.	619
J. M. McArthur	
Hurst Exponent	620
Sid-Ali Ouadfeul and Leila Aliouane	
Hyperspectral Remote Sensing	625
Daniele Marinelli, Francesca Bovolo and Lorenzo Bruzzone	
Hypothesis Testing	630
Rogério G. Negri	
Hypsometry	633
Sebastiano Trevisani and Lorenzo Marchi	
Imputation	637
Javier Palarea-Albaladejo	
Independent Component Analysis	642
Jaya Sreevalsan-Nair	
Induction, Deduction, and Abduction	644
Frits Agterberg	
International Generic Sample Number	656
Sarah Ramdeen, Kerstin Lehnert, Jens Klump and Lesley Wyborn	
Interpolation	660
Jaya Sreevalsan-Nair	
Interquartile Range	664
Alejandro C. Frery	
Inverse Distance Weight	666
Narayan Panigrahi	
Inversion Theory	672
R. N. Singh and Ajay Manglik	
Inversion Theory in Geoscience	678
Shib Sankar Ganguli and V. P. Dimri	
Iterative Weighted Least Squares	688
Sara Taskinen and Klaus Nordhausen	
Journel, André	693
R. Mohan Srivastava	
K-Means Clustering	695
Jaya Sreevalsan-Nair	
K-Medoids	697
Jaya Sreevalsan-Nair	

K-nearest Neighbors	700
Jaya Sreevalsan-Nair	
Kolmogorov, Andrey N.	702
Frits Agterberg	
Korvin, Gabor	703
B. S. Daya Sagar	
Krige, Daniel Gerhardus	704
Richard Minnitt	
Kriging	705
Xavier Freulon and Nicolas Desassis	
Krumbein, William Christian	713
E. H. Timothy Whitten	
Laplace Transform	715
Jaya Sreevalsan-Nair	
Least Absolute Deviation	717
A. Erhan Tercan	
Least Absolute Value	720
U. Radojičić and Klaus Nordhausen	
Least Mean Squares	724
Mark A. Engle	
Least Median of Squares	728
Sara Taskinen and Klaus Nordhausen	
Least Squares	731
Mark A. Engle	
LiDAR	735
Jaya Sreevalsan-Nair	
Linear Unmixing	739
Katherine L. Silversides	
Linearity	741
Christien Thiart	
Local Singularity Analysis	744
Wenlei Wang, Shuyun Xie and Zhijun Chen	
Locally Weighted Scatterplot Smoother	748
Klaus Nordhausen and Sara Taskinen	
Logistic Attractors	751
Balakrishnan Ashok	
Logistic Regression	756
D. Arun Kumar	
Logistic Regression, Weights of Evidence, and the Modeling Assumption of Conditional Independence	759
Helmut Schaeben	
Log-Likelihood Ratio Test	766
Alejandro C. Frery	

Lognormal Distribution	769
Glòria Mateu-Figueras and Ricardo A. Olea	
Lorenz, Edward Norton	773
Kerry Emanuel	
Loss Function	774
Tanujit Chakraborty and Uttam Kumar	
Machine Learning	781
Feifei Pan	
Mandelbrot, Benoit B.	784
Katepalli R. Sreenivasan	
Markov Chain Monte Carlo	785
Swathi Padmanabhan and Uma Ranjan	
Markov Chains: Addition	791
Adway Mitra	
Markov Random Fields	795
Uma Ranjan	
Marsily, Ghislain de	799
Craig T. Simmons	
Mathematical Geosciences	801
Qiuming Cheng	
Mathematical Minerals	818
John H. Doveton	
Mathematical Morphology	820
Jean Serra	
Matheron, Georges	835
Jean Serra	
Matrix Algebra	836
Deeksha Aggarwal and Uttam Kumar	
Maximum Entropy Method	842
Dionissios T. Hristopulos and Emmanouil A. Varouchakis	
Maximum Entropy Spectral Analysis	845
Eulogio Pardo-Igúzquiza and Francisco J. Rodríguez-Tovar	
Maximum Likelihood	851
Jaya Sreevalsan-Nair	
McCammon, Richard B.	853
Michael E. Hohn	
Membership Functions	854
Candan Gokceoglu	
Merriam, Daniel F.	857
D. Collins	
Metadata	858
Simon J. D. Cox	

Mine Planning	860
N. Caciagli	
Mineral Prospectivity Analysis	862
Emmanuel John M. Carranza	
Minimum Entropy Deconvolution	868
Jaya Sreevalsan-Nair	
Minimum Maximum Autocorrelation Factors	870
U. Mueller	
Mining Modeling	875
Youhei Kawamura	
Modal Analysis	881
Frits Agterberg	
Moments	883
Jessica Silva Lomba and Maria Isabel Fraga Alves	
Monte Carlo Method	890
Klaus Mosegaard	
Moran's Index	897
Giuseppina Giungato and Sabrina Maggio	
Morphological Closing	901
Aditya Challa, Sravan Danda and B. S. Daya Sagar	
Morphological Dilation	902
Sravan Danda, Aditya Challa and B. S. Daya Sagar	
Morphological Erosion	906
Aditya Challa, Sravan Danda and B. S. Daya Sagar	
Morphological Filtering	910
Jean Serra	
Morphological Opening	921
Sravan Danda, Aditya Challa and B. S. Daya Sagar	
Morphological Pruning	923
Sin Liang Lim and B. S. Daya Sagar	
Morphometry	928
Sebastiano Trevisani and Igor V. Florinsky	
Moving Average	934
Frits Agterberg	
Multidimensional Scaling	938
Francky Fouedjio	
Multifractals	945
Renguang Zuo	
Multilayer Perceptrons	951
C. Özgen Karacan	
Multiphase Flow	954
Juliana Y. Leung	

Multiple Correlation Coefficient	958
U. Mueller	
Multiple Point Statistics	960
Jef Caers, Gregoire Mariethoz and Julian M. Ortiz	
Multiscaling	970
Jaya Sreevalsan-Nair	
Multivariate Analysis	974
Monica Palma and Sabrina Maggio	
Multivariate Data Analysis in Geosciences, Tools	981
John H. Schuenemeyer	
Němec, Václav	985
Jan Harff and Niichi Nichiwaki	
Neural Networks	986
Vladik Kreinovich	
Nonlinear Mapping Algorithm	990
Jie Zhao, Wenlei Wang, Qiuming Cheng and Yunqing Shao	
Nonlinearity	994
Wenlei Wang, Jie Zhao and Qiuming Cheng	
Normal Distribution	999
Jaya Sreevalsan-Nair	
Object Boundary Analysis	1003
Hannes Thiergärtner	
Object-Based Image Analysis	1008
D. Arun Kumar, M. Venkatanarayana and V. S. S. Murthy	
Olea, Ricardo A.	1012
C. Özgen Karacan	
Ontology	1013
Torsten Hahmann	
Open Data	1017
Lucia R. Profeta	
Optimization in Geosciences	1020
Ilyas Ahmad Huqqani and Lea Tien Tay	
Ordinary Differential Equations	1024
R. N. Singh and Ajay Manglik	
Ordinary Least Squares	1032
Christopher Kotsakis	
Partial Differential Equations	1039
R. N. Singh and Ajay Manglik	
Particle Swarm Optimization in Geosciences	1047
Joseph Awange, Béla Paláncz and Lajos Völgyesi	
Pattern	1053
Uttam Kumar	

Pattern Analysis	1057
Sanchari Thakur, LakshmiKanthan Muralikrishnan, Bijal Chudasama and Alok Porwal	
Pattern Classification	1061
Katherine L. Silversides	
Pattern Recognition	1063
Rogério G. Negri	
Pawlowsky-Glahn, Vera	1065
Ricardo A. Olea	
Pengda, Zhao	1066
Qiuming Cheng and Frits Agterberg	
Plurigaussian Simulations	1067
Nasser Madani	
Point Pattern Statistics	1073
Dietrich Stoyan	
Polarimetric SAR Data Adaptive Filtering	1079
Mohamed Yahia, Tarig Ali, Md Maruf Mortula and Riadh Abdelfattah	
Pore Structure	1083
Weiren Lin and Nana Kamiya	
Porosity	1086
Pham Huy Giao and Pham Huy Nguyen	
Porous Medium	1090
Jennifer McKinley	
Power Spectral Density	1092
Abhey Ram Bansal and V. P. Dimri	
Predictive Geologic Mapping and Mineral Exploration	1095
Frits Agterberg	
Principal Component Analysis	1108
Alessandra Menafoglio	
Probability Density Function	1112
Abraão D. C. Nascimento	
Proximity Regression	1116
Jaya Sreevalsan-Nair	
Q-Mode Factor Analysis	1119
Norman MacLeod	
Quality Assurance	1126
N. Caciagli	
Quality Control	1130
N. Caciagli	
Quantitative Geomorphology	1135
Vikrant Jain, Shantamoy Guha and B. S. Daya Sagar	
Quantitative Stratigraphy	1152
Felix Gradstein and Frits Agterberg	

Quantitative Target Selection	1166
Brandon Wilson, Emmanuel John M. Carranza and Jeff B. Boisvert	
Quaternions and Rotations	1169
Helmut Schaeben	
Radial Basis Functions	1175
Michael Hillier	
Random Forest	1182
Emil D. Attanasi and Timothy C. Coburn	
Random Function	1185
J. Jaime Gómez-Hernández	
Random Variable	1188
Ricardo A. Olea	
Rank Score Test	1192
Alejandro C. Frery	
Rao, C. R.	1194
B. L. S. Prakasa Rao	
Rao, S. V. L. N.	1195
B. S. Daya Sagar	
Realizations	1196
C. Özgen Karacan	
Reduced Major Axis Regression	1199
Muhammad Bilal, Md. Arfan Ali, Janet E. Nichol, Zhongfeng Qiu, Alaa Mhawish and Khaled Mohamed Khedher	
Regression	1203
D. Arun Kumar, G. Hemalatha and M. Venkatanarayana	
Remote Sensing	1206
Ranganath Navalgund and Raghavendra P. Singh	
Reproducible Workflow	1209
Anirudh Prabhu and Peter Fox	
Rescaled Range Analysis	1213
Gabor Korvin	
Reyment, Richard A.	1218
Peter Bengtson	
R-Mode Factor Analysis	1219
Norman MacLeod	
Robust Statistics	1225
Peter Filzmoser	
Rock Fracture Pattern and Modeling	1229
Katsuaki Koike and Jin Wu	
Rodionov, Dmitriy Alekseevich	1234
Hannes Thiergärtner	
Rodriguez-Iturbe, Ignacio	1235
B. S. Daya Sagar	

Root Mean Square	1236
Tianfeng Chai	
Sampling Importance: Resampling Algorithms	1239
Anindita Dasgupta and Uttam Kumar	
Scaling and Scale Invariance	1244
S. Lovejoy	
Scattergram	1256
Richard J. Howarth	
Schuenemeyer, John H.	1259
Donald Gautier	
Schwarzacher, Walther	1260
Jennifer McKinley	
Sedimentation and Diagenesis	1261
Nana Kamiya and Weiren Lin	
Self-Organizing Maps	1264
Sid-Ali Ouadfeul, Leila Aliouane and Mohamed Zinelabidine Doghmane	
Sensitivity Analysis	1271
Valentina Ciriello and Daniel M. Tartakovsky	
Sequence Classification	1273
Mehala Balamurali	
Sequential Gaussian Simulation	1276
J. Jaime Gómez-Hernández	
Serra, Jean	1279
Philippe Salembier	
Shape	1280
Norman MacLeod	
Signal Analysis	1288
Mohamed Zinelabidine Doghmane, Sid-Ali Ouadfeul and Leila Aliouane	
Signal Processing in Geosciences	1297
E. Chandrasekhar and Rizwan Ahmed Ansari	
Simulated Annealing	1320
Jaya Sreevalsan-Nair	
Simulation	1322
Behnam Sadeghi and Julian M. Ortiz	
Singular Spectrum Analysis	1328
U. Radojčić, Klaus Nordhausen and Sara Taskinen	
Singularity Analysis	1332
Behnam Sadeghi and Frits Agterberg	
Smoothing Filter	1339
Alejandro C. Frery	
Spatial Analysis	1342
Qiyu Chen, Gang Liu, Xiaogang Ma and Xiang Que	

Spatial Autocorrelation	1345
Donato Posa and Sandra De Iaco	
Spatial Data	1353
Sabrina Maggio and Claudia Cappello	
Spatial Data Infrastructure and Generalization	1358
Jagadish Boodala, Onkar Dikshit and Nagarajan Balasubramanian	
Spatial Statistics	1362
Noel Cressie and Matthew T. Moores	
Spatiotemporal	1373
Sandra De Iaco, Donald E. Myers and Donato Posa	
Spatiotemporal Analysis	1382
Shrutilipi Bhattacharjee, Johannes Madl, Jia Chen and Varad Kshirsagar	
Spatiotemporal Modeling	1386
Shrutilipi Bhattacharjee, Johannes Madl, Jia Chen and Varad Kshirsagar	
Spatiotemporal Weighted Regression	1390
Xiang Que, Xiaogang Ma, Chao Ma, Fan Liu and Qiyu Chen	
Spectral Analysis	1396
Katherine L. Silversides	
Spectrum-Area Method	1398
Behnam Sadeghi	
Splines	1403
Bengt Fornberg	
Standard Deviation	1408
Alejandro C. Frery	
Stationarity	1410
Francky Fouedjio	
Statistical Bias	1414
Gilles Bourgault	
Statistical Computing	1428
Alan D. Chave	
Statistical Inferential Testing	1439
R. Webster	
Statistical Outliers	1443
Mehala Balamurali and Raymond Leung	
Statistical Quality Control	1451
Jörg Benndorf	
Statistical Rock Physics	1456
Gabor Korvin	
Statistical Seismology	1472
Jiancang Zhuang	
Stereographic Projections	1486
Frits Agterberg	

Stereology	1490
Leszek Wojnar	
Stochastic Geometry in the Geosciences	1501
Dietrich Stoyan	
Stoyan, Dietrich	1510
Jan Harff	
Stratigraphic Sequence	1511
Katherine L. Silversides	
Structuring Element	1513
B. S. Daya Sagar and D. Arun Kumar	
Sums of Geologic Random Variables, Properties	1516
G. M. Kaufman	
Support Vector Machines	1520
Rogério G. Negri	
t-Distributed Stochastic Neighbor Embedding	1527
Mehala Balamurali	
Text Mining	1535
Chengbin Wang and Xiaogang Ma	
Thickening	1537
B. S. Daya Sagar and D. Arun Kumar	
Thiergärtner, Hannes	1539
Heinz Burger	
Three-Dimensional Geologic Modeling	1540
Qiyu Chen, Gang Liu, Xiaogang Ma and Junqiang Zhang	
Time Series Analysis	1545
Abraão D. C. Nascimento	
Time Series Analysis in the Geosciences	1551
Klaus Nordhausen	
Tobler, Waldo	1559
Michael Batty	
Topology in Geosciences	1561
Florian Wellmann	
Total Alkali-Silica Diagram	1566
Ricardo A. Olea	
Trend Surface Analysis	1567
Frits Agterberg	
Tukey, John Wilder	1575
Frits Agterberg	
Turbulence	1576
Frits Agterberg	
Turcotte, Donald L.	1578
Lawrence Cathles and John Rundle	

Unbalanced Analysis of Variance	1579
R. Webster	
Uncertainty Quantification	1583
Behnam Sadeghi, Eric Grunsky and Vera Pawlowsky-Glahn	
Unit Regional Production Value	1589
Gunes Ertunc	
Unit Regional Value	1591
Frits Agterberg	
Univariate	1593
Alejandro C. Frery	
Universality	1597
Dionissios T. Hristopulos	
Upper Approximation	1600
Touhid Mohammad Hossain and Junzo Watada	
Variance	1605
Claudia Cappello and Monica Palma	
Variogram	1609
Behnam Sadeghi	
Very Fast Simulated Reannealing	1614
Dionissios T. Hristopulos	
Virtual Globe	1619
Jaya Sreevalsan-Nair	
Vistelius, Andrey Borisovich	1623
Stephen Henley	
Watson, Geoffrey S.	1625
Noel Cressie and Carol A. Gotway Crawford	
Wavelets in Geosciences	1626
Guoxiong Chen and Henglei Zhang	
Wave-Number Filtering	1636
Sid-Ali Ouadfeul and Leila Aliouane	
Whitten, Eric Harold Timothy	1642
John Cubitt	
Zadeh, Lotfi A.	1643
Behnam Sadeghi	
Z-Transform	1644
Sathishkumar Samiappan, Rajendra Mohan Panda and B. S. Daya Sagar	
Author Index	1647
Subject Index	1651

About the Editors



B. S. Daya Sagar (born February 24, 1967) is a Full Professor of the Systems Science and Informatics Unit at the Indian Statistical Institute. Sagar received his MSc (1991) and PhD (1994) degrees in Geoengineering and Remote Sensing from the Andhra University, India. He worked at Andhra University (1992–1998), The National University of Singapore (1998–2001), and Multimedia University, Malaysia (2001–2007). Sagar has made significant contributions to the field of mathematical earth sciences, remote sensing, spatial data sciences, and mathematical morphology. He has published over 90 papers in journals and has authored or guest-edited 14 books or journal special issues. He authored *Mathematical Morphology in Geomorphology and GISci* (CRC Press, 2013, p. 546) and *Handbook of Mathematical Geosciences* (Springer, 2018, p. 942). He was elected a Fellow of Royal Geographical Society, an IEEE Senior Member, a Fellow of the Indian Geophysical Union, and a Fellow of the Indian Academy of Sciences. He was awarded the Dr. Balakrishna Memorial Award – 1995, the IGU-Krishnan Medal – 2002, the IAMG “Georges Matheron Award – 2011,” the IAMG Certificate of Appreciation – 2018 Award, and the IEEE-GRSS Distinguished Lecturership Award. He is a member of the AGU’s Honors and Recognition Committee (2022–23). He is (was) on the Editorial Boards of *Computers and Geosciences*, *Frontiers: Environmental Informatics*, and *Mathematical Geosciences*. He is the Co-Editor-in-Chief of the Springer’s *Encyclopedia of Mathematical Geosciences*.



Qiuming Cheng is currently a Professor at the School of Earth Science and Engineering, Sun Yat-Sen University, Zhuhai, and the Founding Director of the State Key Lab of Geological Processes and Mineral Resources, China University of Geosciences (Beijing, Wuhan). He received his PhD in Earth Science from the University of Ottawa under the supervision of Dr. Frits Agterberg in 1994. Qiuming was a faculty member at York University, Toronto, Canada (1995–2015). Qiuming has specialized in mathematical geoscience with a research focus on nonlinear mathematical modeling of earth processes and geoinformatics for mineral resources prediction. He has authored over 300 research articles. He received the IAMG’s Krumbein Medal – 2008 and the AAG’s Gold Medal – 2020. He is a member of the Chinese Academy of Sciences and a foreign member of the Academia Europaea. He was President of the IAMG (2012–2016) and President of the IUGS (2016–2020). He

has promoted broadening the scope of MG from more traditional statistical geology to interdisciplinary mathematical geosciences and the collaboration among geounions and other organizations on big data and geo-intelligence in solid earth science and geoscientific input to the Future Earth Program. He initialized and established the IUGS big science program on Deep-time Digital Earth (DDE).



Jennifer McKinley is Professor of Mathematical Geoscience, in Geography and Director of the Centre for GIS and Geomatics at Queen's University Belfast (QUB), UK. As a Chartered Geologist, her research has focused on the application of spatial analysis techniques, including geostatistics, spatial data analysis, compositional data analysis, and Geographical Information Science (GIS), to natural resource management, soil and water geochemistry, environmental and criminal forensics, human health and the environment, and nature-based solutions for urban environments. Jennifer's international leadership roles include Councilor of the International Union of Geosciences (IUGS 2020–2024), President of the Governing Council of the Deep-time Digital Earth Initiative (DDE 2020–2024), and Past President of the International Association of Mathematical Geoscientists (IAMG 2016–2020). Jennifer has served on learned committees including the Royal Irish Academy and Geological Society of London and sits on the Giant's Causeway UNESCO World Heritage Site Steering Group.



Frederik Pieter Agterberg (born November 15, 1936) is a Dutch-Canadian mathematical geoscientist. He obtained BSc (1957), MSc (1959), and PhD (1961) degrees at Utrecht University. After 1 year as WARF postdoctoral fellow at the University of Wisconsin, he joined the Canadian Geological Survey (GSC) as “petrological statistician.” In 1969, he formed the GSC Geomathematics Section, which he headed until his retirement in 1996 when he became a GSC Emeritus Scientist. “Frits” taught Statistics in Geology course at the University of Ottawa from 1968 to 1992 before becoming an Adjunct Professor in the Earth Science Department at the same university. From 1983 to 1989 he was also Adjunct Professor in Mathematics Department at Carleton University. He was instrumental in establishing the International Association for Mathematical Geosciences (IAMG) in 1968, received the IAMG's William Christian Krumbein Medal in 1978, and was IAMG President from 2004 to 2008. He was named IAMG Distinguished Lecturer in 2004 and IAMG Honorary Member in 2017. Frits was made Correspondent of the Royal Netherlands Academy of Arts and Sciences in 1981. He is the sole author of 4 books, editor or co-editor of 11 other books, and author or co-author of over 300 scientific publications. He was co-editor of the *Handbook of Mathematical Geosciences* published in 2018.

Editorial Board

Eric Grunsky Department of Earth and Environmental Sciences, University of Waterloo, Waterloo, ON, Canada

Lim Sin Liang Faculty of Engineering, Multimedia University, Persiaran Multimedia, Selangor, Malaysia

Gang Liu School of Computer Science, China University of Geosciences, Wuhan, Hubei, China

Xiaogang (Marshall) Ma Department of Computer Science, University of Idaho, Moscow, ID, USA

Jaya Sreevalsan-Nair Graphics-Visualization-Computing Lab, International Institute of Information Technology Bangalore, Electronic City, Bangalore, Karnataka, India

Ricardo A. Olea Geology, Energy and Minerals Science Center, U.S. Geological Survey, Reston, VA, USA

Dinesh Sathyamoorthy Science and Technology Research Institute for Defence (STRIDE), Ministry of Defence, Malaysia

John H. Schuenemeyer Southwest Statistical Consulting, LLC, Cortez, CO, USA

Contributors

Rifaat Abdalla Department of Earth Sciences, College of Science, Sultan Qaboos University, Al-Khoudh, Muscat, Oman

Riadh Abdelfattah COSIM Lab, Higher School of Communications of Tunis, Université of Carthage, Carthage, Tunisia

Departement of ITI, Télécom Bretagne, Institut de Télécom, Brest, France

Mohamed Abioui Department of Earth Sciences, Faculty of Sciences, Ibn Zohr University, Agadir, Morocco

Deeksha Aggarwal Spatial Computing Laboratory, Center for Data Sciences, International Institute of Information Technology Bangalore (IIITB), Bangalore, Karnataka, India

Mohammad Ali Aghighi School of Minerals and Energy Resources Engineering, University of New South Wales, Sydney, NSW, Australia

Frits Agterberg Central Canada Division, Geomathematics Section, Geological Survey of Canada, Ottawa, ON, Canada

Md. Arfan Ali School of Marine Sciences, Nanjing University of Information Science and Technology, Nanjing, China

Tarig Ali GIS and Mapping Laboratory, American University of Sharjah United Arab Emirates, Sharjah, UAE

Department of Civil Engineering, American University of Sharjah, Sharjah, UAE

Andisheh Alimoradi Department of Mining Engineering, Imam Khomeini International University, Qazvin, Iran

Leila Aliouane LABOPHT, Faculty of Hydrocarbons and Chemistry, University of Boumerdes, Boumerdes, Algeria

Faculty of Hydrocarbons and Chemistry, University M'hamed Bougara of Boumerdes, Boumerdes, Algeria

Abdullah Alowairdhi Department of Computer Science, University of Idaho, Moscow, ID, USA

Farzin Amirkhani Shiraz Mine Environment and Hydrogeology Research Laboratory (MEHR Lab), University of Tehran, Tehran, Iran

Rizwan Ahmed Ansari Department of Electrical Engineering, Veermata Jijabai Technological Institute, Mumbai, India

Daisy Arroyo Department of Statistics, Universidad de Concepción, Concepcion, Chile

Igor Artemenko Heat and Mass Exchange in Geo-systems Department, Scientific Centre for Aerospace Research of the Earth, National Academy of Sciences of Ukraine, Kiev, Ukraine

D. Arun Kumar Department of Electronics and Communication Engineering and Center for Research and Innovation, KSRM College of Engineering, Kadapa, Andhra Pradesh, India

Balakrishnan Ashok Centre for Complex Systems and Soft Matter Physics, International Institute of Information Technology Bangalore, Electronics City, Bangalore, Karnataka, India

Emil D. Attanasi US Geological Survey, Reston, VA, USA

Joseph Awange School of Earth and Planetary Sciences, Discipline of Spatial Sciences, Curtin University, Perth, WA, Australia

Leonardo Azevedo CERENA, DECivil, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

John Bacon-Shone Social Sciences Research Centre, The University of Hong Kong, Hong Kong, China

Mehala Balamurali Rio Tinto Centre for Mine Automation, Australian Centre for Field Robotics, Faculty of Engineering, The University of Sydney, Sydney, NSW, Australia

Nagarajan Balasubramanian Department of Civil Engineering, Indian Institute of Technology Kanpur, Kanpur, UP, India

Abhey Ram Bansal Gravity and Magnetic Group, CSIR- National Geophysical Research Institute, Hyderabad, India

Ekta Baranwal Department of Civil Engineering, Jamia Millia Islamia, New Delhi, India

Michael Batty University College London, London, UK

Peter Bengtson Institute of Earth Sciences, Heidelberg University, Heidelberg, Germany

Jörg Benndorf TU Bergakademie Freiberg, Freiberg, Germany

Mohammed Benssaou Department of Earth Sciences, Faculty of Sciences, Ibn Zohr University, Agadir, Morocco

Keith Beven Lancaster University, Lancaster, UK

Shrutilipi Bhattacharjee National Institute of Technology Karnataka (NITK), Surathkal, India

Muhammad Bilal School of Marine Sciences, Nanjing University of Information Science and Technology, Nanjing, China

Jeff B. Boisvert University of Alberta, Edmonton, AB, Canada

Jagadish Boodala Department of Civil Engineering, Indian Institute of Technology Kanpur, Kanpur, UP, India

Gilles Bourgault Computer Modelling Group Ltd., Calgary, AB, Canada

Francesca Bovolo Center for Digital Society, Fondazione Bruno Kessler, Trento, Italy

Lorenzo Bruzzone Department of Information Engineering and Computer Science, University of Trento, Trento, Italy

Antonella Buccianti Department of Earth Sciences, University of Florence, Florence, Italy

Heinz Burger Geoscience, Free University Berlin, Berlin, Germany

N. Caciagli Metals Exploration, BHP Ltd, Toronto, ON, Canada
Barrick Gold Corp, Toronto, ON, USA

Jef Caers Department of Geological Sciences, Stanford University, Stanford, CA, USA

Claudia Cappello Dipartimento di Scienze dell'Economia, Università del Salento, Lecce, Italy

Emmanuel John M. Carranza Department of Geology, Faculty of Natural and Agricultural Sciences, University of the Free State, Bloemfontein, Republic of South Africa
University of KwaZulu-Natal, Durban, South Africa

Lawrence Cathles Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, NY, USA

Tianfeng Chai Cooperative Institute for Satellite Earth System Studies (CISESS), University of Maryland, College Park, MD, USA
NOAA/Air Resources Laboratory (ARL), College Park, MD, USA

Tanujit Chakraborty Spatial Computing Laboratory, Center for Data Sciences, International Institute of Information Technology Bangalore (IIITB), Bangalore, Karnataka, India

Aditya Challa Computer Science and Information Systems, APPCAIR, Birla Institute of Technology and Science, Pilani, Goa, India

E. Chandrasekhar Department of Earth Sciences, Indian Institute of Technology Bombay, Mumbai, India

Alan D. Chave Department of Applied Ocean Physics and Engineering, Deep Submergence Laboratory, Woods Hole Oceanographic Institution, Woods Hole, MA, USA

Guoxiong Chen State Key Laboratory of Geological Processes and Mineral Resources, China University of Geosciences, Wuhan, China

Jia Chen Technical University of Munich, Munich, Germany

Qiyu Chen School of Computer Science, China University of Geosciences, Wuhan, Hubei, China

Zhijun Chen China University of Geosciences, Wuhan, China

Qiuming Cheng School of Earth Science and Engineering, Sun Yat-Sen University, Zhuhai, China

State Key Lab of Geological Processes and Mineral Resources, China University of Geosciences, Beijing, China

George Christakos Department of Geography, San Diego State University, San Diego, CA, USA

Bijal Chudasama Geological Survey of Finland, Espoo, Finland

Valentina Ciriello Dipartimento di Ingegneria Civile, Chimica, Ambientale e dei Materiali (DICAM), Università di Bologna, Bologna, Italy

Timothy C. Coburn Department of Systems Engineering, Colorado State University, Fort Collins, CO, USA

David R. Cohen School of Biological, Earth and Environmental Sciences, The University of New South Wales, Sydney, NSW, Australia

D. Collins Emeritus Associate Scientist, The University of Kansas, Lawrence, KS, USA

Simon J. D. Cox CSIRO, Melbourne, VIC, Australia

Max Craglia European Commission –Joint Research Centre (JRC), Ispra, Italy

Carol A. Gotway Crawford Albuquerque, NM, USA

Noel Cressie School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW, Australia

John Cubitt Wrexham, UK

Alexander Damm Department of Geography, University of Zurich, Zürich, Switzerland
Eawag, Swiss Federal Institute of Aquatic Science & Technology, Surface Waters – Research and Management, Dübendorf, Switzerland

Sravan Danda Computer Science and Information Systems, APPCAIR, Birla Institute of Technology and Science, Pilani, Goa, India

Anindita Dasgupta Spatial Computing Laboratory, Center for Data Sciences, International Institute of Information Technology Bangalore (IIITB), Bangalore, Karnataka, India

John C. Davis Heinemann Oil GmbH, Baldwin City, KS, USA

B. S. Daya Sagar Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, Karnataka, India

Sandra De Iaco Department of Economics, Section of Mathematics and Statistics, University of Salento, National Biodiversity Future Center, Lecce, Italy

Nicolas Desassis Mines ParisTech, PSL University, Centre de Géosciences, Fontainebleau, France

Clayton V. Deutsch Centre for Computational Geostatistics, University of Alberta, Edmonton, AB, Canada

Liping Di Center for Spatial Information Science and Systems, George Mason University, Fairfax, VA, USA

Onkar Dikshit Department of Civil Engineering, Indian Institute of Technology Kanpur, Kanpur, UP, India

Roussos Dimitrakopoulos COSMO – Stochastic Mine Planning Laboratory, Department of Mining and Materials Engineering, McGill University, Montreal, Canada

V. P. Dimri CSIR-National Geophysical Research Institute, Hyderabad, Telangana, India

Mohamed Zinelabidine Doghmane Department of Geophysics, FSTGAT, University of Science and Technology Houari Boumediene, Algiers, Algeria

Faramarz Doulati Ardejani School of Mining, College of Engineering, University of Tehran, Tehran, Iran

Mine Environment and Hydrogeology Research Laboratory (MEHR Lab), University of Tehran, Tehran, Iran

John H. Doveton Kansas Geological Survey, University of Kansas, Lawrence, KS, USA

P. A. Dowd School of Civil, Environmental and Mining Engineering, The University of Adelaide, Adelaide, SA, Australia

Juan José Egozcue Department of Civil and Environmental Engineering, U. Politècnica de Catalunya, Barcelona, Spain

Kerry Emanuel Lorenz Center, Massachusetts Institute of Technology, Cambridge, MA, USA

Mark A. Engle Department of Earth, Environmental and Resource Sciences, University of Texas at El Paso, El Paso, Texas, USA

Department of Geological Sciences, University of Texas at El Paso, El Paso, TX, USA

Oktay Erten Centre for Computational Geostatistics, University of Alberta, Edmonton, AB, Canada

Gunes Ertunc Department of Mining Engineering, Hacettepe University, Ankara, Turkey

Peter Filzmoser Institute of Statistics & Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria

Igor V. Florinsky Institute of Mathematical Problems of Biology, Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, Pushchino, Moscow Region, Russia

Bengt Fornberg Department of Applied Mathematics, University of Colorado, Boulder, CO, USA

Francky Fouedjio Department of Geological Sciences, Stanford University, Stanford, CA, USA

Peter Fox Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY, USA

Maria Isabel Fraga Alves CEAUL & DEIO, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal

Andrew U. Frank Geoinformation, Technical University, Vienna, Austria

Roger H. French Material Sciences and Engineering, Case Western Reserve University, Cleveland, OH, USA

Alejandro C. Frery School of Mathematics and Statistics, Victoria University of Wellington, Wellington, New Zealand

Xavier Freulon Mines ParisTech, PSL University, Centre de Géosciences, Fontainebleau, France

Alice-Agnes Gabriel Department of Earth and Environmental Sciences, LMU Munich, Munich, Germany

John Gamon Center for Advanced Land Management Information Technologies, School of Natural Resources, University of Nebraska–Lincoln, Lincoln, NE, USA
Department of Earth and Atmospheric Sciences and Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada

Shib Sankar Ganguli CSIR-National Geophysical Research Institute, Hyderabad, Telangana, India

Fuchang Gao Department of Mathematics and Statistical Science, University of Idaho, Moscow, ID, USA

Michael T. Gastner Division of Science, Yale-NUS College, Singapore, Singapore

Donald Gautier DonGautier L.L.C., Palo Alto, CA, USA

Pham Huy Giao PetroVietnam University (PVU), Baria-Vung Tau, Vietnam
Vietnam Petroleum Institute (VPI), Hanoi, Vietnam

Giuseppina Giungato Department of Economics, Section of Mathematics and Statistics, University of Salento, Lecce, Italy

Candan Gokceoglu Department of Geological Engineering, Hacettepe University, Ankara, Turkey

J. Jaime Gómez-Hernández Research Institute of Water and Environmental Engineering, Universitat Politècnica de València, Valencia, Spain

Caterina Gozzi Department of Earth Sciences, University of Florence, Florence, Italy

Felix M. Gradstein Founding Member and Past Chair, Geologic Time Scale Foundation, Emeritus Natural History Museum, University of Oslo, Oslo, Norway

Dario Grana Department of Geology and Geophysics, University of Wyoming, Laramie, WY, USA

Cedric M. Griffiths Predictive Geoscience, StrataMod Pty. Ltd., Greenmount, WA, Australia
Department of Exploration Geophysics, Curtin University, Perth, Australia

Eric Grunsky Department of Earth and Environmental Sciences, University of Waterloo, Waterloo, ON, Canada

Shantamoy Guha Discipline of Earth Sciences, IIT Gandhinagar, Gandhinagar, India

Utkarsh Gupta Department of Computational and Data Sciences, Indian Institute of Science, Bengaluru, Karnataka, India

Torsten Hahmann Spatial Informatics, School of Computing and Information Science, University of Maine, Orono, ME, USA

Jan Harff Institute of Marine and Environmental Sciences, University of Szczecin, Szczecin, Poland

Junyu He Ocean College, Zhejiang University, Zhoushan, China

G. Hemalatha Department of Electronics and Communication Engineering, KSRM College of Engineering, Kadapa, Andhra Pradesh, India

Stephen Henley Resources Computing International Ltd, Matlock, UK

Michael Hillier Geological Survey of Canada, Ottawa, ON, Canada

Michael E. Hohn Morgantown, WV, USA

Touhid Mohammad Hossain Department of Computer and Information Sciences, Faculty of Science and Information Technology, Universiti Teknologi PETRONAS, Seri Iskandar, Malaysia

Richard J. Howarth Department of Earth Sciences, University College London (UCL), London, UK

Dionissios T. Hristopulos School of Electrical and Computer Engineering, Technical University of Crete, Chania, Crete, Greece

Karel Hron Department of Mathematical Analysis and Applications of Mathematics, Palacký University, Olomouc, Czech Republic

Fang Huang CSIRO Mineral Resources, Kensington, WA, Australia

Ilyas Ahmad Huqqani School of Electrical and Electronics Engineering, Universiti Sains Malaysia, Penang, Malaysia

James Irving Institute of Earth Sciences, University of Lausanne, Lausanne, Switzerland

Vikrant Jain Discipline of Earth Sciences, IIT Gandhinagar, Gandhinagar, India

Qingren Jia College of Electronic Science and Technology, National University of Defense Technology, Changsha, China

Behshad Jodeiri Shokri Department of Mining Engineering, Hamedan University of Technology, Hamedan, Iran

Nana Kamiya Earth and Resource System Laboratory, Graduate School of Engineering, Kyoto University, Kyoto, Japan
Laboratory of Earth System Science, School of Science and Engineering, Doshisha University, Kyotanabe, Japan

Kasturi Devi Kanniah TropicalMap Research Group, Centre for Environmental Sustainability and Water Security (IPASA), Faculty of Built Environment and Surveying, Universiti Teknologi Malaysia, Johor Bahru, Malaysia

C. Özgen Karacan U.S. Geological Survey, Geology, Energy and Minerals Science Center, Reston, VA, USA

G. M. Kaufman Sloan School of Management, MIT, Cambridge, MA, USA

Youhei Kawamura Division of Sustainable Resources Engineering, Hokkaido University, Sapporo, Hokkaido, Japan

Khaled Mohamed Khedher Department of Civil Engineering, College of Engineering, King Khalid University, Abha, Saudi Arabia
Department of Civil Engineering, High Institute of Technological Studies, Mrezgua University Campus, Nabeul, Tunisia

Peter K. Kitanidis Stanford University, Civil and Environmental Engineering & Computational and Mathematical Engineering, Stanford, CA, USA

Michael Klichowicz Institute of Mineral Processing Machines and Recycling Systems Technology, Technische Universität Bergakademie Freiberg, Freiberg, Germany

Jens Klump Mineral Resources, CSIRO, Perth, WA, Australia

Katsuaki Koike Department of Urban Management, Graduate School of Engineering, Kyoto University, Kyoto, Japan

Gabor Korvin Earth Science Department, King Fahd University of Petroleum and Minerals, Dhahran, Kingdom of Saudi Arabia

Yuriy Kostyuchenko Heat and Mass Exchange in Geo-systems Department, Scientific Centre for Aerospace Research of the Earth, National Academy of Sciences of Ukraine, Kiev, Ukraine
International Institute for Applied Systems Analysis (IIASA), Laxenbourg, Austria

Christopher Kotsakis Department of Geodesy and Surveying, School of Rural and Surveying Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece

Vladik Kreinovich Department of Computer Science, University of Texas at El Paso, El Paso, TX, USA

Vera Krieger Institute of Bio- and Geosciences, IBG-2: Plant Sciences, Forschungszentrum Jülich GmbH, Jülich, Germany

Uwe Kroner Technische Universität Bergakademie Freiberg, Freiberg, Germany

Varad Kshirsagar Birla Institute of Technology and Science, Pilani, Hyderabad, India

Uttam Kumar Spatial Computing Laboratory, Center for Data Sciences, International Institute of Information Technology Bangalore (IIITB), Bangalore, Karnataka, India

Kerstin Lehnert Columbia University, Lamont Doherty Earth Observatory, Palisades, NY, USA

Juliana Y. Leung University of Alberta, Edmonton, AB, Canada

Raymond Leung Rio Tinto Centre for Mine Automation, Australian Centre for Field Robotics, Faculty of Engineering, The University of Sydney, Sydney, NSW, Australia

Holger Lieberwirth Institute of Mineral Processing Machines and Recycling Systems Technology, Technische Universität Bergakademie Freiberg, Freiberg, Germany

Sin Liang Lim Faculty of Engineering, Multimedia University, Cyberjaya, Selangor, Malaysia

Weiren Lin Earth and Resource System Laboratory, Graduate School of Engineering, Kyoto University, Kyoto, Japan

Fan Liu Google, Sunnyvale, CA, USA

Gang Liu State Key Laboratory of Biogeology and Environmental Geology, Wuhan, Hubei, China

School of Computer Science, China University of Geosciences, Wuhan, Hubei, China

S. Lovejoy Physics, McGill University, Montreal, Canada

Xiaogang Ma Department of Computer Science, University of Idaho, Moscow, ID, USA

Chao Ma State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation, Chengdu University of Technology, Chengdu, China

Norman MacLeod Department of Earth Sciences and Engineering, Nanjing University, Nanjing, Jiangsu, China

Nasser Madani School of Mining and Geosciences, Nazarbayev University, Nur-Sultan city, Kazakhstan

Johannes Madl Technical University of Munich, Munich, Germany

- Sabrina Maggio** Dipartimento di Scienze dell'Economia, Università del Salento, Lecce, Italy
- Soroush Maghsoudy** School of Mining, College of Engineering, University of Tehran, Tehran, Iran
Mine Environment and Hydrogeology Research Laboratory (MEHR Lab), University of Tehran, Tehran, Iran
- Ajay Manglik** CSIR-National Geophysical Research Institute, Hyderabad, India
- Lorenzo Marchi** Consiglio Nazionale delle Ricerche – Istituto di Ricerca per la Protezione Idrogeologica, Padova, Italy
- Denis Marcotte** Department of Civil, Geological and Mining Engineering, Polytechnique Montreal, Montreal, QC, Canada
- Gregoire Mariethoz** Institute of Earth Surface Dynamics, University of Lausanne, Lausanne, Switzerland
- Daniele Marinelli** Department of Information Engineering and Computer Science, University of Trento, Trento, Italy
- Glòria Mateu-Figueras** Department of Informatics, Applied Mathematics and Statistics, University of Girona, Girona, Spain
Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Girona, Spain
- Paolo Mazzetti** Institute of Atmospheric Pollution Research, Earth and Space Science Informatics Laboratory (ESSI-Lab), National Research Council of Italy, Rome, Italy
- J. M. McArthur** Earth Sciences, UCL, London, UK
- Jennifer McKinley** Geography, School of Natural and Built Environment, Queen's University, Belfast, UK
- Alessandra Menafoglio** MOX-Department of Mathematics, Politecnico di Milano, Milano, Italy
- Martin Meschede** Institute of Geography and Geology, University of Greifswald, Greifswald, Germany
- Alaa Mhawish** School of Marine Sciences, Nanjing University of Information Science and Technology, Nanjing, China
- Richard Minnitt** School of Mining Engineering, University of the Witwatersrand, Johannesburg, South Africa
- Adway Mitra** Centre of Excellence in Artificial Intelligence, Indian Institute of Technology Kharagpur, Kharagpur, India
- Konstantinos Modis** School of Mining and Metallurgical Engineering, National Technical University of Athens, Athens, Greece

Marcus Mohr Department of Earth and Environmental Sciences, LMU Munich, Munich, Germany

Gianna Serafina Monti Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy

Matthew T. Moores School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW, Australia

Md Maruf Mortula Department of Civil Engineering, American University of Sharjah, Sharjah, UAE

Klaus Mosegaard Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark

U. Mueller School of Science, Edith Cowan University, Joondalup, WA, Australia

Onno Muller Institute of Bio- and Geosciences, IBG-2: Plant Sciences, Forschungszentrum Jülich GmbH, Jülich, Germany

LakshmiKanthan Muralikrishnan Cybertech Systems and Software Limited, Thane, Maharashtra, India

V. S. S. Murthy KSRM College of Engineering, Kadapa, Andhra Pradesh, India

Donald E. Myers Department of Mathematics, University of Arizona, Tucson, AZ, USA

Abraão D. C. Nascimento Universidade Federal de Pernambuco, Recife, Brazil

Stefano Nativi Joint Research Centre (JRC), European Commission, Ispra, Italy

Ranganath Navalgund Indian Space Research Organisation (ISRO), Bengaluru, India

Rogério G. Negri São Paulo State University (UNESP), Institute of Science and Technology (ICT), São José dos Campos, São Paulo, Brazil

William I. Newman Departments of Earth, Planetary, & Space Sciences, Physics & Astronomy, and Mathematics, University of California, Los Angeles, CA, USA

Pham Huy Nguyen Imperial College, London, UK

Niichi Nichiwaki Professor Emeritus of Nara University, Nara, Japan

Janet E. Nichol Department of Geography, School of Global Studies, University of Sussex, Brighton, UK

Klaus Nordhausen Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland

Ricardo A. Olea Geology, Energy and Minerals Science Center, U.S. Geological Survey, Reston, VA, USA

Henning Omre Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

Julian M. Ortiz The Robert M. Buchan Department of Mining, Queen's University, Kingston, ON, Canada

Sid-Ali Ouadfeul Algerian Petroleum Institute, Sonatrach, Boumerdes, Algeria

Swathi Padmanabhan Indian Institute of Science, Bangalore, Karnataka, India

Béla Paláncz Department of Geodesy and Surveying, Budapest University of Technology and Economics, Budapest, Hungary

Javier Palarea-Albaladejo Biomathematics and Statistics Scotland, Edinburgh, UK

Monica Palma Università del Salento-Dip. Scienze dell'Economia, Complesso Ecotekne, Lecce, Italy

Guocheng Pan Hanking Industrial Group Limited, Shenyang, Liaoning, China

Feifei Pan Rensselaer Polytechnic Institute, Troy, NY, USA

Rajendra Mohan Panda Geosystems Research Institute, Mississippi State University, Starkville, MS, USA

H. S. Pandalai Department of Earth Sciences, Indian Institute of Technology Bombay, Mumbai, India

Narayan Panigrahi Scientist and Head of GIS Division, Center for Artificial Intelligence and Robotics (CAIR), Bangalore, India

Madhurima Panja Spatial Computing Laboratory, Center for Data Sciences, International Institute of Information Technology Bangalore (IIITB), Bangalore, Karnataka, India

Eulogio Pardo-Igúzquiza Instituto Geológico y Minero de España (IGME), Madrid, Spain

Vera Pawlowsky-Glahn Department of Computer Science, Applied Mathematics and Statistics, Universitat de Girona, Girona, Spain

Daniele Pedretti Dipartimento di Scienze della Terra "A. Desio", Università degli Studi di Milano, Milan, Italy

Alok Porwal Centre for Studies in Resources Engineering, Indian Institute of Technology Bombay, Mumbai, India

Centre for Exploration Targeting, University of Western Australia, Crawley, WA, Australia

Donato Posa Department of Economics, Section of Mathematics and Statistics, University of Salento, Lecce, Italy

Amin Beiranvand Pour Institute of Oceanography and Environment (INOS), Universiti Malaysia Terengganu (UMT), Terengganu, Malaysia

Anirudh Prabhu Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY, USA

B. L. S. Prakasa Rao CR Rao Advanced Institute of Mathematics, Statistics and Computer Science, Hyderabad, India

Lucia R. Profeta Lamont-Doherty Earth Observatory, Columbia University in the City of New York, New York, NY, USA

Zhongfeng Qiu School of Marine Sciences, Nanjing University of Information Science and Technology, Nanjing, China

Xiang Que Computer and Information College, Fujian Agriculture and Forestry University, Fuzhou, Fujian, China

Department of Computer Science, University of Idaho, Moscow, ID, USA

Juan Quiros-Vargas Institute of Bio- and Geosciences, IBG-2: Plant Sciences, Forschungszentrum Jülich GmbH, Jülich, Germany

U. Radojičić Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria

Majid Rahimzadegan Water Resources Engineering and Management Department, Faculty of Civil Engineering, K. N. Toosi University of Technology, Tehran, Iran

Sarah Ramdeen Columbia University, Lamont Doherty Earth Observatory, Palisades, NY, USA

Uma Ranjan Department of Computational and Data Sciences, Indian Institute of Science, Bengaluru, Karnataka, India

Uwe Rascher Institute of Bio- and Geosciences, IBG-2: Plant Sciences, Forschungszentrum Jülich GmbH, Jülich, Germany

Francisco J. Rodríguez-Tovar Universidad de Granada, Granada, Spain

Hamid Roshan School of Minerals and Energy Resources Engineering, University of New South Wales, Sydney, NSW, Australia

John Rundle Departments of Physics and Earth and Planetary Sciences, Earth and Physical Sciences, University of California Davis, Davis, CA, USA

Behnam Sadeghi EarthByte Group, School of Geosciences, University of Sydney, Sydney, NSW, Australia

Earth and Sustainability Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW, Australia

Ramendra Sahoo Discipline of Earth Sciences, IIT Gandhinagar, Gandhinagar, India
Department of Earth and Climate Science, IISER Pune, Pune, India

Philippe Salembier Department of Signal Theory and Communication, Universitat Politècnica de Catalunya, BarcelonaTech, Barcelona, Spain

Sathishkumar Samiappan Geosystems Research Institute, Mississippi State University, Starkville, MS, USA

Helmut Schaeben Technische Universität Bergakademie Freiberg, Freiberg, Germany

Bernhard S. A. Schuberth Department of Earth and Environmental Sciences, LMU Munich, Munich, Germany

John H. Schuenemeyer Southwest Statistical Consulting, LLC, Cortez, CO, USA

Jean Serra Centre de morphologie mathématique, Ecoles des Mines, Paristech, Paris, France

Fojan Shafaei School of Mining, College of Engineering, University of Tehran, Tehran, Iran
Mine Environment and Hydrogeology Research Laboratory (MEHR Lab), University of Tehran, Tehran, Iran

Majid Shahhosseiny School of Mining, College of Engineering, University of Tehran, Tehran, Iran
Mine Environment and Hydrogeology Research Laboratory (MEHR Lab), University of Tehran, Tehran, Iran

Yunqing Shao China University of Geosciences, Beijing, China

Mi Shu Tencent Technology (Beijing) Co., Ltd, Beijing, China

Bastian Siegmann Institute of Bio- and Geosciences, IBG-2: Plant Sciences, Forschungszentrum Jülich GmbH, Jülich, Germany

Jorge F. Silva Information and Decision System Group (IDS), Department of Electrical Engineering, University of Chile, Santiago, Chile

Jessica Silva Lomba CEAUL, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal

Katherine L. Silversides Australian Centre for Field Robotics, Rio Tinto Centre for Mining Automation, The University of Sydney, Camperdown, NSW, Australia

Craig T. Simmons College of Science and Engineering & National Centre for Groundwater Research and Training, Flinders University, Adelaide, SA, Australia

Donald A. Singer U.S. Geological Survey, Cupertino, CA, USA

R. N. Singh Discipline of Earth Sciences, Indian Institute of Technology, Gandhinagar, Palaj, Gandhinagar, India

Raghavendra P. Singh Space Applications Centre (ISRO), Ahmedabad, India

Sanjay Singh Signal and Image Processing Group, Space Applications Centre (ISRO), Ahmedabad, GJ, India

Lowell Kent Smith Emeritus, University of Redlands, Redlands, CA, USA

Amílcar Soares CERENA, DECivil, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

Indu Solomon Spatial Computing Laboratory, Center for Data Sciences, International Institute of Information Technology Bangalore (IIITB), Bangalore, Karnataka, India

Katepalli R. Sreenivasan New York University, New York, NY, USA

Jaya Sreevalsan-Nair Graphics-Visualization-Computing Lab, International Institute of Information Technology Bangalore, Electronic City, Bangalore, Karnataka, India

Pradeep Srivastava Earth Sciences, Indian Institute of Technology, Gandhinagar, Gujarat, India

R. Mohan Srivastava TriStar Gold Inc, Toronto, ON, Canada

Tobias Stephan Department of Geoscience, University of Calgary, Calgary, AB, Canada

Dietrich Stoyan Institut für Stochastik, TU Bergakademie Freiberg, Freiberg, Germany

Shaoqiang Su Computer and Information College, Fujian Agriculture and Forestry University, Fuzhou, Fujian, China

Key Laboratory of Fujian Universities for Ecology and Resource Statistics, Fuzhou, China

A. Subramanyam Department of Mathematics, Indian Institute of Technology Bombay, Mumbai, India

Ziheng Sun Center for Spatial Information Science and Systems, George Mason University, Fairfax, VA, USA

Pejman Tahmasebi College of Engineering and Applied Science, University of Wyoming, Laramie, WY, USA

Daniel M. Tartakovsky Department of Energy Resources Engineering, Stanford University, Stanford, CA, USA

Sara Taskinen Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland

Lea Tien Tay School of Electrical and Electronics Engineering, Universiti Sains Malaysia, Penang, Malaysia

A. Erhan Tercan Department of Mining Engineering, Hacettepe University, Ankara, Turkey

Sanchari Thakur University of Trento, Trento, Italy

Christien Thiart Department of Statistical Sciences, University of Cape Town, Cape Town, South Africa

Hannes Thiergärtner Department of Geosciences, Free University of Berlin, Berlin, Germany

Rahisha Thottolil Spatial Computing Laboratory, Center for Data Sciences, International Institute of Information Technology Bangalore (IIITB), Bangalore, Karnataka, India

Yuanyuan Tian School of Geographical Science and Urban Planning, Arizona State University, Tempe, AZ, USA

Shohei Albert Tomita Department of Urban Management, Graduate School of Engineering, Kyoto University, Kyoto, Japan

Sebastiano Trevisani Università Iuav di Venezia, Venice, Italy

Emmanouil A. Varouchakis School of Mineral Resources Engineering, Technical University of Crete, Chania, Greece

Parvatham Venkatachalam Centre of Studies in Resources Engineering, Indian Institute of Technology, Mumbai, India

M. Venkatanarayana Department of Electronics and Communication Engineering and Center for Research and Innovation, KSRM College of Engineering, Kadapa, Andhra Pradesh, India

Jay M. Ver Hoef Alaska Fisheries Science Center, NOAA Fisheries, Seattle, WA, USA

Eric P. Verrecchia Institute of Earth Surface Dynamics, University of Lausanne, Lausanne, Switzerland

Jaume Soler Villanueva E.T.S. de Ingeniería de Caminos, Universitat Politècnica de Catalunya, Barcelona, Spain

Lajos Völgyesi Department of Geodesy and Surveying, Budapest University of Technology and Economics, Budapest, Hungary

Chengbin Wang State Key Laboratory of Geological Processes and Mineral Resources & School of Earth Resources, China University of Geosciences, Wuhan, China

Ran Wang Center for Advanced Land Management Information Technologies, School of Natural Resources, University of Nebraska–Lincoln, Lincoln, NE, USA

Wenlei Wang Institute of Geomechanics, Chinese Academy of Geological Sciences, Beijing, China

Junzo Watada Graduate School of Information, Production Systems, Waseda University, Fukuoka, Japan

R. Webster Rothamsted Research, Harpenden, UK

Florian Wellmann Computational Geoscience and Reservoir Engineering, RWTH Aachen University, Aachen, Germany

Tao Wen Department of Earth and Environmental Sciences, Syracuse University, Syracuse, NY, USA

Johannes Wendebourg TotalEnergies Exploration, Paris, France

E. H. Timothy Whitten Riverside, Widecombe-in-the-Moor, Devon, UK

Brandon Wilson University of Alberta, Edmonton, AB, Canada

Leszek Wojnar Krakow University of Technology, Krakow, Poland

Jin Wu China Railway Design Corporation, Tianjin, China

Lesley Wyborn Research School of Earth Sciences, Australian National University, Canberra, ACT, Australia

Shuyun Xie State Key Laboratory of Geological Processes and Mineral Resources (GPMR), Faculty of Earth Sciences, China University of Geosciences, Wuhan, China

Mohamed Yahia GIS and Mapping Laboratory, American University of Sharjah United Arab Emirates, Sharjah, UAE

Laboratoire de recherche Modélisation analyse et commande de systèmes- MACS, Ecole Nationale d'Ingénieurs de Gabes – ENIG, Université de Gabes Tunisia, Zrig Eddakhlania, Tunisia

Lingqing Yao COSMO – Stochastic Mine Planning Laboratory, Department of Mining and Materials Engineering, McGill University, Montreal, Canada

Jeffrey M. Yarus Material Sciences and Engineering, Case Western Reserve University, Cleveland, OH, USA

Jordan M. Yarus Geoscience Lead Site Reliability Engineering Halliburton, Houston, TX, USA

Hongfei Zhang State Key Laboratory of Biogeology and Environmental Geology, Wuhan, Hubei, China

School of Earth Science, China University of Geosciences, Wuhan, Hubei, China

Henglei Zhang Institution of Geophysics and Geomatics, China University of Geosciences, Wuhan, China

Junqiang Zhang School of Computer Science, China University of Geosciences, Wuhan, China

Jie Zhao School of the Earth Sciences and Resources, China University of Geosciences, Beijing, China

Jiancang Zhuang The Institute of Statistical Mathematics, Research Organization of Information and Systems, Tokyo, Japan

Sergio Zlotnik E.T.S. de Ingeniería de Caminos, Universitat Politècnica de Catalunya, Barcelona, Spain

Renguang Zuo State Key Laboratory of Geological Processes and Mineral Resources, China University of Geosciences, Wuhan, China

Accuracy and Precision

P. A. Dowd

School of Civil, Environmental and Mining Engineering,
The University of Adelaide, Adelaide, SA, Australia

Definitions

Accuracy is a measure of how close a measured value of a variable is to its true value. In the context of the mathematical geosciences, it could also be a measure of how close an estimated value of a variable is to the unknown true value at a given location when the estimate is obtained as some function of the measured data values. In the latter sense, the accuracy comprises two components: the accuracy of the measurements of the individual data values and the accuracy of the estimate based on those values.

Precision is a measure of the extent to which independent measurements of the same variable agree when using the same measuring device and procedure, i.e., the degree of reproducibility. Precision is also used to refer to the number of significant digits used in the specified measured value. For example, when comparing two measuring devices, if one can measure in smaller increments than the other, it has a higher precision of measurement. To distinguish this definition from the previous one, it will be referred to as the resolution of the measuring device.

As accuracy is independent of precision, a measurement may be precise but inaccurate or it may be accurate but not precise. It is difficult to achieve high accuracy without a good level of precision. A common way of illustrating the concepts of accuracy and precision is to use the game of darts where the objective is to throw multiple darts at a dart board with the objective of hitting the bull's eye at the center of the board. In terms of the accuracy and precision of the dart thrower, there are four general outcomes: (Fig. 1)

Both accuracy and precision are related to uncertainty. Uncertainty arises from the inability to measure a variable exactly and is quantified as the range of values of the variable within which the true value lies. The uncertainty may be due to the calibration of a measuring device or the manner in which a sample is collected and/or analyzed. Uncertainty also refers to the range of values of an estimated value of a variable when the estimate is obtained as some function of a set of measured data values.

The most common measure of accuracy is the relative error in a set of n measurements:

$$\frac{\sum_{i=1}^n |\text{measured value}_i - \text{true value}|}{\text{true value}}$$

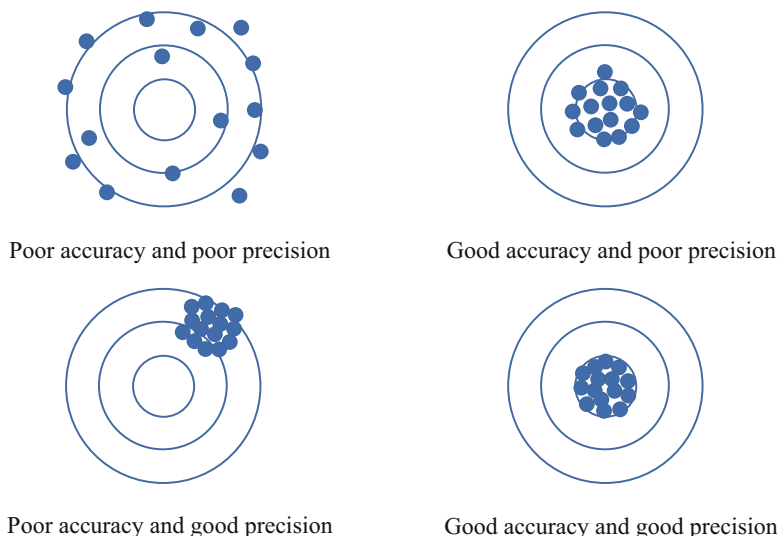
The most common measure of precision is the coefficient of variation:

$$\frac{\sqrt{\frac{\sum_{i=1}^n (\text{measured value}_i - \text{mean measured value})^2}{n-1}}}{\text{mean measured value}}$$

The Mathematical Geosciences Context

Geoscience variables may be quantitative or qualitative. These variables may be measured directly by collecting physical samples and subjecting them to some process or they may be measured indirectly by using one or more of many sensing devices.

Quantitative variables are numerical and they measure specific quantities. They are generally measured on specific volumes and their values are averages over those volumes. For example, the gold content of a specific volume of rock is measured as the proportion of the volume of rock that is gold. As the variance of such variables is inversely proportional to the volume on which they are measured, values measured on different volumes should not be combined without accounting for the volume-variance effect.

Accuracy and Precision, Fig. 1

Qualitative (or categorical) variables may also be measured on specific volumes and they may be nominal (e.g., rock types or the presence or absence of a characteristic such as lithology) or ordinal (i.e., an ordered rank or sequence in which the distance between the elements is irrelevant). Examples of the latter include measures of mineral hardness or rock strength characteristics. Quantifying the uncertainty of qualitative variables is more problematic see, for example, Pulido et al. (2003) for examples of methods, and Lindsay et al. (2012) for quantifying geological uncertainty in three-dimensional modelling.

All geoscience variables are spatial in the sense that their values are a function of the location at which they are measured, and they generally have a level of spatial correlation that reflects the structure of the geological or geomorphological setting in which they occur.

The accuracy of a measurement of a variable can only be known if the true value of the variable is known. If the true value is not known, which is the case for many geoscience variables, then the accuracy of a measurement can only be estimated. Assuming that the measurement instrument is properly calibrated, and the approved measurement process is followed, a practical method of quantifying accuracy is by repeatedly measuring the same sample and expressing the accuracy as:

$$\frac{\sum_{i=1}^n |\text{measured}_i - \text{average}|}{\text{average}}$$

However, in the geosciences, often only some representative proportion of the entire sample is measured. In addition, the direct quantitative measurement of many types of samples requires destructive testing of a subsample of the total sample. A common example is measuring the mineral content of a cylindrical drill core sample. In common practice, the core is

split longitudinally, and one half is kept for reference. The other half-core is crushed to a given particle size and a subsample is taken, which is then crushed to a smaller particle size and again resampled. This procedure is repeated a number of times until a final subsample is obtained and chemically analyzed for mineral content. Each step of the procedure may introduce error and/or affect the representativity of the subsample and hence affect accuracy and precision. For a detailed account of the sampling of particulate materials, see Gy (1977) and Royle (2001). For a detailed approach to accuracy and precision in sampling particulate materials for analytical purposes, see Gy (1998). For quantifying uncertainty in analytical measurements see Ellison and Williams (2012).

Resolution of the Measuring Device: Significant Figures

It is useful to distinguish between exact numbers and measured numbers. An exact number is a number that is perfectly known. It could be a number obtained by counting items, such as the number of samples prepared for analysis. It could also be a defined number, such as $1 \text{ m} = 100 \text{ cm}$. Exact numbers have an infinite number of significant figures.

The interest here is in measured numbers, which are not exactly known because of the limitations of the measuring device and process, and in values that are calculated from measured numbers. An indication of the uncertainty of a measured value is the number of significant figures used to report it. Results of calculations should not be reported to a greater level of precision than the data. The rules for determining the number of significant figures are:

- Nonzero digits are always significant.
- Zeros between significant digits are significant. For example, 3045 has four significant digits.
- Zeros to the left of the first nonzero digit are not significant. For example, 0.00542 has three significant digits. This is clearer when using exponential notation: 5.42×10^{-3} .
- A final zero and all trailing zeros in the decimal portion of a number are significant. For example, the final two zeros in 0.0004700 in which there are four significant (4700) digits and 6.720×10^4 in which all zeros are significant.
- Trailing zeros are not significant in numbers without decimal points. For example, 250 has two significant figures and 5000 or 5×10^3 has only one significant digit.

Sensed Data

Some variables are not directly measured but are extracted from a signal generated by a sensing device. Such a device might be used locally (e.g., lowered into a drillhole to sense mineral grades at incremental depths) or used remotely (e.g., from a satellite or an aircraft to sense features of the surface of the Earth). Widely used subsurface sensing techniques include seismic tomography, magnetotellurics, and ground-penetrating radar. These signal data must be processed to extract the required quantitative or qualitative data. For quantitative variables, such as mineral grades, the depth that the signal penetrates the rock may differ depending on the local characteristics of the rock and thus each signal measurement at different locations may refer to different volumes. Such measurements cannot be combined without accounting for the volume-variance relationship.

Quantification of the accuracy and precision of sensed geoscience data has largely been limited to the surface of the Earth using techniques such as Landsat and LiDAR. These techniques can detect and measure surface features and generate two-dimensional and three-dimensional data with relatively high accuracy and precision. Congalton and Green (2020) is a good reference for quantifying the accuracy and precision of these techniques.

Quantifying the accuracy and precision of subsurface sensed data is more complex. In many cases, the detection of significant qualitative features is relatively accurate. However, it is much more difficult to quantify the accuracy and precision of quantitative measurements. The increasing use of sensed quantitative subsurface data requires some meaningful measure of accuracy and precision, especially when these data are integrated with directly measured data, which will have different levels of accuracy and precision.

An indication of the increasing use of subsurface sensed data for minerals, energy, and water resources, including the identification and reduction of all sources of uncertainty, can be found in the Deep Earth Imaging Project (CSIRO 2016).

Integrating Sensed and Directly Measured Data

Until recently, sensed data in the geosciences have largely been limited to qualitative geological and geomorphological characteristics and to quantitative variables on the surface, or near subsurface, of the earth particularly for applications such as hydrology (e.g., Reichle 2008). For subsurface applications, the exceptions have been quantitative variables, such as elastic moduli and rock-strength parameters, that can be derived from sensed variables such as shear velocity, compressive velocity, and bulk density; see Dowd (1997) for an example of deriving these variables from measured variables. Increasingly, direct quantitative subsurface sensing devices (e.g., measuring mineral grades while drilling) are being developed and implemented. In economic geology and mineral resource applications these sensed data are used together with directly measured values of the same characteristic or property (e.g., sensed mineral grade and directly measured grade) usually at different locations. This requires the two types of data to be integrated in a way that accommodates the different accuracies and uncertainties of the measured values and the different volumes of measurements. This is an area of ongoing research in many areas of the mathematical geosciences. See Careassi et al. (2018) for approaches in environmental modelling.

Using Integrated Data Types for Estimation

Quantitative data may be used to estimate the value of a variable at an unsampled location on the same volume as the data or, more often, on a significantly larger volume. Qualitative data may be used to estimate the presence or absence of a characteristic or feature at an unsampled location or to estimate the boundaries of features or in the various forms of geological domaining. These (statistical or geostatistical) estimates are, or should be, accompanied by a measure of the accuracy of the estimate. In this context, the accuracy refers to the uncertainty of the estimate. The estimate will also be reported with a level of precision.

For various reasons, more often than not, the measure of the uncertainty of the estimate does not include the uncertainty of the data used in the estimate. See de Marsily (1986) for kriging with uncertain data and an application in Cecinati et al. (2018). As the amount of sensed data increases, there is a corresponding need to account appropriately for the relative uncertainties of directly measured and sensed data especially in mineral resource applications and more generally in economic geology.

The data are also used to inform the model used for estimation; for example, the variogram used in geostatistical estimation (cf. Journel and Huijbregts 1978, 1984). The amount of data available, and their relative spatial locations,

will affect the quantification of spatial variability in the variogram model, which in turn will affect the accuracy of estimates. The various forms of kriging yield estimation variances that quantify the accuracy of the estimate, but the quantification does not generally include the uncertainty of the parameters in the variogram model. In addition, sparse data at significant distances from the location at which the estimate is made will overestimate low values and underestimate high values (Goovaerts 1997), thus affecting the accuracy and precision of the estimate. There is a need, especially in economic geology and mineral resource estimation, for the adequate incorporation of these sources of uncertainty in the estimates.

Summary

This entry summarizes the concepts of accuracy and precision in the context of the mathematical geosciences. Both concepts relate to direct, and sensed, in situ measurements of variables, to the integration of these two types of measurement, to the quantification of the spatial variability of these two types of measurement, and to the use of the measurements to estimate values of variables at unsampled locations. It also identifies areas that require further research to enable a greater understanding of accuracy and precision and to include additional sources of uncertainty that affect accuracy and precision.

Cross-References

- [Geologic Time Scale Estimation](#)
- [Geostatistics](#)
- [Kriging](#)
- [Variogram](#)
- [Variance](#)

Bibliography

- Careassi A, Bocquet M, Bertino L, Evenson G (2018) Data assimilation in the geosciences: an overview of methods, issues, and perspectives. *WIREs Clim Change* 9:5
- Cecinati F, Moreno-Ródenas AM, Rico-Ramirez MA, ten Veldhuis M-C, Langeveld JG (2018) Considering rain gauge uncertainty using kriging for uncertain data. *Atmosphere* 9:446
- Congalton RG, Green K (2020) Assessing the accuracy of remotely sensed data: principles and practices, 3rd edn. CRC Press, 348pp
- CSIRO (2016). <https://research.csiro.au/dei/about-us/>
- De Marsily G (1986) Quantitative hydrogeology: groundwater hydrology for engineers. Academic, San Diego, 440pp
- Dowd PA (1997) Geostatistical characterization of three-dimensional spatial heterogeneity of rock properties at Sellafield. *Trans Inst Min Metall, Section A* 106:A133–A147
- Ellison SLR, Williams A (eds) (2012) Quantifying uncertainty in analytical measurement. EURACHEM/CITAC Guide CG 4, 3rd edn. ISBN 978-0-948926-30-3

- Gy P (1977) The sampling of particulate materials – theory and practice. Elsevier, 431pp. ISBN: 978-0444420794
- Gy P (1998) Sampling for analytical purposes. Wiley, 153pp. ISBN: 0-471-97956-2
- Journel AG, Huijbregts CJ (1978) Mining geostatistics. Academic, 610pp. ISBN-13: 978-0123910509. Re-published by Blackburn Press (2004), ISBN-13: 978-1930665910
- Lindsay MD, Aillères L, Jessell MW, de Kemp MW, Betts PG (2012) Locating and quantifying geological uncertainty in three-dimensional models: analysis of the Gippsland Basin, South-Eastern Australia. *Dent Tech* 546–547:10–27
- Pulido A, Ruisánchez I, Boqué R, Rius FX (2003) Uncertainty of results in routine qualitative analysis. *Trends Anal Chem* 22(10):64–654
- Reichle RH (2008) Data assimilation methods in the Earth sciences. *Adv Water Resour* 31:1411–1418
- Royle AG (2001) Simulations of gold exploration samples. *Trans Inst Min Metall: Section B Appl Earth Sci* 110(3):136–144

Additive Logistic Normal Distribution

Gianna Serafina Monti¹, Glòria Mateu-Figueras^{2,3} and Karel Hron⁴

¹Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy

²Department of Informatics, Applied Mathematics and Statistics, University of Girona, Girona, Spain

³Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Girona, Spain

⁴Department of Mathematical Analysis and Applications of Mathematics, Palacký University, Olomouc, Czech Republic

Synonyms

Logistic normal; Logratio normal; Normal on the simplex

Definition

The *additive logistic normal* model is a family of distributions of compositional data defined on the simplex using the logratio approach (Aitchison 1986; Pawlowsky-Glahn et al. 2015) and the multivariate normal distribution for real random vectors. It is the most used distribution on the simplex introduced by Aitchison (1982) as an alternative to the well-known Dirichlet model. For the construction of the distribution, standard strategy of the logratio methodology using the principle of working on coordinates is applied; specifically, the multivariate normal distribution in the real space is used to model the additive logratio representation of the random composition.

Originally, the model was defined using the additive logratio (alr) representation, and hence the name additive

logistic normal was assigned. However, other orthonormal logratio representations, originally known as isometric logratio (ilr) representation (Egozcue et al. 2003), can be used too. For this reason, the term isometric logistic normal or, more general, the term logistic normal is adopted for this model as well. Later, motivated by the principle of working on coordinates the term normal distribution on the simplex comes to emphasize the idea of staying on the simplex (Mateu-Figueras et al. 2013). Nowadays, the term logratio normal is also preferred by some authors (Comas-Cufi et al. 2016) to highlight the use of logratio coordinates instead of logistic transformation.

Sample Space of Compositional Data and Its Geometry

The content of this section is largely covered in the chapter ▶ “Additive Logistic Skew-Normal Distribution,” and more extended in the chapter ▶ “Compositional Data”. For readers’ convenience, we introduce here only notation and some essential concepts useful in the following.

The simplex is the sample space of random compositions with D parts. It is denoted as

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, \dots, x_D), x_i > 0, i = 1, 2, \dots, D, \sum_{i=1}^D x_i = \kappa \right\},$$

where κ is a constant, usually taken to be 1 or 100.

The simplex $\mathcal{S}^D$ has a $(D - 1)$ -dimensional real Euclidean vector space structure with the following operations (Pawlowsky-Glahn et al. 2015):

- Perturbation: $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, \dots, x_D y_D)$,
- Powering: $\alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, \dots, x_D^\alpha)$,
- Inner product: $\langle \mathbf{x}, \mathbf{y} \rangle_a = \sum_{i=1}^D \ln \frac{x_i}{g_m(\mathbf{x})} \ln \frac{y_i}{g_m(\mathbf{y})}$,

for $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$, a scalar $\alpha \in \mathbb{R}$, $g_m(\mathbf{x})$ the geometric mean of the parts of $\mathbf{x}$, and $\mathcal{C}$ the *closure* operator which normalizes any vector to the constant sum κ .

This algebraic-geometric structure of the sample space assures existence of a basis and the corresponding coordinate representation of compositions. These coordinates are real vectors that obey the standard Euclidean geometry. Consequently, once a basis has been chosen, all standard geometric operations, or statistical methods, can be applied to compositional data in coordinates and transferred to the simplex by preserving their properties. This is known as the *principle of working on coordinates* (see the ▶ “Compositional Data” entry). Some frequently used coordinate representations are:

- $\text{alr}(\mathbf{x}) = \left(\ln \frac{x_1}{x_D}, \dots, \ln \frac{x_{D-1}}{x_D} \right) \in \mathbb{R}^{D-1}$,

- $\text{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{g_m(\mathbf{x})}, \dots, \ln \frac{x_D}{g_m(\mathbf{x})} \right) \in \mathbb{R}^D$, where $\sum_{i=1}^D \ln \frac{x_i}{g_m(\mathbf{x})} = 0$,
- $h(\mathbf{x}) = (\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, \dots, \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a) \in \mathbb{R}^{D-1}$, where $\{\mathbf{e}_1, \dots, \mathbf{e}_{D-1}\}$ is an orthonormal basis on $\mathcal{S}^D$.

Note that vectors $\text{alr}(\mathbf{x})$ and $\text{clr}(\mathbf{x})$, where the latter stands for centered logratio (clr) coefficients, are non-orthogonal logratio representations. In particular the clr representation has one extra dimension, which makes it a constrained vector. Vector $h(\mathbf{x})$ is an orthonormal logratio representation. It contains the coordinates of $\mathbf{x}$ with respect to the orthonormal basis $\{\mathbf{e}_1, \dots, \mathbf{e}_{D-1}\}$. Egozcue et al. (2003) introduced it using a particular orthonormal basis naming it *isometric logratio* (ilr($\mathbf{x}$)) representation. Quite recently, Martín-Fernández (2019) renamed it as *orthogonal logratio* (olr($\mathbf{x}$)) representation to emphasize this intrinsic property. There are some popular possibilities how to build such an orthonormal basis (Egozcue and Pawlowsky-Glahn 2005; Filzmoser et al. 2018). In order to avoid confusion, here we will denote a generic orthonormal representation as $h(\mathbf{x})$. Also, given a real vector of olr coordinates $\mathbf{y}$ we will denote the corresponding composition as $h^{-1}(\mathbf{y})$. The different logratio coordinate representations can be linked through the change of basis matrix.

We briefly review the concept of *subcomposition*, used in the remainder of the chapter. Given a random composition $\mathbf{X} \in \mathcal{S}^D$, a C -part subcomposition is a subvector in $\mathcal{S}^C$ formed only by C parts ($C < D$). A subcomposition can be obtained as $\mathcal{C}(\mathbf{S}\mathbf{X})$ where $\mathbf{S}$ is a (C, D) -matrix with C elements equal 1 (one in each row and at most one in each column) and the remaining elements equal 0.

Furthermore, to define probability laws it is important to establish a measure with respect to which our density function is expressed. Traditionally, a Lebesgue measure λ has been used in the simplex, but the *Aitchison measure* λ_a is preferred in the context of the Aitchison geometry (Mateu-Figueras and Pawlowsky-Glahn 2008). It could be proved that λ_a is absolutely continuous with respect to λ . The relationship between them is described by the Jacobian

$$\frac{d\lambda_a}{d\lambda} = \frac{1}{\sqrt{D} \prod_{i=1}^D x_i}, \quad (1)$$

and allows to express density functions on $\mathcal{S}^D$ with respect to either λ or λ_a .

The Logistic Normal Model

Definition

The additive logistic normal (aln) model is defined by Aitchison (1982) using the alr coordinates. Indeed, a D -part random composition $\mathbf{X}$ is said to have an aln distribution when its additive logratio representation, $\text{alr}(\mathbf{X})$, has a $(D - 1)$ -dimensional normal distribution in the space of

coordinates. The corresponding density function on $\mathcal{S}^D$ with respect to the λ measure is

$$f(\mathbf{x}) = \frac{(2\pi)^{-(D-1)/2} |\Sigma|^{-1/2}}{\prod_{i=1}^D x_i} \exp\left(-\frac{1}{2} (\text{alr}(\mathbf{x}) - \boldsymbol{\mu})' \Sigma^{-1} (\text{alr}(\mathbf{x}) - \boldsymbol{\mu})\right). \quad (2)$$

This distribution will be denoted as $\mathbf{X} \sim \mathcal{N}_{\mathcal{S}}^D(\boldsymbol{\mu}, \Sigma)$ (Mateu-Figueras et al. 2013). Note that the total number of parameters of an aln distribution is $(D-1)(D+2)/2$.

We could express density (2) in terms of any orthonormal logratio representation $h(\mathbf{X})$. We have to use only the matrix relationship between the two logratio representations $h(\mathbf{X}) = \mathbf{A} \text{alr}(\mathbf{X})$ (see the ► “Additive Logistic Skew-Normal Distribution” entry for the matrix $\mathbf{A}$) and the linear transformation property of the normal distribution. The resulting density is defined on the simplex and also expressed with respect to the Lebesgue measure λ :

$$f(\mathbf{x}) = \frac{(2\pi)^{-(D-1)/2} |\mathbf{Y}|^{-1/2}}{\sqrt{D} \prod_{i=1}^D x_i} \exp\left(-\frac{1}{2} (h(\mathbf{x}) - \boldsymbol{\xi})' \mathbf{Y}^{-1} (h(\mathbf{x}) - \boldsymbol{\xi})\right), \quad (3)$$

where

$$\boldsymbol{\xi} = \mathbf{A}\boldsymbol{\mu}, \quad \mathbf{Y} = \mathbf{A}\Sigma\mathbf{A}'. \quad (4)$$

A similar strategy could be applied to obtain the density in terms of the clr representation, but it has a limited use in practice as a degenerate density is obtained due to the zero-sum constraint of the vector $\text{clr}(\mathbf{X})$.

Both densities (2) and (3) define exactly the same probability law on $\mathcal{S}^D$ and are defined with respect to the Lebesgue

measure λ . Using the relationship (1), it is easy to express it with respect to the Aitchison measure λ_a as

$$f_a(\mathbf{x}) = (2\pi)^{-(D-1)/2} |\mathbf{Y}|^{-1/2} \exp\left(-\frac{1}{2} (h(\mathbf{x}) - \boldsymbol{\xi})' \mathbf{Y}^{-1} (h(\mathbf{x}) - \boldsymbol{\xi})\right); \quad (5)$$

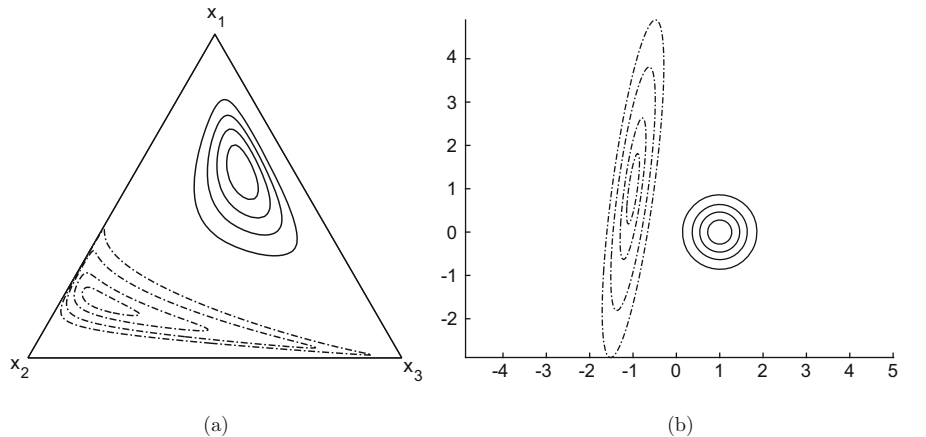
here the subindex a is only used to remind that the density is expressed with respect to λ_a . Mateu-Figueras et al. (2013) define the logistic normal model using density (5) and call it the *normal distribution on the simplex*. Even though this change of measure gives the same law of probability, it produces changes in some characteristic values of the distribution. An interested reader can find a detailed explanation in Mateu-Figueras et al. (2013). We advise that if we want to study properties of the distribution directly on the simplex, it is better to use the density with respect to the Aitchison measure as it is compatible with the simplex space structure. Moreover, the Aitchison measure enables further generalization of the logistic normal distribution, if parts of the random composition are weighted, for example, according to measurement precision (Egozcue and Pawlowsky-Glahn 2016; Hron et al. 2022). However, for computation of probability of an event we have to use the density with respect to the Lebesgue measure in order to compute the integrals.

In Fig. 1 we represent the isodensity curves of two logistic normal densities on (a) $\mathcal{S}^3$ with respect to λ_a and on (b) $\mathbb{R}^2$ using the orthonormal logratio representation with respect to the particular orthonormal basis defined in Egozcue et al. (2003). Note that using the coordinate representation, the typical circumferences and ellipses from the multivariate normal in real space are obtained as isodensity curves. The isodensity curves on the simplex are the corresponding logratio circumferences and logratio ellipses.

Additive Logistic Normal

Distribution, Fig. 1 Isodensity curves of two logistic normal models in (a) on $\mathcal{S}^3$ and (b) on $\mathbb{R}^2$ using an orthonormal logratio representation. The solid line stands for $\boldsymbol{\xi} = (1, 0)$, $\mathbf{Y} = \mathbf{I}$ and the dash-dotted line for

$$\boldsymbol{\xi} = (-1, 1), \mathbf{Y} = \begin{pmatrix} 0.2 & 0.8 \\ 0.8 & 6 \end{pmatrix}$$



Key Properties

Real random vectors obtained as linear transformation of a multivariate normal distribution on real space are still multivariate normal distributed. This is the well-known linear transformation property of the multivariate normal. Applying this property to the logratio vectors, we can easily prove the closure under perturbation, power transformation, and subcomposition of the logistic normal family of distributions. We list here the main properties using the density in terms of the logratio representation $h(\mathbf{x})$ with respect to a generic orthonormal basis, but the density in terms of the $\text{alr}(\mathbf{x})$ could also be used. An interested reader can find a detailed proof of each property in Mateu-Figueras et al. (2013).

Property 1 (Closure under $\oplus$ and $\odot$) Let $\mathbf{X} \sim \mathcal{N}_{\mathcal{S}}^D(\xi, \mathbf{Y})$ be a random composition, $\mathbf{a} \in \mathcal{S}^D$, and $b \in \mathbb{R}$. Then, the D -part random composition $\mathbf{X}^* = \mathbf{a} \oplus (b \odot \mathbf{X}) \sim \mathcal{N}_{\mathcal{S}}^D(h(\mathbf{a}) + b\xi, b^2\mathbf{Y})$.

From this property, and using the density function with respect to the Aitchison measure λ_a , we obtain the invariance under perturbation operation, that is, $f_a(\mathbf{a} \oplus \mathbf{x}) = f_a(\mathbf{x})$. This has important consequences, because when working with compositional data, it is usual to center the input data set which could be expressed as perturbation by a compositional vector (center of the distribution). However, this property would not be obtained using the density with respect to the measure λ . Accordingly, although both densities define the same probability law, this is one of differences between them.

Property 2 (Closure under subcomposition) Let $\mathbf{X} \sim \mathcal{N}_{\mathcal{S}}^D(\xi, \mathbf{Y})$ and $\mathbf{X}_S = \mathcal{C}(\mathbf{S}\mathbf{X})$ be a C -part random subcomposition. Then $\mathbf{X}_S \sim \mathcal{N}_{\mathcal{S}}^D(\xi_S, \mathbf{Y}_S)$ with

$$\xi_S = \mathbf{U}^* \mathbf{S} \mathbf{U} \xi, \quad \mathbf{Y}_S = (\mathbf{U}^* \mathbf{S} \mathbf{U}) \mathbf{Y} (\mathbf{U}^* \mathbf{S} \mathbf{U})',$$

where the columns of matrices $\mathbf{U}$ and $\mathbf{U}^*$ contain the clr representation of the corresponding orthonormal basis in $\mathcal{S}^D$ and $\mathcal{S}^C$.

The center of a random composition $\mathbf{X}$ plays the role of the expected value for a real random vector (Pawlowsky-Glahn et al. 2015). Mateu-Figueras et al. (2013) show that this center can be expressed as the expected value using the density function with respect to λ_a ; for this reason it is denoted as $E_a(\mathbf{X})$.

Property 3 (Location) Let $\mathbf{X} \sim \mathcal{N}_{\mathcal{S}}^D(\xi, \mathbf{Y})$, then mode and expected value with respect to the measure λ_a coincide and are $\text{mode}_a(\mathbf{X}) = E_a(\mathbf{X}) = h^{-1}(\xi)$ independently of the chosen orthonormal coordinates $h(\mathbf{x})$.

The expected value with respect to the Lebesgue measure corresponds to the standard expected value for real random vectors. Although it exists, there is no closed form for this

value because the corresponding integral is not reducible to any simple form, and numerical integration should be used. Nevertheless, for random compositions, the use of $E_a(\mathbf{X})$ is recommended with the same arguments as, when working with a compositional data set, the center, that is the sample geometric mean, is used instead of the arithmetic mean.

Variability of random compositions can be measured with the metric variance, also called the total variance (Aitchison 1986).

Property 4 (Metric variance) Let $\mathbf{X} \sim \mathcal{N}_{\mathcal{S}}^D(\xi, \mathbf{Y})$, then $\text{Mvar}(\mathbf{X}) = \text{trace}(\mathbf{Y})$.

Given a compositional data set, the parameters of the logistic normal model, that is, the vector of expected values ξ and the covariance matrix $\mathbf{Y}$, may be estimated using the maximum likelihood method based on the logratio representation of the data set. If we change the logratio representation, the estimates will be related through the corresponding transformation matrix $\mathbf{A}$ as done in (4), the only deviations will be due to the numerical nature of the procedure. Thus, the estimates are invariant under the choice of the logratio representation.

Finally, given a compositional data set, we can validate the assumption of the logistic normality via testing for multivariate normality of the olr coordinate representation using a proper goodness-of-fit test. In this case, any univariate normality tests applied to each marginal are not invariant under the choice of the logratio representation, but the χ^2 goodness-of-fit test applied to the Mahalanobis distances or the multivariate Shapiro-Wilk test of normality is invariant.

Applications to Geosciences

In this section, by means of some compositional data analysis, we briefly show the central role of the logistic normal distribution in geosciences.

As the normal distribution in real space, the logistic normal distribution is used in many statistical compositional applications and represents an essential distribution for statistical inference on the simplex, such as regression analysis, hypotheses testing, and construction of confidence intervals.

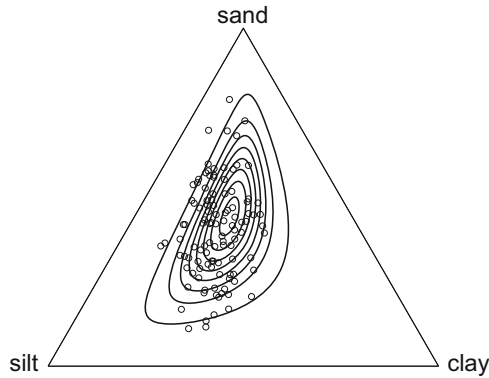
We consider, as a first example, the GEMAS data set (Reimann et al. 2014a, b), a geochemical data set on agricultural and grazing land soil, collected as GEMAS data in the R-package robCompositions (Templ et al. 2011), to show the fit of the logratio normal distribution. Here the complementary particle size distributions are considered, that is, the three-part compositions $\mathbf{x} = (\text{sand}, \text{silt}, \text{clay})$. For this example only samples from Italy are used and filtered from outliers (Filzmoser and Hron 2008), resulting in $n = 109$ soil samples. Using the orthonormal logratio representation called pivot coordinates

$h(\mathbf{x}) = \left(\sqrt{2/3} \ln(\text{sand}/(\text{silt} \cdot \text{clay})^2), \sqrt{1/2} \ln(\text{silt}/\text{clay}) \right)$ (Filzmoser et al. 2018), the following parameter estimates are obtained: $\hat{\xi} = (0.245, 0.346)$ and $\hat{\Sigma} = \begin{pmatrix} 0.285 & 0.031 \\ 0.031 & 0.084 \end{pmatrix}$. Goodness-of-

fit tests applied to each marginal show no significant departures from normality (p -values > 0.1 for the Anderson-Darling and Shapiro-Wilk test). The Anderson-Darling test applied to the Mahalanobis distances to check the χ^2_2 assumption shows no significant departure (p -value > 0.1). The fitted isodensity curves on $\mathcal{S}^3$ are displayed in Fig. 2.

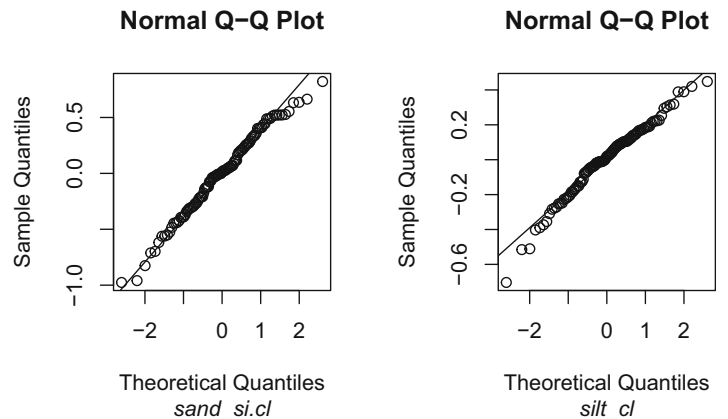
For the multivariate analysis of variance (MANOVA) to test whether significant differences exist in the three-part compositions of particle size distributions among soil groups, the multivariate logistic normality is one of the fundamental assumptions. In other words, we assume that the dependent variables, that is, the (sand, silt, clay) compositions follow the multivariate logistic normality within groups defined by the soil classes. The Shapiro-Wilk test for multivariate normality applied to the two pivot coordinates shows no significant strong departures from the null hypothesis (p -values $= 0.1$).

Figure 3 reports Normal Q-Q plots of the residuals of the MANOVA for the two pivot coordinates, which confirm that a



Additive Logistic Normal Distribution, Fig. 2 Fitted isodensity curves for the logistic normal model to the composition (sand, silt, clay) of the GEMAS data set. Samples from Italy were taken and filtered from outliers ($n = 109$)

Additive Logistic Normal Distribution, Fig. 3 Q-Q plots of the residuals of the MANOVA (GEMAS data set)



normal distribution is reasonable in both cases. Finally, the MANOVA test suggests significant differences between the soil groups (p -value < 0.001 for the Pillai test). Note that we would obtain the same inference in multivariate linear regression, using the soil classes as categorical covariate.

Similar to the MANOVA, also Linear Discriminant Analysis (LDA) for compositions, principal component analysis, or canonical correlation are other techniques which use the assumption of logistic normality to represent group distributions.

For demonstration purposes, we use a second example related to whole-rock composition of lavas from the Atacazo and Ninahuilca volcanoes, Ecuador (Hidalgo et al. 2008). In particular, we look at the major elements (sub)composition with five parts: Fe_2O_3 , MgO , K_2O , La , and Yb . The logistic normality assumption is confirmed by the univariate p -values of the Shapiro-Wilk tests applied to each pivot coordinate listed in Table 1; all reported p -values are greater than 0.05 within each level of the grouping variable, related to different volcanoes. Moreover, the last row of Table 1 reports results of the Shapiro-Wilk test for multivariate normality within the two groups, which again are compatible with the assumption.

The purpose of LDA in this example is to find the linear combinations of the original variables, namely the four pivot coordinates, that gives the best possible separation between groups in the data set. The LDA results offer a satisfactory separation of the two groups; in fact the model accuracy is equal to 0.945. The discriminant scores themselves are approximately normally distributed within groups as we can see from Fig. 4.

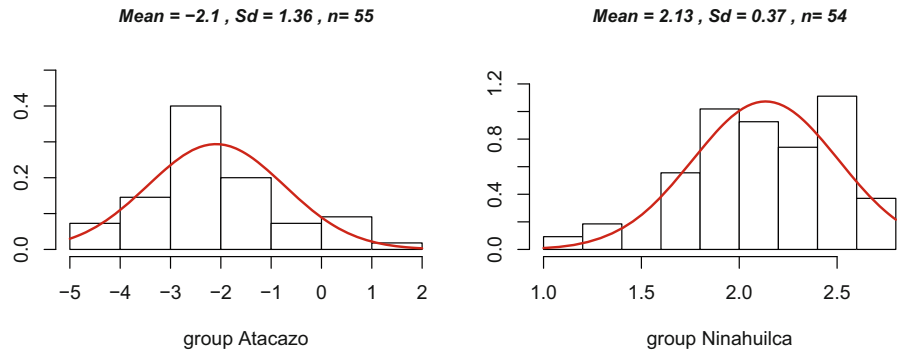
Summary

In this chapter the logistic normal model, a follower of the additive logistic normal distribution compatible with the algebraic-geometric structure of the simplex, was presented as a general tool for distributional modeling of compositional data in geosciences. It is flexible enough to model data

Additive Logistic Normal Distribution, Table 1 Results of univariate, and multivariate (last row), Shapiro-Wilk normality test (test statistics *W* and *p*-values) within groups

Pivot coordinates	Group Atacazo		Group Ninahuilca	
	<i>W</i>	<i>p</i> -value	<i>W</i>	<i>p</i> -value
$\sqrt{4/5} \ln (\text{Fe}_2\text{O}_3/\text{MgO} \cdot \text{K}_2\text{O} \cdot \text{La} \cdot \text{Yb})^4)$	0.958	0.051	0.975	0.309
$\sqrt{3/4} \ln (\text{MgO}/\text{K}_2\text{O} \cdot \text{La} \cdot \text{Yb})^3)$	0.967	0.130	0.985	0.739
$\sqrt{2/3} \ln (\text{K}_2\text{O}/(\text{La} \cdot \text{Yb})^2)$	0.98	0.482	0.968	0.162
$\sqrt{1/2} \ln (\text{La}/\text{Yb})$	0.954	0.033	0.984	0.684
Multivariate test	0.762	<0.001	0.962	0.085

Additive Logistic Normal Distribution, Fig. 4 Histograms of discriminant scores with superimposed normal curve, whose parameters are estimated from the respective scores means and standard deviations



coming from a wide-ranging applications, providing a sensible alternative to the well-known Dirichlet distribution which is not fully compatible with the logratio approach (Pawlowsky-Glahn et al. 2015). For all these reasons, logistic normal is considered the main distribution for compositional data.

Cross-References

- Additive Logistic Skew-Normal Distribution
- Compositional Data

Bibliography

Aitchison J (1982) The statistical analysis of compositional data (with discussion). *J R Stat Soc Ser B* 44:139–177

Aitchison J (1986) The statistical analysis of compositional data. Monographs on statistics and applied probability. Chapman & Hall Ltd, London. (Reprinted in 2003 with additional material by The Blackburn Press), 416 p

Comas-Cufí M, Martín-Fernández JA, Mateu-Figueras G (2016) Logratio methods in mixture models for compositional data sets. *SORT* 40:349–374

Egozcue JJ, Pawlowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis. *Math Geol* 37:795–828

Egozcue JJ, Pawlowsky-Glahn V (2016) Changing the reference measure in the simplex and its weighting effects. *Aust J Stat* 45(4):25–44

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. *Math Geol* 35:279–300

Filzmoser P, Hron K (2008) Outlier detection for compositional data using robust methods. *Math Geosci* 40(3):233–248

Filzmoser P, Hron K, Templ M (2018) Applied compositional data analysis. Springer, Cham

Hidalgo S, Monzier M, Almeida E, Chazot G, Eissen JP, van der Plicht J, Hall M (2008) Late Pleistocene and Holocene activity of the Atacazo-Ninahuilca Volcanic Complex (Ecuador). *J Volc Geoth Res* 176:16–26

Hron K, Menafoglio A, Palarea-Albaladejo J, Filzmoser P, Talská R, Egozcue JJ (2022) Weighting of parts in compositional data analysis: advances and applications. *Math Geosci* 54:71–93

Martín-Fernández JA (2019) Comments on: Compositional data: the sample space and its structure, by Egozcue and Pawlowsky-Glahn. *TEST* 28:653–657

Mateu-Figueras G, Pawlowsky-Glahn V (2008) A critical approach to probability laws in geochemistry. *Math Geosci* 40(5):489–502

Mateu-Figueras G, Pawlowsky-Glahn V, Egozcue JJ (2013) The normal distribution in some constrained sample spaces. *SORT* 37:29–56

Pawlowsky-Glahn V, Egozcue J, Tolosana-Delgado R (2015) Modeling and analysis of compositional data. Wiley, Chichester

Reimann C, Birke M, Demetriades A, Filzmoser P, O’Connor P (eds) (2014a) Chemistry of Europe’s agricultural soils, Part A. Schweizerbart Science Publishers, Stuttgart

Reimann C, Birke M, Demetriades A, Filzmoser P, O’Connor P (eds) (2014b) Chemistry of Europe’s agricultural soils, Part B. Schweizerbart Science Publishers, Stuttgart

Templ M, Hron K, Filzmoser P (2011) robCompositions: an R-package for robust statistical analysis of compositional data. In: Pawlowsky-Glahn V, Buccianti A (eds) Compositional data analysis: theory and applications. Wiley, Chichester, pp 341–355

Additive Logistic Skew-Normal Distribution

Glòria Mateu-Figueras^{1,2} and Karel Hron³

¹Department of Informatics, Applied Mathematics and Statistics, University of Girona, Girona, Spain

²Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Girona, Spain

³Department of Mathematical Analysis and Applications of Mathematics, Palacký University, Olomouc, Czech Republic

Synonyms

Logistic skew-normal; Logratio skew-normal; Skew-normal on the simplex

Definition

The *additive logistic skew-normal* model is a family of distributions of compositional data defined on the simplex using the logratio approach (Aitchison 1986; Pawłowsky-Glahn et al. 2015) and the multivariate skew-normal distribution for real random vectors (Azzalini and Capitanio 1999). For the construction of the distribution standard strategy of the logratio methodology using the principle of working on coordinates is applied, specifically, the multivariate skew-normal distribution in the real space is used to model the additive logratio representation of the random composition. This model represents a generalization of the additive logistic normal model because it is contained as a particular case. This distribution allows us to model compositional data sets when their logratio representation displays low or moderate level of skewness and as such, it is appropriate for modeling of compositional data in geosciences.

Originally, the model was defined using the additive logratio representation, hence the name additive logistic skew-normal was assigned. However, other logratio representations such as the orthonormal logratio (olr) representation (Martín-Fernández 2019), originally known as isometric logratio (ilr) representation (Egozcue et al. 2003), can also be used. For this reason, the term logistic skew-normal or logratio skew-normal can be used to refer this distribution model.

Sample Space of Compositional Data and Its Geometry

The sample space of random compositions, where the logistic skew-normal model is defined, is called the simplex and is denoted as

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, \dots, x_D), x_i > 0, \sum_{i=1}^D x_i = \kappa \right\},$$

where κ is a constant, usually taken to be one (proportional representation of compositional data) or one hundred (representation in percentages). Below the algebraic geometric structure of the simplex, contained also in the entry ▶ “Compositional Data”, is briefly recalled for the purpose of this chapter.

The simplex $\mathcal{S}^D$ has a $(D - 1)$ -dimensional real Euclidean vector space structure with the following operations (Pawłowsky-Glahn et al. 2015):

- Perturbation: $\mathbf{x} \oplus \mathbf{y} = C(x_1 y_1, \dots, x_D y_D)$,
- Powering: $\alpha \odot \mathbf{x} = C(x_1^\alpha, \dots, x_D^\alpha)$,

for $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ and a scalar $\alpha \in \mathbb{R}$. The operator C is the *closure* and normalizes the compositional vector in the argument by dividing each component by the sum of all components and multiplying them by the constant κ . Also, an inner product can be defined, that induces a norm and a distance (see the ▶ “Compositional Data” entry for details).

This algebraic-geometric structure of the sample space assures existence of a basis and the corresponding coordinate representation of compositions. These coordinates are real vectors that obey the standard Euclidean geometry, with the sum of two vectors and the multiplication of a vector by a scalar instead of perturbation and powering. Consequently, once a basis has been chosen, all standard geometric operations or statistical methods can be applied to coordinates and transferred to the simplex by preserving their properties. This is known as the *Principle of working on coordinates* (see the ▶ “Compositional Data” entry).

Some frequently used coordinate representations defined by Aitchison (1986) involve logratios:

$$\text{alr}(\mathbf{x}) = \left(\ln \frac{x_1}{x_D}, \dots, \ln \frac{x_{D-1}}{x_D} \right)$$

$$\text{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{g_m(\mathbf{x})}, \dots, \ln \frac{x_D}{g_m(\mathbf{x})} \right)$$

The *additive logratio* (alr) representation is a $(D - 1)$ -dimensional vector which corresponds to coordinates with respect to an oblique basis. Consequently, the geometric operations are preserved as $\text{alr}(\mathbf{x} \oplus \mathbf{y}) = \text{alr}(\mathbf{x}) + \text{alr}(\mathbf{y})$ and $\text{alr}(\alpha \odot \mathbf{x}) = \alpha \cdot \text{alr}(\mathbf{x})$, but distances and orthogonal projections computed using the alr coordinates will not be the same as computed with the original compositions and using the Aitchison distance or inner product. This is the main drawback of the additive logratio coordinates. The *centered logratio* (clr) representation is a D -dimensional vector obtained as coefficients with respect to a generating system,

not coordinates of a basis. We have one extra dimension and for this reason the clr vector is constrained: the sum of its components equals to zero. Although this representation is widely used in the literature because geometric operations, distances and inner product are preserved, its use to define families of distributions is very limited because degenerate densities are obtained.

We can also consider coordinates with respect to an orthonormal basis, initially called as *isometric logratio* (ilr) representation, but recently renamed to *orthogonal logratio* (olr) representation which better reflects its intrinsic properties. Egozcue et al. (2003) proposed a particular orthonormal basis for construction of these coordinates, but there are also other popular possibilities how to build such a basis. One of them, explained in the ► “Compositional Data” entry, is using a sequential binary partition of a generic composition and called balances (Egozcue and Pawłowsky-Glahn 2005). Some other olr coordinate systems, called pivot coordinates, highlight the role of single compositional parts and were proposed in Filzmoser et al. (2018). In order to avoid confusion, here we will denote a generic orthonormal representation as $h(\mathbf{x})$. Also, given a real vector of olr coordinates $\mathbf{y}$ we will denote the corresponding composition as $h^{-1}(\mathbf{y})$.

Different logratio coordinate representations can be linked through a matrix which stands for the change of the basis. For example, we can construct a $(D-1) \times (D-1)$ matrix $\mathbf{A}$ which relates the alr representation with a particular orthonormal representation $h(\mathbf{x})$ as

$$h(\mathbf{x}) = \mathbf{A} \text{alr}(\mathbf{x}) \text{ or } \text{alr}(\mathbf{x}) = \mathbf{A}^{-1}h(\mathbf{x}) \quad (1)$$

where $\mathbf{A} = \mathbf{U}'\mathbf{F}^*$, $\mathbf{U}$ is a $(D, D-1)$ -matrix which contains in columns the clr representation of the orthonormal basis and $\mathbf{F}^*$ is the $(D, D-1)$ -matrix

$$\mathbf{F}^* = \frac{1}{D} \begin{pmatrix} D-1 & -1 & \dots & -1 \\ -1 & D-1 & \dots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \dots & D-1 \\ -1 & -1 & \dots & -1 \end{pmatrix}$$

Given a D -part random composition $\mathbf{X}$, we can be interested in the composition formed only by C -parts ($C < D$). This can be regarded as a projection on a simplex with fewer parts and, accordingly, a random *subcomposition* is obtained. Formally, the formation of a subcomposition can be achieved as $C(\mathbf{S}\mathbf{X})$ where $\mathbf{S}$ is a (C, D) -matrix with C elements equal 1 (one in each row and at most one in each column) and the remaining elements equal 0.

To define probability laws it is important to establish a measure with respect to which our density function is expressed. Traditionally, a Lebesgue measure λ has been

used in the simplex but the alternative measure called *Aitchison measure* and denoted as λ_a is preferred in context of the Aitchison geometry (Mateu-Figueras and Pawłowsky-Glahn 2008). This measure is defined using the Lebesgue measure in the space of olr coordinates and it is compatible with the vector space structure defined above. It could be proved that the Aitchison measure λ_a is absolutely continuous with respect to the Lebesgue measure λ . The relationship between them is

$$\frac{d\lambda_a}{d\lambda} = \frac{1}{\sqrt{D}\prod_{i=1}^D x_i}. \quad (2)$$

This relationship allows to express density functions on $\mathcal{S}^D$ with respect to λ or λ_a . Specifically, the chain rule for measures relates two probability densities, $dP/d\lambda_a$ and $dP/d\lambda$, as $dP/d\lambda = (dP/d\lambda_a)(d\lambda_a/d\lambda)$.

The Multivariate Skew-Normal Model

According to the definition given by Azzalini and Capitanio (1999), a $(D-1)$ -variate random vector $\mathbf{Y}$ follows a multivariate *skew-normal* (sn) distribution on $\mathbb{R}^{D-1}$ with parameters $\boldsymbol{\mu}$, Σ , and $\boldsymbol{\alpha}$, if its density function is given by

$$f(\mathbf{y}) = 2(2\pi)^{-(D-1)/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})' \Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu})\right) \Phi(\boldsymbol{\alpha}' \boldsymbol{\omega}^{-1}(\mathbf{y} - \boldsymbol{\mu})), \quad (3)$$

where Φ is the standard normal distribution function and $\boldsymbol{\omega}$ is the square root of the diagonal matrix formed from Σ . The $(D-1)$ -variate vector $\boldsymbol{\alpha}$ regulates the shape of the density and indicates the direction of maximum skewness. The sn family is only able to model moderate levels of skewness as the skewness coefficient of each univariate sn marginal takes values in the interval $(-0.995, 0.995)$. When $\boldsymbol{\alpha} = 0$, the multivariate normal density with parameters $\boldsymbol{\mu}$ and Σ is obtained. In the following the notation $\mathbf{Y} \sim \mathcal{SN}^{D-1}(\boldsymbol{\mu}, \Sigma, \boldsymbol{\alpha})$ is used.

The moment generating function provided in Azzalini and Capitanio (1999) is

$$M\mathbf{Y}(t) = 2 \exp(t' \boldsymbol{\mu} + (1/2)t' \Sigma t) \Phi(\boldsymbol{\delta}' \boldsymbol{\omega} t), \quad (4)$$

where

$$\boldsymbol{\delta} = \frac{\boldsymbol{\omega}^{-1} \Sigma \boldsymbol{\omega}^{-1} \boldsymbol{\alpha}}{\sqrt{1 + \boldsymbol{\alpha}' \boldsymbol{\omega}^{-1} \Sigma \boldsymbol{\omega}^{-1} \boldsymbol{\alpha}}} \quad (5)$$

The expected value and the covariance matrix derived from (4) are

$$E(\mathbf{Y}) = \boldsymbol{\mu} + \boldsymbol{\omega} \sqrt{2/\pi} \boldsymbol{\delta}, \quad \text{Var}(\mathbf{Y}) = \boldsymbol{\Sigma} - (2/\pi) \boldsymbol{\omega} \boldsymbol{\delta} \boldsymbol{\delta}' \boldsymbol{\omega}$$

The appealing properties of this distribution are listed by Azzalini and Capitanio (1999). Most of them are analogues of the multivariate normal properties, in particular, random vectors obtained as linear transformation of the form $\mathbf{Y}^* = \mathbf{A}\mathbf{Y}$ for any real matrix $\mathbf{A}$, are still multivariate sn distributed. Another interesting property in connection with the multivariate normal random compositions is that the random variable

$$Z = (\mathbf{Y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \boldsymbol{\mu}), \quad (6)$$

follows a χ^2 distribution with $D - 1$ degrees of freedom.

Given a random sample, the maximum likelihood (ML) method is proposed to estimate the parameters but numerical procedures have to be used. The method of moments is proposed to obtain starting values for the iterative procedure. The R-package “sn” is available and contains a suite of functions for handling univariate and multivariate sn distributions (Azzalini 2020).

The Logistic Skew-Normal Model

Definition

By analogy with the additive logistic normal (aln) model defined by Aitchison (1986) using the alr coordinates, Mateu-Figueras et al. (2005) use the same strategy to define the additive logistic skew-normal (alsn) model. Indeed, a D -part random composition $\mathbf{X}$ is said to have an alsn distribution when its additive logratio representation, $\text{alr}(\mathbf{X})$, has a $(D - 1)$ -dimensional sn distribution in the space of coordinates. The corresponding density function on S^D with respect to the Lebesgue measure λ is

$$f(\mathbf{x}) = \frac{2(2\pi)^{-(D-1)/2} |\boldsymbol{\Sigma}|^{-1/2}}{\prod_{i=1}^D x_i} \exp\left(-\frac{1}{2} (\text{alr}(\mathbf{x}) - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\text{alr}(\mathbf{x}) - \boldsymbol{\mu})\right) \Phi(\boldsymbol{\alpha}' \boldsymbol{\omega}^{-1} (\text{alr}(\mathbf{x}) - \boldsymbol{\mu})). \quad (7)$$

This distribution will be denoted as $\mathbf{X} \sim \mathcal{SN}_S^D(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$. A particular member of this class of distributions is determined by $(D - 1)(D + 4)/2$ parameters, $D - 1$ more than the aln class due to the $\boldsymbol{\alpha}$ vector. The aln model is obtained as a particular case when $\boldsymbol{\alpha} = \mathbf{0}$. For this reason, we can say that the alsn model is a generalization of the aln model.

We could express the density function in terms of any orthonormal logratio representation. We have only to use the matrix relationship between $h(\mathbf{X})$ and $\text{alr}(\mathbf{X})$ coordinates (1) and the linear transformation property of the sn

distribution. Accordingly, for $h(\mathbf{X}) = \mathbf{A} \text{alr}(\mathbf{X})$ we can derive the previous density function in terms of $h(\mathbf{X})$. It is defined on the simplex and also expressed with respect to the Lebesgue measure λ ,

$$f(\mathbf{x}) = \frac{2(2\pi)^{-(D-1)/2} |\mathbf{Y}|^{-1/2}}{\sqrt{D} \prod_{i=1}^D x_i} \exp\left(-\frac{1}{2} (h(\mathbf{x}) - \boldsymbol{\xi})' \mathbf{Y}^{-1} (h(\mathbf{x}) - \boldsymbol{\xi})\right) \Phi(\boldsymbol{\varrho}' \mathbf{v}^{-1} (h(\mathbf{x}) - \boldsymbol{\xi})), \quad (8)$$

where $\mathbf{v}$ stands for the square root of the diagonal matrix formed from $\mathbf{Y}$. The relationship between the parameters of both densities is

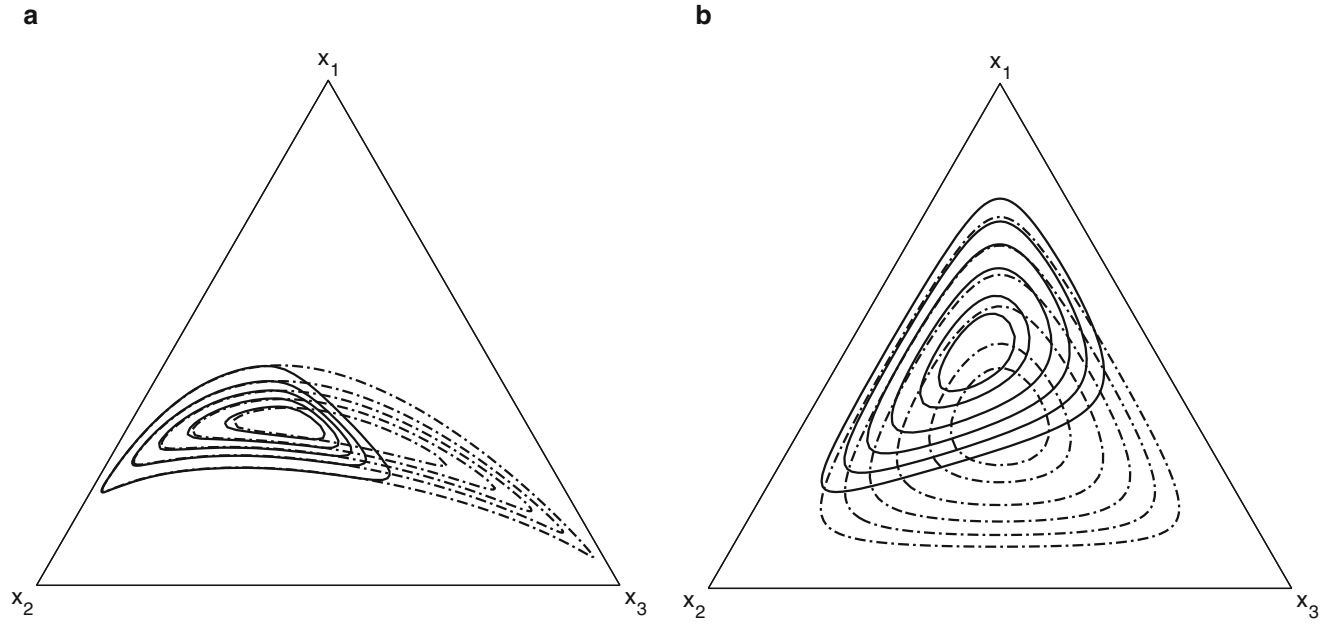
$$\boldsymbol{\xi} = \mathbf{A}\boldsymbol{\mu}, \quad \mathbf{Y} = \mathbf{A}\boldsymbol{\Sigma}(\mathbf{A})', \quad \boldsymbol{\varrho} = \mathbf{v}(\mathbf{A}^{-1})' \boldsymbol{\omega}^{-1} \boldsymbol{\alpha} \quad (9)$$

Both densities (7) and (8) define exactly the same probability law on S^D and are defined with respect to the Lebesgue measure λ . Using the relationship (2), it is easy to express it with respect to the Aitchison measure λ_a as

$$f_a(\mathbf{x}) = 2(2\pi)^{-(D-1)/2} |\mathbf{Y}|^{-1/2} \exp\left(-\frac{1}{2} (h(\mathbf{x}) - \boldsymbol{\xi})' \mathbf{Y}^{-1} (h(\mathbf{x}) - \boldsymbol{\xi})\right) \Phi(\boldsymbol{\varrho}' \mathbf{v}^{-1} (h(\mathbf{x}) - \boldsymbol{\xi})); \quad (10)$$

here the subindex a is only used to remind that the density is expressed with respect to λ_a . Mateu-Figueras and Pawlowsky-Glahn (2007) define the logistic skew-normal model using this density and call it the *skew-normal distribution on the simplex* by analogy to the normal on the simplex law defined using the same strategy (Mateu-Figueras et al. 2013). Even though this change of measure gives us the same law of probability, it produces changes in some characteristic values of the distribution. An interested reader can find a detailed explanation in Mateu-Figueras and Pawlowsky-Glahn (2007) and Mateu-Figueras et al. (2013). If we want to study properties of the distribution directly on the simplex, it is better to use the density with respect to the Aitchison measure because it is compatible with the simplex space structure. Moreover, the Aitchison measure enables further generalization of the logistic skew-normal distribution, if parts of the random composition are weighted, e.g., according to measurement prevision (Egozcue and Pawlowsky-Glahn 2016; Talská et al. 2020). However, for computation of probability of an event we have to use the density with respect to the Lebesgue measure in order to compute the integrals.

In Fig. 1 we present ternary diagrams with isodensity curves for two logistic skew-normal densities with respect



Additive Logistic Skew-Normal Distribution, Fig. 1 In solid line, isodensity curves of two logistic skew-normal in S^3 with: (a) $\xi = (-0.05, -0.08)$, $\Upsilon = \begin{pmatrix} 0.2 & -0.5 \\ -0.5 & 1.5 \end{pmatrix}$, $\varrho = (-5, 5)$; (b) $\xi = (1.5, 0.5)$, $\Upsilon = Id$,

$\varrho = (1.5, 2)$. In dashed-dotted line, the isodensity curves for the corresponding logistic normal model obtained with $\varrho = (0, 0)$

to the Aitchison measure. We can compare it with the corresponding logistic normal density obtained by taking $\alpha = 0$. Observe that using the logistic skew-normal models, we will be able to capture more forms of variability besides the typical arc-shape or rounded-shape forms of the logistic normal model.

Key Properties

Using the linear transformation property of the multivariate sn distribution, it is easy to prove the closure under perturbation, power transformation and subcomposition of the logistic skew-normal family of distributions. We list here the main properties using the density in terms of the logratio representation $h(\mathbf{x})$ with respect to a generic orthonormal basis, but the density in terms of the alr($\mathbf{x}$) could also be used. An interested reader can find a detailed proof of each property in Mateu-Figueras and Pawłowsky-Glahn (2007) and Mateu-Figueras et al. (2013).

Property 1 (Closure Under $\oplus$ and $\odot$). Let $\mathbf{X} \sim \mathcal{SN}_S^D(\xi, \Upsilon, \varrho)$ be a random composition, $\mathbf{a} \in S^D$ and $b \in \mathbb{R}$. Then, the D -part random composition $\mathbf{X}^* = \mathbf{a} \oplus (b \odot \mathbf{X}) \sim \mathcal{SN}_S^D(h(\mathbf{a}) + b\xi, b^2\Upsilon, \varrho)$.

From this property and using the density function with respect to the Aitchison measure λ_a we obtain the invariance under perturbation operation, i.e., $f_a(\mathbf{a} \oplus \mathbf{x}) = f_a(\mathbf{x})$. This has

important consequences, because when working with compositional data, it is usual to center the input data set which could be expressed as perturbation by a compositional vector (center of the distribution). However, this property would be not obtained using the density with respect to the measure λ . Accordingly, although both densities define the same probability law, this is one of differences between them.

Property 2 (Closure Under Subcomposition). Let $\mathbf{X} \sim \mathcal{SN}_S^D(\xi, \Upsilon, \varrho)$ and $\mathbf{X}_S = C(\mathbf{S}\mathbf{X})$ be a C -part random subcomposition. Then $\mathbf{X}_S \sim \mathcal{SN}_S^D(\xi_S, \Upsilon_S, \varrho_S)$ with

$$\begin{aligned} \xi_S &= \mathbf{U}^* \mathbf{S} \mathbf{U} \xi, & \Upsilon_S &= (\mathbf{U}^* \mathbf{S} \mathbf{U}) \Upsilon (\mathbf{U}^* \mathbf{S} \mathbf{U})', \\ \varrho_S &= \frac{v_S \Upsilon_S^{-1} \mathbf{B}' \varrho}{\sqrt{1 + \varrho' (\mathbf{v}^{-1} \Upsilon \mathbf{v}^{-1} - \mathbf{B} \Upsilon_S^{-1} \mathbf{B}') \varrho}}, \end{aligned}$$

where $\mathbf{B} = \mathbf{v}^{-1} \Upsilon (\mathbf{U}' \mathbf{S}' \mathbf{U}^*)$, v_S and $\mathbf{v}$ are the squared roots of the diagonal matrices formed by Υ_S and Υ , respectively. Matrices $\mathbf{U}$ and $\mathbf{U}^*$ contain in columns the clr representation of the corresponding orthonormal basis in S^D and S^C .

The center of a random composition $\mathbf{X}$ plays the role of the expected value for a real random vector (Pawłowsky-Glahn et al. 2015). Mateu-Figueras et al. (2013) show that this center can be expressed as the expected value using the density function with respect to λ_a ; for this reason it is denoted $E_a(\mathbf{X})$.

Property 3 (Location). Let $\mathbf{X} \sim \mathcal{SN}_S^D(\xi, \Upsilon, \varrho)$, then the expected value with respect to the measure λ_a is $E_a(\mathbf{X}) = h^{-1}(\beta)$ with $\beta = \xi + \mathbf{v}\theta\sqrt{2/\pi}$, where $\mathbf{v}$ stands for the squared root of the diagonal matrix formed by Υ and θ is as (5) but using parameters ξ , Υ and ϱ .

The expected value with respect to the Lebesgue measure corresponds to the standard expected value for real random vectors. For a logistic skew-normal random composition, there is no closed form for this value, although it exists. The reason is that the corresponding integral it is not reducible to any simple form and numerical integration should be used. Nevertheless, for random compositions, the use of $E_a(\mathbf{X})$ is recommended with the same arguments as, when working with a compositional data set, the center (the geometric mean) is used instead of the arithmetic mean.

Variability of random compositions can be measured with the metric variance, also called the total variance (Aitchison 1986).

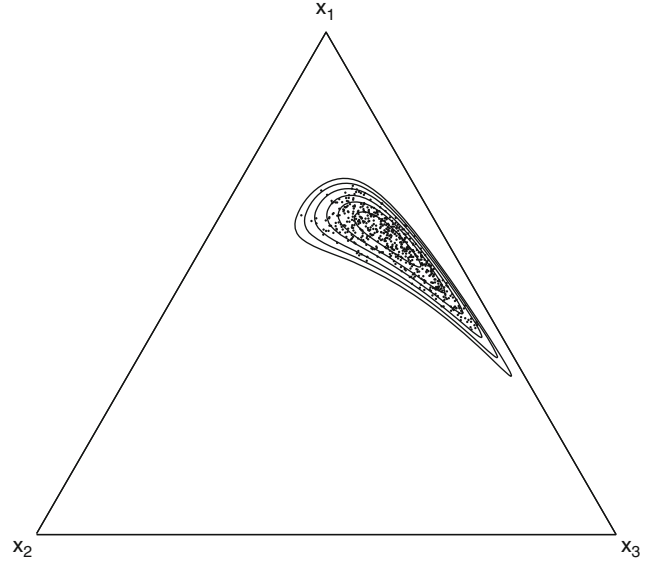
Property 4 (Metric Variance). Let $\mathbf{X} \sim \mathcal{SN}_S^D(\xi, \Upsilon, \varrho)$, then the metric variance of $\mathbf{X}$ is $\text{Mvar}(\mathbf{X}) = \text{trace}(\Upsilon - (2/\pi)\mathbf{v}\theta\theta'\mathbf{v})$, where $\mathbf{v}$ stands for the squared root of the diagonal matrix formed by Υ and θ is as (5) but using parameters ξ , Υ , and ϱ .

Given a compositional data set, the parameters of the logistic skew-normal models can be estimated using the ML method based on the logratio representation of the data set. These estimates cannot be expressed in analytic form and numerical methods have to be used to compute them. If we change the logratio representation, the estimates will be related thought the corresponding transformation matrix $\mathbf{A}$ as we do in (9), the only deviations will be due to the numerical nature of the procedure. Thus, the estimates are invariant under the choice of the logratio representation. The logistic normal model is obtained when $\varrho = \mathbf{0}$ and, consequently, a likelihood ratio test can be used to decide if a logistic skew-normal model provides a significantly better fit.

Finally, given a compositional data set, we can validate the assumption of the logistic skew-normality via testing for multivariate skew-normality of the olr coordinate representation using a proper goodness-of-fit test. One possibility is to apply a goodness-of-fit test for the underlying χ^2 distribution with $D - 1$ degrees of freedom of the random variable $\mathbf{Z}$ (6) using an olr coordinate representation and the corresponding parameter estimates.

Application to Geochemical Data

For demonstration of fitting the skew-normal distribution on the simplex to real-world geochemical data, we use the Kola data resulting from a large geochemical mapping in the Kola peninsula of Northern Europe (Reimann et al. 1998). In total,



Additive Logistic Skew-Normal Distribution, Fig. 2 Subcomposition (Al, Si, K) of the *bhorizon* data set and the fitted isodensity curves for the logistic skew-normal model

approximately 600 soil samples were taken and analyzed for more than 50 chemical elements. The three major lithogenic elements Al, Si, and K from the *bhorizon* data set (Filzmoser 2020) are used, filtered from outlying observations (Filzmoser and Hron 2008). Using the particular olr representation $(\sqrt{1/6} \ln((x_1 x_2)/x_3^2), \sqrt{1/2} \ln(x_1/x_2))$ and the Rpackage *sn*, the following estimated parameters are obtained: $\hat{\xi} = (0.127, 1.156)$, $\hat{\Upsilon} = \begin{pmatrix} 0.332 & -0.139 \\ -0.139 & 0.074 \end{pmatrix}$ and $\hat{\varrho} = (-2.658, -0.928)$. Each marginal of the two-dimensional logratio data set exhibits a slight skewness that the skew-normal model can capture. The values of the estimated parameters depend on the particular logratio representation we use but the final model on S^3 is invariant under this choice. The fitted isodensity curves on S^3 for these three-part compositions are displayed in Fig. 2.

Summary

In the chapter the logistic skew normal model, a follower of the additive logistic skew normal distribution compatible with the algebraic-geometric structure of the simplex, was presented as a general tool for distributional modeling of compositional data in geosciences. It is flexible enough to model data coming from a wide range of applications and to provide a sensible alternative to the well-known Dirichlet distribution

which is not fully compatible with the logratio approach (Pawlowsky-Glahn et al. 2015).

Cross-References

► Compositional Data

Bibliography

- Aitchison J (1986) The statistical analysis of compositional data. In: Monographs on statistics and applied probability. Chapman & Hall, London, 416 p. (Reprinted in 2003 with additional material by The Blackburn Press)
- Azzalini A (2020) The R package sn: the skew-normal and related distributions such as the Skew- t (version 1.6-2). Università di Padova, Italia, URL <http://azzalini.stat.unipd.it/SN>
- Azzalini A, Capitanio A (1999) Statistical applications of the multivariate skew-normal distribution. *J R Statist Soc B* 61(3):579–602
- Egozcue JJ, Pawlowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis. *Math Geol* 37:795–828
- Egozcue JJ, Pawlowsky-Glahn V (2016) Changing the reference measure in the simplex and its weighting effects. *Aust J Stat* 45(4):25–44
- Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. *Math Geol* 35:279–300
- Filzmoser P (2020) StatDA: statistical analysis for environmental data. <https://CRAN.R-project.org/package=StatDA>, R package version 1.7.4
- Filzmoser P, Hron K (2008) Outlier detection for compositional data using robust methods. *Math Geosci* 40(3):233–248
- Filzmoser P, Hron K, Templ M (2018) Applied compositional data analysis. Springer, Cham
- Martín-Fernández JA (2019) Comments on: compositional data: the sample space and its structure, by Egozcue and Pawlowsky-Glahn. *TEST* 28:653–657
- Mateu-Figueras G, Pawlowsky-Glahn V (2007) The skew-normal distribution on the simplex. *Commun Stat Theory Methods* 36(9):1787–1802
- Mateu-Figueras G, Pawlowsky-Glahn V (2008) A critical approach to probability laws in geochemistry. *Math Geosci* 40(5):489–502
- Mateu-Figueras G, Pawlowsky-Glahn V, Barceló-Vidal C (2005) The additive logistic skew-normal distribution on the simplex. *Stoch Env Res Risk A* 18:205–214
- Mateu-Figueras G, Pawlowsky-Glahn V, Egozcue JJ (2013) The normal distribution in some constrained sample spaces. *SORT* 37:29–56
- Pawlowsky-Glahn V, Egozcue J, Tolosana-Delgado R (2015) Modeling and analysis of compositional data. Wiley, Chichester
- Reimann C, Åyräs M, Chekushin V, Bogatyrev I, Boyd R, Caritat P, Dutter R, Finne TE, Halleraker JH, Jæger Ø, Kashulina G, Lehto O, Niskavaara H, Pavlov VK, Räisänen ML, Strand T, Volden T (1998) Environmental geochemical atlas of the central parts of the Barents region. Geological Survey of Norway, Trondheim
- Talská R, Menafoglio A, Hron K, Egozcue JJ, Palarea-Albaladejo J (2020) Weighting the domain of probability densities in functional data analysis. *Stat* 9(1):e283

Agterberg, Frits

Qiuming Cheng

School of Earth Science and Engineering, Sun Yat-Sen University, Zhuhai, China

State Key Lab of Geological Processes and Mineral Resources, China University of Geosciences, Beijing, China



Fig. 1 Photo of Frits Agterberg taken in China University of Geosciences, Beijing, in November 2017

Biography

Frederik Pieter (Frits) Agterberg was born in 1936 in Utrecht, the Netherlands. He studied geology and geophysics at Utrecht University, obtaining B.Sc. (1957), M.Sc. (1959), and Ph.D. (1961). These three degrees were obtained “cum laude” (with distinction). After a 1-year Wisconsin Alumni Research Foundation postdoctorate fellowship at the University of Wisconsin, he joined the Geological Survey of Canada in 1962, initially as petrological statistician working on the Canadian contribution to the International Upper Mantle Project. Later, he formed and headed the Geomathematics Section of the Geological Survey of Canada in Ottawa (1971–1996).

He has made a major contribution to the geomathematical literature in numerous areas, often as the first to explore new applications and methods, and always with mathematical rigor blended with practical examples. The major research

areas include statistical frequency distributions applied to geoscience data, mineral-resources quantitative assessment, stratigraphic analysis and timescales, and fractal and multi-fractal modeling. As one of the founders of the International Association for Mathematical Geosciences (IAMG). Frits Agterberg, in his association with the IAMG, has been a key figure in its ongoing success and is widely recognized as one of the fathers of mathematical geology.

Agterberg has authored or coauthored over 350 scientific journal articles and books, including *Computer Programs for Mineral Exploration* published in 1989 by Science (Agterberg 1989), which has become a routinely used computer program for mineral resources assessments worldwide. The books he had published include the textbook *Geomathematics: Mathematical Background and Geo-Science Applications*, published in 1974 by Elsevier with approximately 10,000 copies sold worldwide (Agterberg 1974), the monograph *Automated Stratigraphic Correlation*, published in 1990 by Elsevier (Agterberg 1990), and the textbook *Geomathematics: Theoretical Foundations, Applications and Future Developments*, published in 2014 by Springer (Agterberg 2014). He has edited or coedited seven other books and many special issues in scientific journals.

Frits Agterberg received several prestigious awards and recognitions, including the third W.C. Krumbein medalist of the International Association for Mathematical Geology in 1978, Best Paper Awards for articles in the journal *Computers & Geosciences* for 1978, 1979, and 1982, correspondent member of the Royal Dutch Academy of Sciences in 1981, and as Honorary Professor of the China University of Geosciences in 1987. A newly discovered fossil, *Adercotrima agterbergi*, was named after him to recognize his contributions to quantitative stratigraphy.

Since 1968, he was associated with the University of Ottawa, where he has taught an undergraduate course on “Statistics in Geology” for 25 years and served as a supervisor for undergraduate and graduate students. Several of his former students now occupy prominent positions in the universities, government organizations, and mining industry in Canada and abroad. Other academic positions included being Distinguished Visiting Research Scientist at the Kansas Geological Survey of the University of Kansas (1969–1970), Adjunct Professor at Syracuse University (1977–1981), Esso Distinguished Lecturer for the Australian Mineral Resource Foundation, University of Sydney (August–November 1980), and Adjunct Research Professor, Department of Mathematics, Carleton University, Ottawa (1986–1994). Since 2000, he has served as a guest and distinguished professor at China

University of Geosciences (CUG) both in Beijing and Wuhan, where he has co-supervised a dozen Ph.D. students with his former student Professor Qiuming Cheng at York University.

Agterberg has lectured in more than 40 short courses worldwide, including several lectures delivered when he was named IAMG Distinguished Lecturer in 2004. From 1979 to 1985, he was leader of the International Geological Correlation Programme’s Project (IGCP) on “Quantitative Stratigraphic Correlation Techniques.” He has served on numerous committees, editorial boards, and councils of national and international organizations. For example, he served as an associate editor of several journals, including the *Canadian Journal of Earth Sciences*, the *Bulletin of the Canadian Institute of Mining and Metallurgy*, *Mathematical Geosciences*, *Computers & Geosciences*, and *Natural Resources Research*, the president of IAMG from 2004 to 2008, and general secretary of IAMG from 2012 to 2016. He chaired the Quantitative Stratigraphy Committee of the International Stratigraphic Commission (ICS), which is part of the International Union of Geological Sciences (IUGS).

In 1996, Frits Agterberg commenced a phased retirement from the Geological Survey of Canada to work as part-time independent geomathematical consultant for the industry. He continues to teach and supervise graduate students at the University of Ottawa. In November of 2017, his Chinese students and friends organized a workshop with a party on the CUGB campus for celebrating his 81st birthday and his contributions to Chinese Mathematical Geosciences. Now he is still tirelessly working on scientific publications, including coediting a large volume of the *Encyclopedia of Mathematical Geosciences*, which includes several hundred international scientists working on over 300 separate articles on different topics of mathematical geosciences.

Bibliography

- Agterberg F (1974) *Geomathematics, mathematical background and geo science application*. Development in geomathematics. Elsevier, Amsterdam
- Agterberg F (1989) *Computer programs for mineral exploration*. Science 245(4913):76–81
- Agterberg F (1990) *Automated stratigraphic correlation*. Elsevier, Amsterdam
- Agterberg F (2014) *Geomathematics: theoretical foundations, applications and future developments*. Springer, Cham, Switzerland

Aitchison, John

John Bacon-Shone

Social Sciences Research Centre, The University of Hong Kong, Hong Kong, China



Fig. 1 John Aitchison, courtesy of Prof. Bacon-Shone

Biography

John Aitchison was born in East Linton, East Lothian, Scotland, on July 22, 1926, and passed away in Glasgow on December 23, 2016.

John first studied mathematics at the University of Edinburgh, where he received his MA, before attending Cambridge on a scholarship, where he initially studied mathematics, before a course from Frank Anscombe diverted him to Statistics. He started his career as a statistician in the Department of Applied Economics at the University of Cambridge, where he wrote his first classic text on the lognormal distribution with Alan Brown (Aitchison and Brown 1957). This book already showed the importance of accounting for restricted sample spaces by transformation or using non-Euclidean distance metrics. He then returned to Scotland, where he was a Lecturer in Statistics in the University of Glasgow. After 5 years at the University of Liverpool, he returned to Glasgow to become Titular Professor of Statistics and Mitchell Lecturer in Statistics and write his second classic book on statistical prediction analysis (Aitchison and Dunsmore 1975). Then John took a big step into the unknown as Professor of Statistics at the University of Hong Kong. During his time in Hong Kong, John presented his seminal paper on statistical compositional data analysis (Aitchison 1982) as a read paper to the Royal Statistical Society in

London, for which he later received the Guy Medal in Silver. This influential paper (over 5000 citations) finally solved Pearson's problem of spurious correlation (Pearson 1897), which Chayes (Chayes 1960) had linked to compositional data. While others had used the logistic transformation for binary outcome probabilities, John alone saw the potential for modeling compositions in general by using the logistic Normal distribution. This paper developed into an essential monograph (Aitchison 1986). This book (which has received over 5000 citations) and many other influential related papers, which together revolutionized the statistical analysis of compositions, led to John receiving the William Christian Krumbein Medal of the International Association for Mathematical Geosciences in 1997 for a "distinguished career as author and researcher in the field of mathematical geology." After retirement from Hong Kong, John became Professor of Statistics in the University of Virginia, before finally returning to Glasgow.

Bibliography

- Aitchison J (1982) The statistical analysis of compositional data. *J R Stat Soc Ser B Methodol* 44(2):139–177
- Aitchison J (1986) The statistical analysis of compositional analysis. Chapman & Hall, London
- Aitchison J, Brown J (1957) The lognormal distribution. Cambridge University Press, Cambridge
- Aitchison J, Dunsmore IR (1975) Statistical prediction analysis. Cambridge University Press, Cambridge
- Chayes F (1960) On correlation between variables of constant sum. *J Geophys Res* 65(12):4185–4193
- Pearson K (1897) On a form of spurious correlation which may arise when indices are used, etc. *Proc R Soc* 60:489–498

Algebraic Reconstruction Techniques

Utkarsh Gupta and Uma Ranjan

Department of Computational and Data Sciences, Indian Institute of Science, Bengaluru, Karnataka, India

Definition

Algebraic reconstruction techniques (ART) were initially proposed for radiology and medical applications but have been extensively used to study various seismological phenomena (Del Pino 1985; Peterson et al. 1985). These algorithms are based on iterative reconstruction techniques and efficiently help invert matrices of large dimensions, taking advantage of the sparse nature of the matrices. These methods gained popularity due to the fact that they were able to handle various data geometries with irregular sampling and limited projection angle (Peterson et al. 1985). In addition, these techniques

are relatively more straightforward to implement compared to transform methods described by Scudder (Scudder 1978).

This entry describes fundamental principles and derivations of ART along with various modifications to ART, namely simultaneous iterative reconstructive technique (SIRT) and simultaneous algebraic reconstructive technique (SART).

Tomographic Projection Model

In iterative reconstruction techniques, one starts by discretizing the problem, as shown in the Fig. 1. We achieve discretization by superimposing a square grid on the image $g(x, y)$; we assume that each square grid cell holds a constant value denoted by g_j (subscript j denotes the j^{th} cell), and N represents the total number of cells. In addition, a ray denotes a line in the $x - y$ plane, and p_i represents the ray sum (also referred as line-integral) along i^{th} ray as shown in Fig. 1. Relation between g'_j s and p'_i s can be expressed as in Eq. 1.

$$\sum_{j=1}^N a_{ij} g_j = p_i, \quad i = 1, 2, \dots, M \quad (1)$$

where M represents the total number of rays across all the projection direction, and a_{ij} represents the weight factor which is assigned for j^{th} cell and i^{th} ray. Eq. 1 represents a large set of linear equations that needs to be inverted to determine the unknowns g'_j s.

For small values of N and M , conventional linear algebra methods like matrix inversion to determine the unknowns g'_j s

can be applied. But, in almost all the practical scenarios, values of N and M are quite large. For example, for the image of dimension 256×256 , then the value of N is as large as 65,000. For real scenarios, M 's magnitude follows the same order as of N . Direct matrix inversion is often impossible even for cases where values of N and M are small due to measurement noise in the projection data or the cases when $M < N$. In all such cases, we depend upon simple least square method. These least square-based methods are not applicable for real scenarios as M and N will be huge.

Therefore, in situations where conventional linear algebra fails to achieve the desired solution, algebraic reconstruction techniques (ART) comes into picture.

ART Algorithm

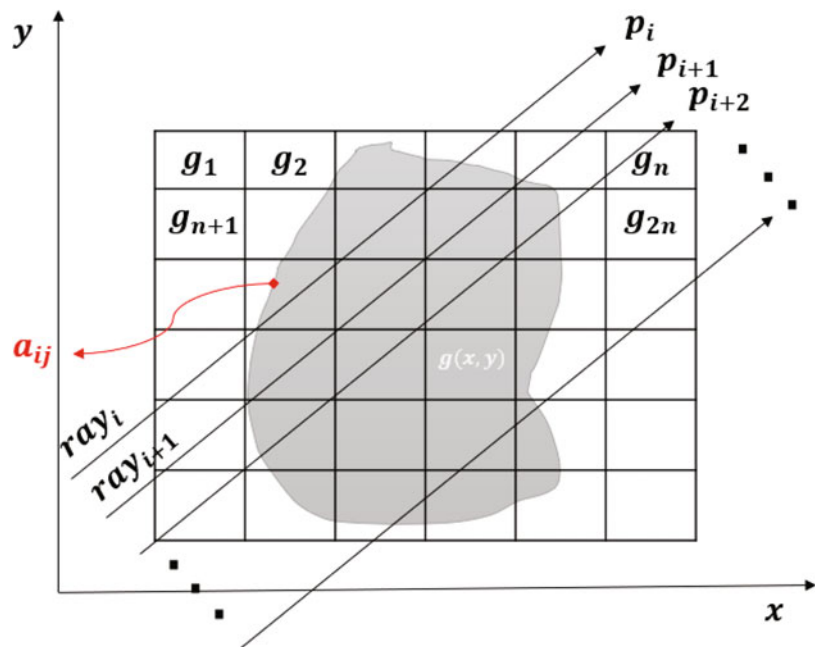
ART algorithm is based on *Methods of Projections*, proposed by Kaczmarz (1937) (hence sometimes referred to as Kaczmarz method). Eq. 1 can be unrolled as set of linear equations as follows:

$$\begin{aligned} a_{11}g_1 + a_{12}g_2 + a_{13}g_3 + \dots + a_{1N}g_N &= p_1, \\ a_{21}g_1 + a_{22}g_2 + a_{23}g_3 + \dots + a_{2N}g_N &= p_2, \\ &\vdots \\ a_{M1}g_1 + a_{M2}g_2 + a_{M3}g_3 + \dots + a_{MN}g_N &= p_M. \end{aligned} \quad (2)$$

Further, we will represent an image $g(x, y)$ as $\vec{g} = [g_1, g_2, g_3, \dots, g_N]$ which is a N -dimensional vector of unknowns g'_j s in a N -dimensional Euclidean space. Each sub-equation in Eq. 2 represents an individual hyperplane.

Algebraic Reconstruction

Techniques, Fig. 1 Illustration of tomographic projection process, with a_{ij} representing the weight contribution for j^{th} pixel and corresponding i^{th} ray



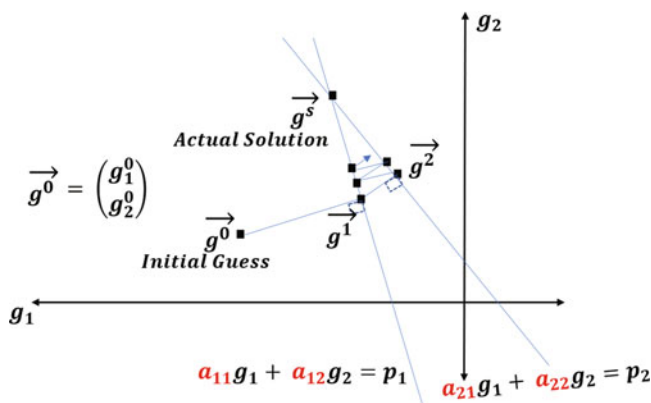
A unique solution will exist only when all the M hyperplanes intersect at a single point.

For better understanding and visualization, we explain the algorithm in two-dimensional Euclidean space with two unknowns g_1 and g_2 . These unknowns satisfy the following equations:

$$\begin{aligned} a_{11}g_1 + a_{12}g_2 &= p_1, \\ a_{21}g_1 + a_{22}g_2 &= p_2. \end{aligned} \quad (3)$$

Sub-equations in Eq. 3 represent the equations of 2D-lines, and the same is graphically shown in Fig. 2. In order to, iteratively solve the pair of linear equations (Eq. 3), we start with an initial guess $\vec{g}^0$, which is projected perpendicularly onto the first equation resulting in $\vec{g}^1$, then reprojecting the resulting point $\left(\vec{g}^1\right)$ onto the second equation and then back projecting onto the first line and so on. If a unique solution exists, the system will always converge to the solution (which in this case is represented as $\vec{g}^s$ in Fig. 2).

We can extend the same process described for solving a pair of linear equations to a large set of linear equations. The method begins with an initial guess represented by vector $\vec{g}^0 = [g_1^0, g_2^0, g_3^0, \dots, g_N^0]$. In most of the computational implementations, $\vec{g}^0$ is represented as a $\vec{0}$ vector. The initial guess is projected onto the first hyperplane of Eq. 2 giving $\vec{g}^1$. Thereafter, $\vec{g}^1$ is projected onto the second hyperplane of Eq. 2 to yield $\vec{g}^2$ and so on. When $\vec{g}^{(i-1)}$ is projected onto i^{th}



Algebraic Reconstruction Techniques, Fig. 2 The Kaczmarz method of solving algebraic equations is illustrated for the case of two unknowns. It begins with some arbitrary initial guesses and then projects onto the line corresponding to the first equation. The resulting point is now projected onto the line representing the second equation. If there are only two equations, this process is continued back and forth, until convergence is achieved. [Modified from Kak and Slaney (2001)]

hyperplane resulting in $\overrightarrow{g^{(i)}}$, this operation can be mathematically expressed as:

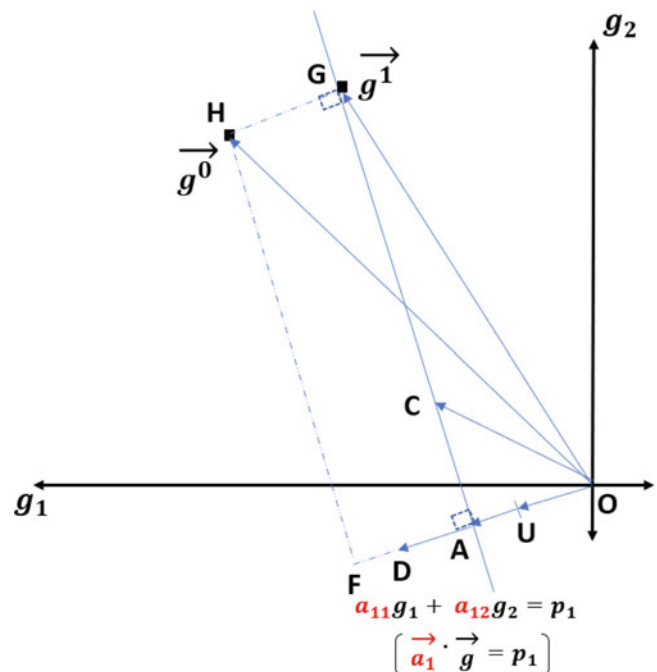
$$\overrightarrow{g^{(i)}} = g^{(i-1)} - \frac{\left(\overrightarrow{g^{(i-1)}} \cdot \overrightarrow{a_i} - p_i\right)}{\left(\overrightarrow{a_i} \cdot \overrightarrow{a_i}\right)} \overrightarrow{a_i} \quad (4)$$

where $(\vec{a_i} \cdot \vec{a_i})$ represents the scalar product operation and $\vec{a_i} = [a_{i1}, a_{i2}, a_{i3} \dots a_{iN}]$. Eq. 4 represents the vectorized update equation for the ART algorithm. Next section covers the derivation for Eq. 4. Vector form of the first hyperplane in Eq. 2 can be written as:

$$\vec{a}_1 \cdot \vec{g} = p_1 \quad (5)$$

where the normal vector to the hyperplane is given by $\vec{a_1} = [a_{11}, a_{12}, a_{13}, \dots, a_{1N}]$ (represented by $\vec{OD}$ in Fig. 3). Eq. 5 simply says that the projection of any vector $\vec{OC}$ (where C is any random point on the hyperplane) on the vector a_1 is of constant length. Unit vector along the direction of normal vector ($\vec{a_1}$) is represented by $\vec{OU}$ given by:

$$\overrightarrow{OU} = \frac{\overrightarrow{a_1}}{\sqrt{\overrightarrow{a_1} \cdot \overrightarrow{a_1}}} \quad (6)$$



Algebraic Reconstruction Techniques, Fig. 3 The hyperplane $\vec{a}_1 \cdot \vec{g} = p_1$ perpendicular to the vector $\vec{OD}$ ($\vec{a}_1$)

Also, the perpendicular distance of the hyperplane from the origin is represented by $\overrightarrow{OA}$ in Fig. 3, is given by $\overrightarrow{OC} \cdot \overrightarrow{OU}$:

$$\begin{aligned} |\overrightarrow{OA}| &= \overrightarrow{OU} \cdot \overrightarrow{OC} = \frac{1}{\sqrt{\overrightarrow{a_1} \cdot \overrightarrow{a_1}}} (\overrightarrow{a_1} \cdot \overrightarrow{OC}) \\ &= \frac{1}{\sqrt{\overrightarrow{a_1} \cdot \overrightarrow{a_1}}} (\overrightarrow{a_1} \cdot \overrightarrow{g}) = \frac{p_1}{\sqrt{\overrightarrow{a_1} \cdot \overrightarrow{a_1}}} \end{aligned} \quad (7)$$

To calculate $\overrightarrow{g}^1$, we subtract the initial guess $\overrightarrow{g}^0$ from the vector $\overrightarrow{HG}$:

$$\overrightarrow{g}^{(1)} = \overrightarrow{g}^{(0)} - \overrightarrow{HG} \quad (8)$$

where the length of the vector $\overrightarrow{HG}$ is represented by:

$$\begin{aligned} |\overrightarrow{HG}| &= |\overrightarrow{OF}| - |\overrightarrow{OA}| \\ &= \overrightarrow{g}^{(0)} \cdot \overrightarrow{OU} - |\overrightarrow{OA}| \end{aligned} \quad (9)$$

Substituting Eqs. 6 and 7 in Eq. 9, we get:

$$|\overrightarrow{HG}| = \frac{\left(\overrightarrow{g}^{(0)} \cdot \overrightarrow{a_1} - p_1 \right)}{\sqrt{\overrightarrow{a_1} \cdot \overrightarrow{a_1}}} \quad (10)$$

Since $\overrightarrow{HG}$ follows the same direction as the unit vector $\overrightarrow{OU}$, we can write:

$$\overrightarrow{HG} = |\overrightarrow{HG}| \overrightarrow{OU} = \frac{\left(\overrightarrow{g}^{(0)} \cdot \overrightarrow{a_1} - p_1 \right)}{(\overrightarrow{a_1} \cdot \overrightarrow{a_1})} \overrightarrow{a_1} \quad (11)$$

Finally substituting Eq. 11 in Eq. 8, we get:

$$\overrightarrow{g}^{(1)} = \overrightarrow{g}^{(0)} - \frac{\left(\overrightarrow{g}^{(0)} \cdot \overrightarrow{a_1} - p_1 \right)}{(\overrightarrow{a_1} \cdot \overrightarrow{a_1})} \overrightarrow{a_1} \quad (12)$$

Eq. 4 represents the generalized form of Eq. 12.

Convergence of ART Algorithm

Tanabe (1971) proved that if a unique solution $\overrightarrow{g}_s$ exists to a system of linear equations described by Eq. 2 (refer Fig. 2), then:

$$\lim_{k \rightarrow +\infty} \overrightarrow{g}^{(kM)} = \overrightarrow{g}_s \quad (13)$$

Few more comments about convergence have been presented below:

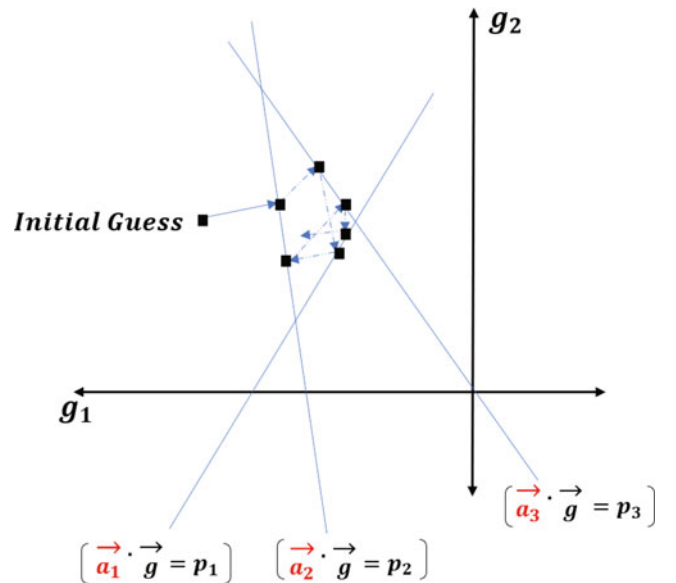
Case 1: Suppose, in Fig. 2, two lines are perpendicular to each other. In that case, it is possible to arrive at the actual solution in just two iterations following the update Eq. 4, and the convergence is independent of the choice of the initial guess.

Case 2: Suppose, in Fig. 2, two lines make a very small angle between them. In that case, we may require a significant number of iterations (k value in Eq. 13) to reach the actual solution depending upon the choice of the initial guess.

Clearly, rate of convergence of ART algorithm is directly influenced by the angles between the hyperplanes. If M hyperplanes in Eq. 2 can be made orthogonal with respect to one another, we can reach the actual solution in just one pass through the M equation (only in case a unique solution exists).

Characteristics of the Solution

Overdetermined System ($M > N$): Not an uncommon case, it is possible that we have more number of hyperplanes (M) than the number of unknowns (N) in Eq. 2, and also projection data ($p_1, p_2, \dots, p_M$) is corrupted by measurement noise. In



Algebraic Reconstruction Techniques, Fig. 4 Illustration of the case where all hyperplanes do not intersect at a single point. In such cases, iterations will continue to oscillate in the neighborhood of the intersections of the hyperplanes

such cases, no unique solution is possible (as shown in Fig. 4). In all such cases, ART iteration never converges to a unique point, but iterations will continue to oscillate in the neighborhood of the intersections of the hyperplanes (as shown in Fig. 4).

Under-determined System ($M < N$): In such cases, there is no unique solution, but in fact, there are an infinite number of solutions possible. It can be shown in such cases that ART iteration converges to a solution $\vec{g}_s$ such that $|\vec{g}^0 - \vec{g}_s|$ is minimized.

Advantages and Disadvantages of ART

The key feature of the iterative steps presented for ART algorithm here is that it is now possible to integrate the prior information about the reconstructed image $g(x, y)$. For example, if it is known that pixels of the reconstructed image cannot take a nonnegative value, then in each iteration, we can set negative components equal to zero. Similar to the above example, we can incorporate other information about the image $g(x, y)$ if known or conveyed in advance.

In applications requiring large-sized reconstructions, difficulties arise in calculation, storage, and efficient retrieval of weight coefficients a_{ij} . In most of the ART implementations, these weights a_{ij} exactly represent the length of the intersection of the i^{th} ray with the j^{th} pixel. It should be noted that line-length is not the only way to describe the distribution of weights; various physics-based concepts like attenuation and point spread function can also be incorporated.

Eq. 4 represents the vectorized update equation for the ART algorithm. Below we describe the ART algorithm as the gray levels/pixel (g_j) wise update equation:

$$g_j^{(i)} = g_j^{(i-1)} + \frac{(p_i - q_i) \vec{a}_{ij}}{\sum_{k=1}^N a_{ik}^2} \vec{a}_{ij} \quad (14)$$

where

$$q_i = \vec{g}^{(i-1)} \cdot \vec{a}_i \quad (15)$$

$$= \sum_{k=1}^N g_k^{(i-1)} a_{ik} \quad (16)$$

These equations represent that when we project $((i-1)^{\text{th}})$ onto the i^{th} equation, the gray level of the j^{th} pixel ($g_j^{(i-1)}$) is corrected by a factor of $\Delta g_j^{(i)}$ where:

$$\Delta g_j^{(i)} = g_j^{(i)} - g_j^{(i-1)} = \frac{p_i - q_i}{\sum_{k=1}^N a_{ik}^2} a_{ij} \quad (17)$$

Here, p_i is the measured line-integral/ray sum along the i^{th} ray, and q_i is the computed line-integral/ray sum for the same ray but using the $(i-1)^{\text{th}}$ solution for the image gray levels.

ART reconstruction, in general, suffers from salt and pepper noise which generally arises due to inconsistencies introduced while modeling the weights (a_{ik}) in the forward projection process. Due to these inconsistencies, computed ray sum (q_i in Eq. 15) is different from the measured ray sum (p_i). In the practical scenario, it is possible to reduce the effect of noise by relaxation, in which instead of updating the gray level by the factor $\Delta g_j^{(i)}$, we update it by the factor of $\alpha \Delta g_j^{(i)}$, where α is less than 1. At the same time, introducing the relaxation parameter α in the update equation leads to the slower convergence of the ART algorithm.

Finally, the update equation for the ART algorithm including the relaxation parameter α is described below:

$$g_j^{(i+1)} = g_j^{(i)} + \alpha \Delta g_j^{(i)} \quad (18)$$

$$\vec{g}^{(i+1)} = \vec{g}^{(i)} + \alpha \frac{(p_i - \vec{g}^{(i)} \cdot \vec{a}_i)}{(\vec{a}_i \cdot \vec{a}_i)} \vec{a}_i \quad (19)$$

Simultaneous Iterative Reconstructive Technique (SIRT)

Simultaneous iterative reconstructive technique (SIRT) (Gilbert 1972) is another iterative algorithm that produces a smoother and better-looking reconstructed image than those produced from the ART algorithm at the cost of slower convergence. We again compute the correction term $\Delta g_j^{(i)}$ (using Eq. 17) in the j^{th} pixel caused by i^{th} equation. But before adding the correction term to the j^{th} gray level, we calculate all the correction terms calculated from all the M equations in Eq. 2, after that, we update the gray level in the j^{th} cell by adding the average of all the obtained correction terms obtained from each equations. This concludes the one iteration of the SIRT algorithm. In the second iteration, we again start with the first equation in Eq. 2, and the process is repeated. Mathematically SIRT update equation can be represented as:

$$g_j^{(k+1)} = g_j^{(k)} + \alpha \frac{\sum_{i=1}^M \Delta g_j^{(i)}}{M} \quad (20)$$

where k represents the iteration number, M signifies number of total projection rays or total number of algebraic equations, and α represents the relaxation parameter.

Simultaneous Algebraic Reconstructive Technique (SART)

Simultaneous algebraic reconstructive technique (SART) (Andersen and Kak 1984) was designed with mainly two objectives in mind: first is to reduce the salt and pepper noise commonly found in ART-type reconstruction algorithms, and secondly to achieve good quality and numerical accuracy in only a single iteration. We will now describe the key steps involved in SART algorithm.

Modeling the Forward Projection Process

For SART-type reconstruction algorithm, we express the image $g(x, y)$ as the linear combination of N basis images ($b_j(x, y)$) weighted by real numbers $\beta_1, \beta_2, \beta_3, \dots, \beta_N$.

This can be mathematically expressed as:

$$g(x, y) \approx \hat{g}(x, y) = \sum_{j=1}^N \beta_j b_j(x, y) \quad (21)$$

$\hat{g}(x, y)$ is a discrete approximation of the actual image $g(x, y)$, and β_j 's are the finite set of numbers which completely describes the image relative to the chosen basis images ($b_j(x, y)$).

If $r_i(x, y) = 0$ is the equation of i^{th} ray, the projection operator R_j along the direction of the ray can be expressed as follows:

$$p_i = R_i g(x, y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) \delta(r_i(x, y)) dx dy \quad (22)$$

Assuming the projection operator R_j as a linear operator, we have:

$$R_i \hat{g}(x, y) = \sum_{j=1}^N \beta_j R_i b_j(x, y) \quad (23)$$

In practical implementation of SART-type algorithm, we compute $R_i \hat{g}(x, y)$ using some approximations a_{ij} to reduce the computational load of the algorithm. Hence, we get:

$$\begin{aligned} p_i &= R_i g(x, y) \approx R_i \hat{g}(x, y) = \sum_{j=1}^N \beta_j R_i b_j(x, y) \\ &\approx \sum_{j=1}^N \beta_j a_{ij} \end{aligned} \quad (24)$$

$$p_i = \sum_{j=1}^N \beta_j a_{ij} \quad (25)$$

Here, a_{ij} represent the line integral of $b_j(x, y)$ along the i^{th} ray. Eq. 25 shares the same structure as Eq. 1 and is yet more generalized since β_j does not represent the image gray levels as g_j in Eq. 1. Eq. 25 and Eq. 1 are equivalent if in case image frame is divided into N identical sub-squares, and we use the basis image as described by Eq. 26.

$$b_j(x, y) = \begin{cases} 1, & \text{inside the } j^{th} \text{ pixel} \\ 0, & \text{otherwise} \end{cases} \quad (26)$$

SART algorithm provides superior reconstruction using the forward projection process described above. Accurate reconstructions are obtained using bilinear elements which are simplest higher order basis functions. It can be shown that using bilinear basis we obtain a more continuous image reconstruction than those which are obtained using pixel basis. However, computing line-integral across these bilinear basis is a computationally expensive task. Hence we approximate the overall ray integral $R_i \hat{g}(x, y)$ by finite sum involving M_i equidistant point $\{\hat{g}(s_{im})\}$ (see Fig. 5a).

$$p_i = \sum_{m=1}^{M_i} \hat{g}(s_{im}) \Delta s \quad (27)$$

$\hat{g}(s_{im})$ is determined using bilinear interpolation using the values β_j of $g(x, y)$ on the four neighboring points on the sample lattice. We get:

$$\hat{g}(s_{im}) = \sum_{j=1}^N d_{ijm} \beta_j \quad (28)$$

where d_{ijm} represents the contribution that is made by the j^{th} image to the m^{th} point on the i^{th} ray. Overall, we can approximate the line-integral p_i as the linear combination of the β_j :

$$p_i = \sum_{m=1}^{M_i} \sum_{j=1}^N d_{ijm} \beta_j \Delta s \quad (29)$$

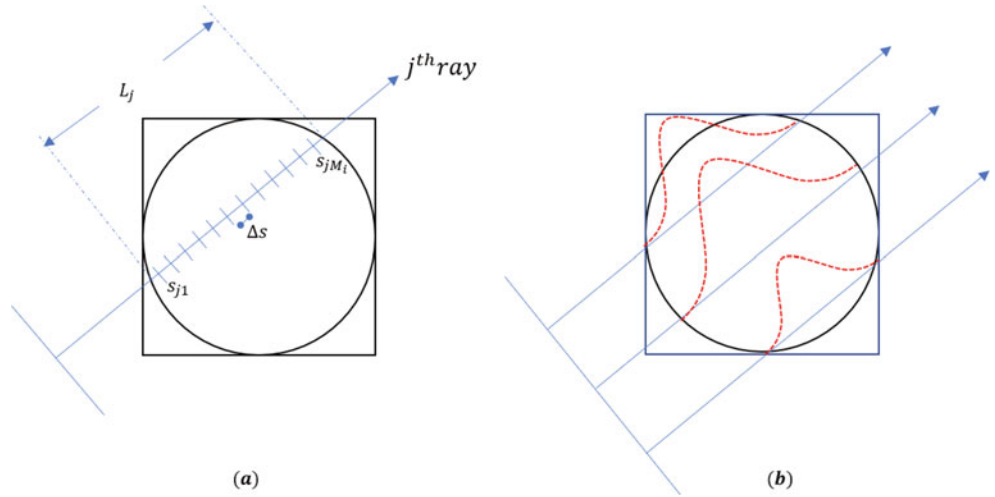
$$p_i = \sum_{j=1}^N \sum_{m=1}^{M_i} d_{ijm} \beta_j \Delta s, 1 \leq i \leq J \quad (30)$$

$$= \sum_{j=1}^N a_{ij} \beta_j \quad (31)$$

Algebraic Reconstruction Techniques, Fig. 5 (a)

Illustration of the ray-integral equations for a set of equidistant points along a straight line cut by the circular reconstruction region.

(b) Longitudinal Hamming window for set of rays



where a_{ij} is computed as the sum of contribution from different points along the ray.

$$a_{ij} = \sum_{m=1}^{M_i} d_{ijm} \Delta s \quad (32)$$

Also, the overall weights (a_{ij}) are adjusted in such a way that $\sum_{j=1}^N a_{ij}$ is actually equal to the physical length L_j . The precise formula used for updating the β'_i 's using SART algorithm is stated as:

$$\vec{\beta}_j^{(k+1)} = \vec{\beta}_j^{(k)} + \frac{\sum_i \left[t_{ij} \frac{(\vec{p}_i - \vec{\beta}^{(k)}) \cdot \vec{a}_i}{\sum_{j=1}^N a_{ij}} \right]}{\sum_i a_{ij}} \quad (33)$$

where

$$t_{ij} = \sum_{m=1}^{M_i} h_{im} d_{ijm} \Delta s \quad (34)$$

where the summation with respect to i is over the rays intersecting the j^{th} image cell for a given projection direction. The sequence h_{im} , for $1 \leq m \leq M_i$, is a Hamming window of length M_i and length of the window varies according to the number of point M_i describing the part of ray inside the reconstruction circle.

On the whole, SART algorithm described using the update Eq. 33 results in a less noisy reconstructed image when compared to images reconstructed using ART and SIRT type algorithms.

Summary and Conclusions

This chapter covers three essential algebraic reconstruction techniques: ART, SIRT, and SART, prominent in geoscience, primarily for seismological research and application. Also, these algorithms form the basis for several other algebraic-based reconstruction techniques proposed for various tasks. For instance, C-SART (constrained simultaneous algebraic reconstruction technique) algorithm has been applied for ionospheric tomography (Hobiger et al. 2008) that has a higher convergence rate than classical SART. These algorithms, being iterative in nature, have the added advantage of requiring less computational storage along with fast convergence to the solution in case of large and sparse linear systems (Xun 1987).

Bibliography

- Andersen AH, Kak AC (1984) Simultaneous algebraic reconstruction technique (SART): a superior implementation of the ART algorithm. *Ultrason Imaging* 6(1):81–94
- Del Pino E (1985) Seismic wave polarization applied to geophysical tomography. In: SEG technical program expanded abstracts 1985. Society of Exploration Geophysicists, Tulsa, pp 620–622
- Dines KA, Lytle RJ (1979) Computerized geophysical tomography. *Proc IEEE* 67(7):1065–1073
- Gilbert P (1972) Iterative methods for the three-dimensional reconstruction of an object from projections. *J Theor Biol* 36(1):105–117
- Hobiger T, Kondo T, Koyama Y (2008) Constrained simultaneous algebraic reconstruction technique (C-SART) – a new and simple algorithm applied to ionospheric tomography. *Earth Planets Space* 60(7): 727–735
- Kak AC, Slaney M (2001) Principles of computerized tomographic imaging. Society for Industrial and Applied Mathematics, Philadelphia

- Karczmarz S (1937) Angenaherte auflosung von systemen linearer gleichungen. Bull Int Acad Pol Sic Let Cl Sci Math Nat 35:355–357
- Peterson JE, Paulsson BN, McEvilly TV (1985) Applications of algebraic reconstruction techniques to crosshole seismic data. Geophysics 50(10):1566–1580
- Scudder HJ (1978) Introduction to computer aided tomography. Proc IEEE 66:628–637
- Tanabe K (1971) Projection method for solving a singular system of linear equations and its applications. Numer Math 17(3):203–214
- Xun L (1987) Art algorithm and its applications in geophysical inversion. Comput Techn Geophys Geochem Explor 9(1):34

Allometric Power Laws

Gabor Korvin

Earth Science Department, King Fahd University of
Petroleum and Minerals, Dhahran, Kingdom of Saudi Arabia

Definition

Allometry has its origin in biology, anatomy, and physiology; it studies the effect of body size on shape and metabolic rate, the laws of growth, and relationships between size and different measurable properties of living organisms. Its laws were first summarized by D’Arcy Thompson in 1917 and Julian Huxley in 1932 (Stevens 2009). Frequently, relationships between size and other measured quantities obey power law equations (*Allometric Power Laws*) of the form $y = kx^\alpha$ (with standard notation $y \propto x^\alpha$ or $y \sim x^\alpha$, where $\propto$ or $\sim$ denote proportionality). The exponent α is in most cases a non-integer real number, which explains the name of the discipline (allometry < Gr. *ἄλλωμετρον* = “strange scale”). APLs of the form $y \propto x^\alpha$ are also called *scaling laws*, because they are *scale-invariant*, and they arise when describing *self-similar objects*, such as earth-materials and formations on the surface and in the interior of the Earth and planets.

Introduction: Allometric Power Laws and Self-Similarity

A *power law* is a functional relationship between two quantities, and it has the form $y = kx^\alpha$ (with standard notation $y \propto x^\alpha$ or $y \sim x^\alpha$, where $\propto$ or $\sim$ denote proportionality). Famous power laws are the *allometric power laws* (APLs in what follows), originally used to express relationships between body size and biological variables (Stevens 2009), such as the empirical rule that metabolic rate q_0 is proportional to body mass M raised to the $(3/4)^{\text{th}}$ power: $q_0 \propto M^{3/4}$, or that heart

rate τ is inversely proportional to the $(1/4)^{\text{th}}$ power of M : $\tau \propto M^{-1/4}$. Dimensional considerations would suggest a metabolic rate proportional to the surface area available for energy transfer, so one would expect that metabolic rates of similar organism should scale as $q_0 \propto M^{2/3}$ rather than the empirical “ $M^{0.75}$ ” law, and the very name of such laws (allometry < Gr. *ἄλλωμετρον* = “strange scale”) expresses the fact that most observed APL exponents are “strange,” in the sense that their values usually cannot be derived from physical principles or dimensional analysis (Korvin 1992: 233).

APLs of the form $y = kx^\alpha$ (where both sides are positive) can be written, taking logarithm, as $\log y = \alpha \log x + \log k$ or $\ln y = \alpha \ln x + \ln k$ that is they show up as straight lines when plotted on double-logarithmic axes, and the *scaling exponent* is obtained from the slope.

APLs are *scale invariant*. Given an APL $f(x) = k \cdot x^\alpha$, multiplying the argument x by a factor λ causes only a proportionate change in the function itself: $(\lambda x) = k(\lambda x)^\alpha = \lambda^\alpha f(x) \propto f(x)$.

The frequent occurrence of such APLs in Geosciences (see Table 1) which connect two measured properties of a geologic object (as, e.g., perimeter and area of islands, width and depth of rivers, etc.) is a consequence of the *similarity* of all geological objects of the same kind (as, e.g., all mountains, all islands, all sea coasts, all meandering rivers), or it arises because of their *self-similarity*, that is the (statistical) similarity of their parts observed at different scales (Korvin 1992: 4–5).

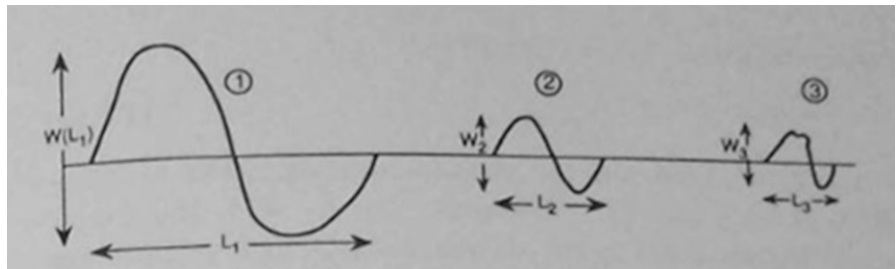
Consider, as an example, the relation between wavelengths L and widths W of meanders of different size (Fig. 1), and assume there exists a functional relation $W = W(L)$ (Eq. 1) between them. Consider three pieces of the same meander, or from different meanders, with wavelengths $L_1 > L_2 > L_3$. Because of the similarity of meanders, one has for any two wavelengths L_a and L_b that $\frac{W(L_a)}{W(L_b)} = f\left(\frac{L_a}{L_b}\right)$ with a yet unknown function $f(x)$, i.e., $W(L_a) = W(L_b)f\left(\frac{L_a}{L_b}\right)$ (Eq. 2), in particular $W_1 = W_3f\left(\frac{L_1}{L_3}\right)$; $W_1 = W_2f\left(\frac{L_1}{L_2}\right)$; $W_2 = W_3f\left(\frac{L_2}{L_3}\right)$ (Eqs. 3 a-c). Combining Eqs. (3. a, b, c) yields $\left(\frac{L_1}{L_3}\right) = f\left(\frac{L_1}{L_2}\right)f\left(\frac{L_2}{L_3}\right)$, that is, letting $\frac{L_1}{L_2} = a$, $\frac{L_2}{L_3} = b$ we see that for all $a, b > 0$ the unknown function $f(x)$ satisfies $f(a \cdot b) = f(a) \cdot f(b)$ (Eq. 4). This is Cauchy’s functional equation (Aczél & Dhombres 1989; Korvin 1992: 73–76) whose only continuous solution (apart from $f(x) \equiv 0$) is $f(x) = x^\alpha$ for some real number α . Writing up Eq. 2 for the case $L_a = L$, $L_b = 1$, $W(1) = w$ we get an APL: $W = wL^\alpha$ (Eq. 5) connecting meander width with wavelength (Korvin 1992: 6). *Note that the value of exponent α cannot be found from scaling arguments!*

Allometric Power Laws, Table 1 Some empirically found APLs in Geosciences (the exponent b is different in each case)

#	Name of the law	Equation	Meaning of the constants	Authors	Discussion of the law
1	Cloud areas and perimeters	$P \propto A^{\frac{D}{2}}$ (Eq. 6)	P = cloud perimeter A = cloud area, $D = 1.35 \pm 0.05$	Lovejoy (1982)	Korvin (1992): 62
2	Destruction of basaltic bodies by high velocity impacts	$M(<s) \propto s^b$ or $N(>m) \propto m^{-b}$ (Eqs. 7a, b)	$M(<s)$ = total mass of fragments of size $< s$ $N(>m)$ = total # of fragments of mass larger than m	Fujiwara et al. (1977)	Korvin (1992): 203
3	Drainage area vs. channel length	$L \propto A_d^{0.5}$ (Eq. 8)	L = channel length A_d = drainage area	Hack (1957)	Korvin (1992): 6
4	Grain size distribution in sedimentary rock	$N(r) \propto \left(\frac{r}{r_0}\right)^{-b}$ (Eq. 9)	$N(r)$ = # of grains longer than r , r_0 reference length	Johnson and Kotz (1970)	Korvin (1992): 203
5	Horton's law of stream numbers	$N_\omega = R_B^{\Omega-\omega}$ (Eq. 10)	N_ω = # of streams of order ω , Ω is the order of the basin, R_B is the bifurcation ratio	Horton (1945)	Scheidegger (1968)
6	Meander width and channel width vs. channel depth	$W_c = 6.8h^{1.54}$ $W_m = 7.44W_c^{1.01}$ (Eqs. 11. a, b)	W_m meander width, W_c channel width, h Channel depth (feet)	Lorenz et al. (1985)	Korvin (1992): 1–3
7	Number of clusters of entrapped oil particles in rock	$n_s \propto s^{-b}$ (Eq. 12)	n_s = # of clusters s = # of oil droplets in a cluster	Sherwood (1986)	Korvin (1992): 193–195
8	Porosity vs. grain radius and eff. hydraulic radius for sandstone and graywacke	$\Phi = 0.5 \left(\frac{r_{\text{grain}}}{r_{\text{eff}}}\right)^{-0.25}$ (Eq. 13)	Φ porosity, r_{grain} grain radius, r_{eff} effective hydraulic radius	Pape et al. (1984)	Korvin (1992): 288
9	Porosity vs. pore-size for sandstone	$\Phi = \left(\frac{l_1}{l_2}\right)^b$ (Eq. 14)	Φ porosity, l_1 and l_2 smallest and largest pore size	Katz and Thompson (1985)	Korvin (1992): 286–287
10	Rivers' width vs. depth (for sinuosity > 1.7)	$W_c \propto h^{1.54}$ (Eq. 15)	W_c bankfull width, h bankfull depth	Leeder (1973)	Korvin (1992): 1
11	Size distribution of caves	$N(l) \propto \left(\frac{l}{l_0}\right)^{-b}$ (Eq. 16)	$N(l)$ # of caves longer than l , l_0 reference length	Curl (1986)	Korvin (1992): 197
12	Size distribution of islands	$Prob(A > a) \propto a^{-b}$ (Eq. 17)	A , a areas	Korčák (1940)	Korvin (1992): 191
13	Undesirable porosity in concrete	$\Phi = \left(\frac{l_1}{l_2}\right)^{1/5}$ (Eq. 18)	Φ porosity, l_1 and l_2 smallest and largest grain size	Caquot (1937)	Korvin (1992): 14

Allometric Power Laws,

Fig. 1 Derivation of the APL between meander width W and wavelength L . (From Korvin 1992: 5)



Several kinds of phenomena can produce, or explain, the observed APLs in earth-processes, earth-formations, and geomaterials. If an earth-process is observed to similarly behave across a range of spatial and/or temporal scales, the created formation usually becomes self-similar, and thus described by APLs. Deterministic chaos, self-organized

criticality (SOC, Hergarten 2002), critical phenomena (as rock fragmentation and failure, Turcotte 1986) are also associated with APLs. Other relevant models leading to APLs include (Mitzenmacher 2004): preferential attachment (applied in transportation geography, geomicrobiology, soil science, hydrology, and geomorphology); minimization of

costs, entropy, etc. (applicable to channel networks and drainage basins); multiplicative cascade models (used in meteorology, soil physics, geochemistry); and DLA (Diffusion Limited Aggregation, Korvin 1992: 349–366, used to model the growth of drainage networks, escarpments, eroded plateaus). However, as we know from geomorphology, similar landforms and other earth-formations might arise through quite different processes. This is the principle of *equifinality* (Bertalanffy 1968): same or similar outcomes can be produced by different processes, and thus there is no way to reconstruct the formative mechanism from the APLs obeyed by the resulting geologic objects.

Examples for Allometric Power Laws in Geosciences

Some examples for APLs are compiled in Table 1. Discussion of these laws and references to their sources are found in Korvin (1992).

Allometric Relations with Modified Power Laws

Empirical data are sometimes better described by approximate power laws instead of the ideal APL $y = ax^k$ because of random errors in the observed values, or their systematic deviation from the power law. Such functional forms, of course, do not have *scaling symmetry* any more. Popular variants are (https://en.wikipedia.org/wiki/Power_law):

Power law with error term: $y = ax^k + \varepsilon$ (Eq. 19).

Broken power law: It is a piece-wise function consisting of one or more power laws, for two power laws, for example:

$$\left. \begin{array}{ll} y \propto x^{\alpha_1} & \text{for } x < x_{th} \\ y \propto x^{\alpha_1 - \alpha_2} x^{\alpha_2} & \text{for } x > x_{th} \end{array} \right\} \text{ (Eq. 20, the threshold } x_{th} \text{ is the cross-over).}$$

Power law with exponential cutoff: $y \propto x^\alpha e^{-\beta x}$ (Eq. 21).

Curved power law: $f(x) \propto x^{\alpha + \beta x}$ (Eq. 22).

Rigout's formula: Rigout (1984, cited in Korvin 1992: 290) described the *Richardson-Mandelbrot plot* (Mandelbrot 1982) for the length of Britain's coastline measured with resolution λ as $L(\lambda) = L_{max} \left\{ 1 + \left(\frac{\lambda}{L_0} \right)^c \right\}^{-1}$ (Eq. 23, L_0 is the *cross-over length*). This approach makes possible to describe two slopes (one seen for small, one for large λ values) on the double-logarithmic plot. The same trick was used by Pape et al. (1984, cited in Korvin 1992: 290) who studied the scaling of pore volume $\Phi(\lambda)$ with resolution λ and found $\Phi(\lambda) = \Phi_{max} \left\{ 1 + \left(\frac{0.5676\lambda}{r_{pore}} \right)^{0.6434} \right\}^{-1}$ (Eq. 24, r_{pore} is pore radius).

Allometric Power Laws for Size Distribution

From among the APLs compiled in Table 1, #7 (number of clusters of entrapped oil particles in rock), #11 (size distribution of caves), and #12 (size distribution of islands, called Korčak's law) are examples for the probability distribution (or the total number, total volume, or total mass) of geologic objects with respect to their size. While Korčak claimed that $b = 0.5$ in his law $Prob(A > a) \propto a^{-b}$ (Eq. 17), careful regression of his data gave $b = 0.48$; Mandelbrot reported $b = 0.65$ for the whole Earth with considerable variation for the different regions, for example, $b = 0.5$ was found for Africa, $b = 0.75$ for Indonesia and North America (Korvin 1992: 191; Mandelbrot 1982). Eq. (17) defines what is called in probability theory the “power law,” “hyperbolic,” or “Pareto” distribution.

The power-law probability distribution $p(x) \propto x^{-\alpha}$ has a well-defined mean over $x \in [1, \infty)$ only if $\alpha > 2$, and it has a finite *variance* only if $\alpha > 3$. For most observed power-law distributions in Geosciences, the exponents α are such that the mean is well-defined but the variance does not exist because typically the exponent falls in the range $2 < \alpha < 3$. It is important to note that for power-law distributions $p(x) = Cx^{-\alpha}$ for $x > x_{min}$ and $\alpha > 1$, the *median does exist*, and $p_{median} = 2^{1/(\alpha-1)} x_{min}$ (Eq. 25). In this case, $p(x)$ can be written in the normalized form $p(x) = \frac{\alpha-1}{\alpha x_{min}} \left(\frac{x}{x_{min}} \right)^{-\alpha}$ (Eq. 26). Its moments can also be computed: $\langle x^m \rangle = \int_{x_{min}}^{\infty} x^m p(x) dx = \frac{\alpha-1}{\alpha-1-m} x_{min}^m$ for $m < \alpha - 1$ (Eq. 27). For $m \geq \alpha - 1$, all moments diverge: in the case $\alpha \leq 2$, the mean and all higher-order moments are infinite; when $2 < \alpha < 3$, the mean exists, but the variance and higher-order moments diverge.

To avoid divergences, power-law distributions can be made more tractable with an *exponential cutoff*, i.e., by using $p(x) \propto x^{-\alpha} e^{-\lambda x}$ (Eq. 28), where the exponential decay $e^{-\lambda x}$ suppresses the power-law-like increase at large values of x . This distribution does not scale for large values of x ; but it is approximately scaling over a finite region before the cutoff (whose size is controlled with the parameter λ).

CAVEAT: The geoscientist should never forget that the straightness of the plotted line is a necessary, but not sufficient, condition for the data to follow a power-law relation, and *many non-power-law distributions will also appear as approximately straight lines on a log-log plot*. For example, *log-normal distributions* are often mistaken for power-law distributions (Clauset et al. 2009, Mitzenmacher 2004, Korvin 1989). Equations and practical advice for fitting a power-law distribution, estimating the uncertainty in the fitted parameters, calculating the p -value for the fitted power-law model can be found in Clauset et al. (2009) and its very useful “Companion paper” (<https://aaronclauset.github.io/powerlaws/>) which provides usable code (in Matlab, Python, R and C++) for the estimation and testing routines.

Concluding Remarks: The Power of Allometric Power Laws

As two case histories will show, APLs are a useful tool in the hand of the practicing or theoretical geoscientist. Perhaps the most beautiful *practical* use of APLs was made in 1985, in a reserve estimation project. J.C. Lorenz and co-workers at the Sandia National Laboratories, Albuquerque, tried to estimate the maximum width of the meanders of an ancient fluvial system (that is the present-day *reservoir extension*) using only the depths of paleostreams measured in wells (their work is reviewed, and referenced, in Korvin 1992: 1–3). Actually, they only used core and log data from a *single well*. They identified 49 fining-upward sand bodies from the meandering fluvial environment of the Mesaverde group, from the histogram of their preserved thicknesses estimated the maximal paleo-channel depth as 3–4 m (10–13 ft). To take into account postdepositional compaction of sand into sandstone, they increased this thickness by 11% which gave 3.3–4.4 m (11–13 ft) depth at the time of deposition. Then they substituted the decompacted channel depth into the empirical APL $W_c = 6.8h^{1.54}$ Eq. (11a, see Law #6 in Table 1). The formula gave an estimated channel width $W_c = 45 - 67m(150 - 220ft)$, plugging this into the APL $W_m = 7.44W_c^{1.01}$ (Eq. 11.b) yielded the predicted meander belt width $W_c = 350 - 520m(1,100 - 1,700ft)$.

Many empirical laws of the physics of sedimentary rock describe some physical property of the rock as a *power of porosity*, as, e.g., the conductivity σ of fluid-filled rock satisfies $\sigma \propto \Phi^{m_1}$ (Archie's law, Eq. 29), for some rocks and some porosity ranges the permeability k satisfies $k \propto \Phi^{m_2}$ (Eq. 30). Such power laws are not considered *allometric power laws* (APL) because while the dependence is power-like, the measured quantity is a property of the whole rock, and not of its size. To find a heuristic explanation of such empirical findings, one must keep in mind that very often we can assume hierarchical (or fractal) structures and processes, and hidden APLs lying behind such power law-like behavior. Consider, e.g., the empirically observed $k \propto \Phi^{m_2}$ relation (Eq. 30) and try to explain it by scaling arguments (Korvin 1992: 292–294). According to the Kozeny-Carman relation: $k = \frac{1}{b} \Phi^3 \cdot \frac{1}{S_{spec}^2} \cdot \frac{1}{\tau^2}$ (Eq. 31), where k is permeability, Φ porosity, S_{spec} specific surface of the pore space, τ tortuosity, and b a constant depending on the shape of the pore-throats available for the flow. Denote by ε the smallest pore size, and consider a piece of rock of fixed size $R, \gg \varepsilon$. If the total pore volume in an R -sized rock scales with fractal dimension D_p , then $\Phi_{total} \propto \left(\frac{R}{\varepsilon}\right)^{D_p}$, and porosity scales with ε as $\Phi \sim \frac{\Phi_{total}}{R^3} \propto \left(\frac{R}{\varepsilon}\right)^{D_p} \cdot R^{-3} = \frac{R^{D_p-3}}{\varepsilon^{D_p}} \propto \varepsilon^{m_4}$ (Eq. 32, in this heuristic consideration $m_1, m_2, \dots$ are the corresponding exponents). If the pore surface has fractal dimension D_s , the specific surface of a piece of rock of size R scales as $S_{spec} \propto \left(\frac{R}{\varepsilon}\right)^{D_s} / R^3 = \frac{R^{D_s-3}}{\varepsilon^{D_s}} \propto \varepsilon^{m_5}$ (Eq. 33). Next, consider tortuosity $\tau (= \text{hydraulic path/straight line path})$ over

distance R . If R is not too great, the hydraulic path is a *planar curve*, that is its fractal dimension is $D_s - 1$ (by Mandelbrot's rule, Mandelbrot 1982: 365, Korvin 1992: 125 Eq. 2.3.1.7) and its length scales as $L_{hydr} \propto \left(\frac{R}{\varepsilon}\right)^{D_s-1}$, so tortuosity satisfies $\tau = \frac{L_{hydr}}{R} \propto \left(\frac{R}{\varepsilon}\right)^{D_s-1} \cdot R^{-1} \propto \varepsilon^{m_6}$ (Eq. 34). Plugging Eqs. (32, 33, 34) into the right-hand side of Eq. (31), we find that permeability also scales as a power of ε : $k \propto \varepsilon^{m_7} = (\varepsilon^{m_4})^{(m_7/m_4)} \propto \Phi^{(m_{10}/m_4)}$ (Eq. 35) ($m_4 = -D_p \approx 3$, certainly $m_4 \neq 0$) which explains the power-like relation $k \propto \Phi^{m_2}$ in (Eq. 30). Thus, fractal geometry, assumption of self-similar pores, and simple scaling arguments yield a heuristic explanation for the empirical power-law relation between porosity and permeability over a limited range of specimen and pore size. Because of the principle of *equifinality*, this – of course – is only one of the many possible explanations.

Cross-References

- Chaos in Geosciences
- Flow in Porous Media
- Fractal Geometry in Geosciences
- Frequency Distribution
- Grain Size Analysis
- Lognormal Distribution
- Pore Structure
- Porosity
- Porous Medium
- Probability Density Function
- Rock Fracture Pattern and Modeling
- Scaling and Scale Invariance
- Singularity Analysis
- Statistical Rock Physics

References

- Aczél J, Dhombres J (1989) Functional equations in several variables. Cambridge University Press, Cambridge
- Caquot A (1937) Rôle des matériaux inertes dans le béton. Mem Soc Ing Civils France:562–582
- Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. SIAM Rev 51(4), 661–703. See also its companion paper: Power-law Distributions in Empirical Data <https://aaronclauset.github.io/powerlaws/>
- Curl RL (1986) Fractal dimensions and geometries of caves. Math Geol 18(8):765–783
- Fujiwara A, Kamimoto G, Tsukamoto A (1977) Destruction of basaltic bodies by high-velocity impact. Icarus 31:277–288
- Hack JT (1957) Studies of longitudinal stream-profiles in Virginia and Maryland. US Geol Surv Prof Pap 294B:45–97
- Hergarten S (2002) Self-organized criticality in earth systems. Springer, New York
- Horton RE (1945) Erosional development of streams and their drainage basins: hydrophysical approach to quantitative morphology. Bull Geol Soc Am 56:275–370

- Johnson NJ, Kotz S (1970) Continuous univariate Distributions-2. Houghton Mifflin, Boston
- Katz AJ, Thompson AH (1985) Fractal sandstone pores: implications for conductivity and pore formation. *Phys Rev Lett* 54:1325–1328
- Korčák J (1940) Deux types fondamentaux de distribution statistique. *Bull Inst Int Stat* 30:295–299
- Korvin G (1989) Fractured but not fractal: fragmentation of the Gulf of Suez basement. *PAGEOPH* 131(1–2):289–305
- Korvin G (1992) Fractal models in the earth sciences. Elsevier, Amsterdam
- Leeder MR (1973) Fluvial fining-upwards cycles and the magnitude of palaeochannels. *Geol Mag* 110(3):265–276
- Lorenz JC, Heinze DM, Clark JA (1985) Determination of width of meander-belt sandstone reservoirs from vertical downhole data, Mesaverde group, Piceance Creek Basin, Colorado. *AAPG Bull* 69(5):710–721
- Lovejoy S (1982) Area-perimeter relation for rain and cloud areas. *Science* 216(4542):185–187
- Mandelbrot B (1982) The fractal geometry of nature. W.H. Freeman & Co., New York
- Mitzenmacher M (2004) A brief history of generative models for power law and lognormal distributions. *Internet Math* 1(2):226–251
- Pape H, Riepe L, Schopper JR (1984) The role of fractal quantities, as specific surface and tortuosities, for physical properties of porous media. *Part Part Syst Charact* 1(1–4):66–73
- Rigout JP (1984) An empirical formulation relating boundary lengths to resolution in specimens showing non-ideally fractal dimensions. *J Microsc* 133:41–54
- Scheidegger AE (1968) Horton's law of stream numbers. *Water Resour Res* 4(3):655–658
- Sherwood JD (1986) Island size distribution in stochastic simulations of the Saffman-Taylor instability. *J Phys A Math Gen* 19(4):L195–L200
- Stevens CF (2009) Darwin and Huxley revisited: the origin of allometry. *J Biol* 8:14
- Turcotte DL (1986) Fractals and fragmentation. *J Geophys Res* B 91: 1921–1926
- von Bertalanffy L (1968) General systems theory, foundations, development, applications. George Braziller, New York

Argand Diagram

James Irving¹ and Eric P. Verrecchia²

¹Institute of Earth Sciences, University of Lausanne, Lausanne, Switzerland

²Institute of Earth Surface Dynamics, University of Lausanne, Lausanne, Switzerland

Synonyms

Argand plane; Cole-Cole plot; Complex plane; Gauss plane; Impedance diagram; Zero-pole diagram; Zero-pole plane; z-plane

Definition

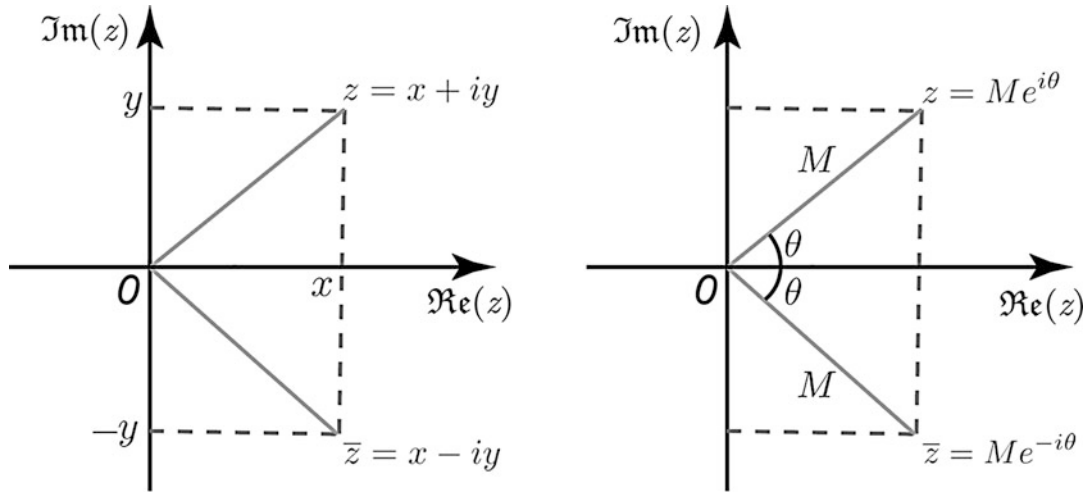
The Argand diagram denotes a graphical method to plot complex numbers, expressed in the form $z = x + iy$, where (x, y) are used as coordinates and plotted on a plane with two orthogonal axes, the abscissa representing the real axis and the ordinate the imaginary axis.

Argand Diagram

The Argand diagram is a way to plot a complex number in the form of $z = x + iy$, using the ordered pair (x, y) as coordinates and where the constant i represents the imaginary unit, i.e., $\sqrt{-1}$. The Argand diagram (or plane) is therefore defined by two orthogonal axes where the abscissa refers to a real axis and the ordinate to an imaginary axis.

Historical Origin of the Argand Diagram

The family name attached to the Argand diagram designates that this geometric representation of complex numbers is credited to Swiss mathematician Jean-Robert Argand, who was born in Geneva in 1768 and died in Paris in 1822. Indeed, this graphical method has been detailed in Argand (1806). Nevertheless, Howarth (2017) suggested that the plot was first introduced by Caspar Wessel (1799), a Norwegian-Danish mathematician. Wessel (1745–1818) presented his work about the analytical representation of the direction of numbers to the Academy of Science in Copenhagen on March 10, 1797; this contribution was printed in 1798 and finally published in 1799. Nevertheless, Valentiner (1897), in the first preface of the French translation of Wessel's essay, noted that Carl Friedrich Gauss (1777–1855) arrived at the same idea in the same year (1799), but without citing any specific source. This remark led some researchers to refer to the Argand plane as the Gauss plane. In the same preface, Valentiner also acknowledged a British mathematician, John Wallis (1616–1703). It seems that Wallis proposed a similar approach, even earlier than Wessel, in his "Treatise of Algebra," published in London in 1685. Valentiner noted that Wessel's essay includes "a theory as complete as Argand's, and in my opinion, more correct [...]" (Valentiner, 1897, p. iii). Regarding the various synonyms for the Argand diagram, such as the complex z-plane, zero-pole diagram, or zero-pole plot, they were introduced during the twentieth century, mostly by mathematicians and geophysicists. For example, according to Howarth (2017), the concept of "complex z-plane," with z referring to $z = x + iy$, was coined by the American mathematician William Fogg Osgood at the turn of the twentieth century (Osgood 1901) and used later by engineers studying signal processing as the "zero-pole diagram" and "zero-pole plot." In geophysics, the concept of "zero-



Argand Diagram, Fig. 1 Left, representation of an imaginary number (z), and its conjugate ($\bar{z}$), as ordered pairs (x, y) , and $(x, -y)$, on an Argand diagram, respectively. The x-axis is the real axis, whereas the

y-axis refers to the imaginary axis. Right: representation of the same numbers in terms of the modulus (M) and the angle (θ) as polar coordinates (M, θ)

pole” usually refers to the z-transform of a discrete impulse response function when it reaches infinite values (Buttkus 2000).

The Argand Diagram: A Geometric Plot

The Argand diagram is a kind of Cartesian plane upon which an imaginary number, $z = x + iy$, is represented with its real part along the x-axis and its imaginary part along the y-axis. It provides a geometrical way of locating imaginary numbers (Fig. 1, left). Therefore, any complex number can be recognized by the coordinates of a unique point in the diagram according to its ordered pair (x, y) . Any complex number with a null real part ($z = 0 + iy$) will lie on the vertical imaginary axis, whereas any real number ($z = x + i0$) will lie on the real horizontal axis. The conjugate of a complex number, i.e., $\bar{z} = x - iy$, can be easily represented by its symmetric location along the real x-axis (Fig. 1, left), i.e., by the $(x, -y)$ coordinates on the Argand diagram. The distance from the origin $(0, 0)$ to a given point (x, y) on the Argand plane is known as the modulus of z (or its absolute value), and is usually denoted by r , M , or $|z|$. The modulus is easily calculated as $|z| = \sqrt{x^2 + y^2}$.

Complex numbers can also be expressed in polar form on the Argand diagram, i.e., using the modulus (M) and the angle (θ) (also called the argument $\text{Arg}(z)$) as polar coordinates, the angle (θ) being measured along the positive direction of the real x-axis. Consequently, the ordered pair (x, y) is replaced by the pair (M, θ) as illustrated in Fig. 1 (right). The ordered pair (x, y) can now be calculated as:

$$x = M \cdot \cos(\theta) \text{ and } y = M \cdot \sin(\theta) \quad (1)$$

Therefore, $z = M \cdot \cos(\theta) + i \cdot M \cdot \sin(\theta)$. Using Euler’s formulae expressing the relationship between trigonometric functions, e (Euler’s number) and i ($\sqrt{-1}$ number), i.e.,

$$\cos(\theta) = \frac{e^{i\theta} + e^{-i\theta}}{2} \text{ and } \sin(\theta) = \frac{e^{i\theta} - e^{-i\theta}}{2i} \quad (2)$$

we arrive at

$$\begin{aligned} z &= M(\cos(\theta) + i \cdot \sin(\theta)) = M \left[\frac{e^{i\theta} + e^{-i\theta}}{2} + i \cdot \frac{e^{i\theta} - e^{-i\theta}}{2i} \right] \\ &= M \cdot e^{i\theta} \end{aligned}$$

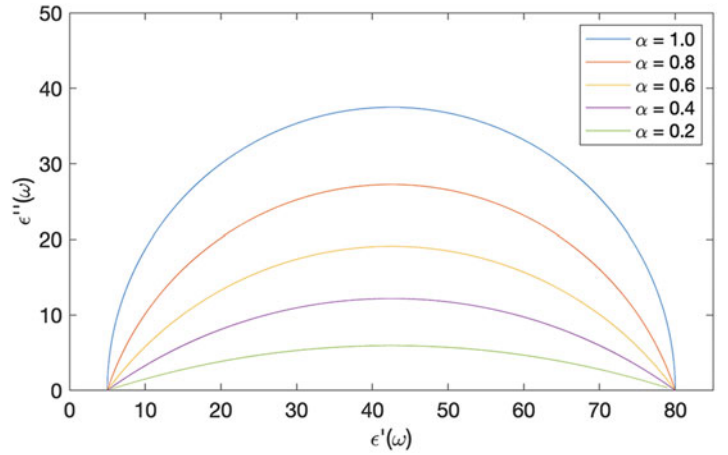
If $z = M \cdot e^{i\theta}$, then $\bar{z} = M \cdot e^{-i\theta}$. Following Eq. (1):

$$\frac{y}{x} = \frac{\sin \theta}{\cos \theta} = \tan \theta \Rightarrow \theta = \arctan\left(\frac{y}{x}\right) \quad (3)$$

Some Applications of the Argand Diagram in Geosciences

In many areas of the geosciences, complex numbers play an important role, and the Argand diagram can serve as a useful tool for data analysis. A wide range of Earth materials, for example, possess dispersive material properties, whereby the steady-state material response to a sinusoidal forcing depends on the frequency at which the forcing is applied. Due to various relaxation phenomena, both the amplitude and phase of the response will vary with frequency, meaning that the corresponding physical property, which links the response and forcing through a constitutive equation, takes the form of a complex and frequency-dependent variable in the Fourier domain. One commonly used formulation to describe such behavior is the Cole-Cole model (Cole and Cole 1941). Originally proposed to describe the relaxation behavior of the dielectric permittivity, it is given by

Argand Diagram, Fig. 2 Plot of the imaginary versus real part of the complex dielectric permittivity corresponding to the Cole-Cole model, for different values of α . Frequency was varied from 1 Hz to 10^{20} Hz, with fixed parameters $\varepsilon_\infty = 5 \text{ F.m}^{-1}$, $\varepsilon_0 = 80 \text{ F.m}^{-1}$, $\tau = 10^{-10} \text{ s}$



$$\varepsilon'(\omega) - i\varepsilon''(\omega) = \varepsilon_\infty + \frac{\varepsilon_s - \varepsilon_\infty}{1 + (i\omega\tau)^\alpha} \quad (4)$$

where $\varepsilon'(\omega)$ and $\varepsilon''(\omega)$ are the real and imaginary parts of the permittivity [F.m^{-1}], ε_∞ and ε_s are the high- and low-frequency limiting values of the real part, ω is the angular frequency [s^{-1}], τ is a relaxation time constant [s], and α is a unitless time-constant distribution parameter that varies between 0 and 1. Plots on an Argand diagram of complex permittivity measurements made at different frequencies provide insight into the nature of the relaxation (Fig. 2). For instance, $\alpha = 1$ in the Cole-Cole equation describes a Debye-type relaxation response involving a single polarization mechanism, and the complex values plot along a semi-circular arc whose intersections with the real axis correspond to ε_∞ and ε_s . Lower values of α result in Argand plots with the same real-axis intersections but lower complex amplitudes, indicating a distribution of relaxations within the considered frequency range. Multiple polarization mechanisms whose relaxations are well separated in frequency may appear as more than one arc. Similar analyses using the Cole-Cole and other related models have been performed for the electrical conductivity as well as for various elastic moduli.

Argand diagrams have also proven highly useful in the context of linear time-invariant (LTI) system identification and digital filter design, for which there are many important applications in the geosciences. In the z -transform domain, the input $X(z^{-1})$, output $Y(z^{-1})$, and impulse response $H(z^{-1})$ of a discrete LTI system are related as follows:

$$Y(z^{-1}) = H(z^{-1}).X(z^{-1}) \quad (5)$$

where the z -transform of time series $f(t)$ with sampling interval T is given by

$$F(z^{-1}) = \sum_{n=-\infty}^{\infty} f(nT)z^{-n} \quad (6)$$

Inspection of the complex roots of the z -transform polynomials in Eq. (5) in the Argand plane can provide important information regarding the system impulse response. De Laine (1970), for example, suggested that the unit hydrograph for a catchment rainfall-runoff model could be estimated via comparison of the roots of the runoff polynomials corresponding to different storm events. Runoff roots having similar locations in the Argand plane were assumed to correspond to the catchment system and not the input. Turner et al. (1989) similarly suggested that roots for a single runoff time series that fell along a “skew circle” in the Argand plane could be attributed to the unit hydrograph and thus allow for its identification. In seismic data analysis, examination of the roots of $H(z^{-1})$ also has important implications for the stability of inverse filtering operations such as deconvolution; if the roots of the wavelet polynomial under consideration lie outside the unit circle $|z| = 1$, the deconvolution operation will be unstable. Further, if one considers of the LTI system as a rational function in the z -transform domain, i.e.,

$$H(z^{-1}) = \frac{A(z^{-1})}{B(z^{-1})} \quad (7)$$

where $A(z^{-1})$ and $B(z^{-1})$ correspond to moving-average and feedback operations, respectively; the positions of the roots of $A(z^{-1})$ (zeros) and those of $B(z^{-1})$ (poles) in the complex plane with respect to the unit circle can be used to understand the system frequency response as well as design effective digital filters for environmental data.

Summary

The Argand diagram provides a means of plotting a complex number in the form of $z = x + iy$, using the ordered pair (x, y) as coordinates and where the constant i represents the imaginary unit, i.e., $\sqrt{-1}$. The Argand diagram (or plane) is

therefore defined by two orthogonal axes, a real one (abscissa) and an imaginary one (ordinate). In many areas of the geosciences, complex numbers play an important role, and the Argand diagram can serve as a useful tool for data analysis. Examples are provided regarding dispersive material properties, linear time-invariant (LTI) system identification, and digital filter design, as well as seismic data analysis.

Bibliography

- Argand JR (1806) Essai sur une manière de représenter les quantités imaginaires dans les constructions géométriques. Mme Veuve Blanc Edn, Paris
- Buttkus B (2000) Spectral analysis and filter theory in applied geophysics. Springer, Berlin
- Cole KS, Cole RH (1941) Dispersion and absorption in dielectrics. *J Chem Phys* 9:341–351
- De Laine RJ (1970) Deriving the unit hydrograph without using rainfall data. *J Hydrol* 10:379–390
- Howarth RJ (2017) Dictionary of mathematical geosciences. Springer-Nature, Cham
- Osgood WF (1901) Note on the functions defined by infinite series whose terms are analytic functions of a complex variable; with corresponding theorems for definite integrals. *The Annals of Mathematics* 31:25–34
- Turner JE, Dooge JCI, Bree T (1989) Deriving the unit hydrograph by root selection. *J Hydrol* 110:137–152
- Valentiner H (1896) Première préface. In: Wessel C Essai sur la représentation analytique de la direction, French translation, Host, Copenhagen, pp III–X
- Wallis J (1685) Treatise of algebra. John Playford printer, London
- Wessel C (1799) Om Directionens analytiske Betegning et Forsog, anvendt fornemmelig til plane og sphæriske Polygoners Oplosning. *Nye Samling af det Kongelige Danske Videnskabernes Selskabs Skrifter* 5:469–518

Artificial Intelligence in the Earth Sciences

Norman MacLeod
Department of Earth Sciences and Engineering, Nanjing University, Nanjing, Jiangsu, China

Definition

There are many definitions of artificial intelligence (AI), but perhaps it is most commonly understood to mean any machine, computer, or software system that accepts information and performs the cognitive functions humans associate with intelligence, such as learning and problem-solving. Part of the difficulty in defining AI more precisely is that no common, universally accepted, formal definitions of the core concepts needed to define human intelligence exist (e.g., thinking, intelligence, experience, consciousness).

Regardless, this difficulty has not prevented philosophers, artists, writers, filmmakers, scientists, engineers, and/or earth scientists from imagining what an artificially intelligent device might be like, how it might be constructed, what it might be used to do, and what dangers it might pose.

Introduction

Few can have failed to notice the plethora of news bulletins, editorials, perspectives, symposia, conferences, course offerings, lectures, and the like that, over the past few years, have announced the imminent arrival of artificially intelligent machines. These are portrayed as being designed to assist with the tasks traditionally assigned to human workers in fields that, thus far, have proven resistant to automation. While not muted (so far) as a replacement for scientists, artists, or senior managers, presumably because of the creativity required by these fields, there is widespread agreement among employers that many repetitive tasks could, and perhaps should, be given over to automation if possible. In addition to creativity, scientific research often involves complex, laborious, and time-consuming searches through datasets, library references, catalogues, image sets, and, increasingly, online resources for bits of information critical to the resolution of a particular problem at hand or hypothesis test. In principle, many researchers would welcome access to automated intelligent agents that could conduct such searches quickly, systematically, and objectively, thus freeing themselves, their colleagues, their assistants, and their students to focus on tasks better suited to their training, knowledge, and core interests. The Massachusetts Institute of Technology's recent announcement that a team of its medical researchers used a deep-learning system named "Hal" successfully to find new generalized antibiotic molecules that kill a number of the most common disease-causing bacteria after screening more than 107,000,000 candidate molecules represents an interesting case in point, especially insofar as the material discovered often exhibited a form structurally divergent from conventional antibiotics (Stokes et al. 2020). Consequently, current debates over artificial intelligence (AI) techniques and technologies are as relevant – though perhaps not quite as threatening – to research scientists as they are to professionals in other areas of human activity.

Curiously, the earth science community appears to be of two minds regarding AI. On one hand, the use of computer software designed to process and identify patterns in earth science data without having to be programmed explicitly to do so is old hat in many fields, so ubiquitous it hardly bears comment. On the other hand, some recent articles have claimed earth science research is either unusually susceptible to the misuse of AI (e.g., Ebert-Uphoff et al. 2019) or focuses on problem classes inherently resistant to AI-based

approaches (e.g., Anonymous 2019; see also Reichstein et al. 2019). Instead, these critics advocate a vaguely defined partnership between traditional “physical process models” and artificial expert systems. At present, the focus of most new AI applications in the earth sciences has centered on machine learning systems and software, especially the newer “deep-learning” artificial neural network architectures. However, machine learning is only one aspect of the much larger artificial intelligence field. In order to determine whether AI has a past, present, or future in earth science research, a more general review is required.

What Is Artificial Intelligence?

The earliest written record of an AI device was the character of ΤΑΛΩΝ or Talon (also Talos) which was a giant robot made of bronze, gifted to Europa by Zeus, that circled the island triannually and threw stones at approaching ships to protect the island from invaders. Modern variations of the Talon myth can be found in Homer’s description of Hephaestus’ serving devices in the *Iliad*, the Jewish parable of the Prague golem, Mary Shelly’s *Frankenstein*, Arthur C. Clark’s HAL from 2001: *A Space Odyssey*, Ridley Scott’s dystopian epic *Blade Runner* (based on the Philip K. Dick novel *Do Androids Dream of Electric Sheep?*), and latter installments of the *Terminator* film franchise.

The first computational steps toward realizing artificial intelligence were taken by Alan Turing, the mathematician and computer-design pioneer who understood that binary arithmetic could be used to perform any conceivable act of mathematic deduction. This insight led to the Church-Turing thesis (Church 1936; Turing 1937, 1950) which was a key advance that ultimately enabled digital computers to be designed. At the same time, several researchers recognized the similarity between Turing’s switch-based computational system and the operation of mammalian neurons which are the fundamental units of the human nervous system. This transfer of insight across disciplines led to Warren McCulloch and Walter Pitt’s (1943) proposal of a design for a Turing-complete artificial neuron and the possibility of constructing a complete, artificial, electronic brain. In the wake of these mid-twentieth-century developments, the MIT and Stanford University computer scientist John McCarthy coined the term “artificial intelligence” to describe machines that can perform tasks characteristic of human intelligence such as planning, understanding language, recognizing objects and sounds, learning, and problem-solving (McCorduck 1979). McCartney’s list of the tasks or capabilities that define AI operations is still reflected in the major components of this field (Fig. 1).

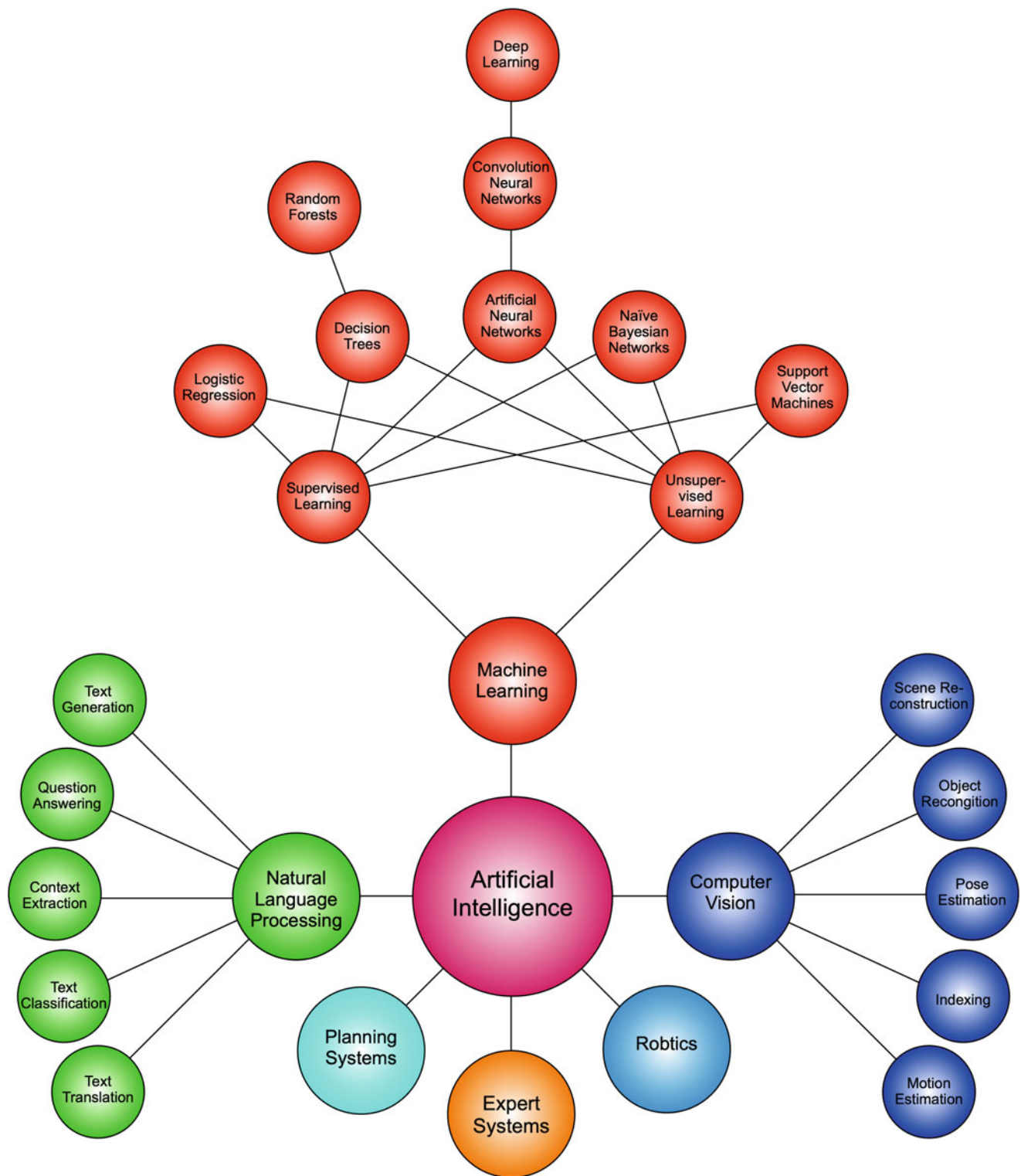
Development and Activity Domains

Two broad classes of AI are recognized: narrow AI and general AI. Narrow AI includes all devices and systems that have some characteristics of human intelligence. Systems that exhibit narrow AI have existed at least since the late 1950s (e.g., Samuel 1959), and it is devices of this sort that are being referred to in contemporary media reports of specific, realized AI applications. The list of fields in which narrow AI systems have, or could be, applied is very long indeed.

No general AI system currently exists, in part owing to lack of agreement of the characteristics such a system would need to possess in order to be so designated, but also owing to the broad assumption that genuine general AI system would need to be self-aware or exhibit consciousness. The basis of fear over the capabilities of a general AI system can be traced to the assumption that a conscious system would seek to establish local or general dominance in the sense many social biological species do. But as Pinker has noted, dominance is usually associated with mammalian male behavior patterns and is derived from genetic strategies that have no necessary correspondence to AI system development goals (<https://bigthink.com/videos/steven-pinker-on-artificial-intelligence-apocalypse?jwsourc=cl>). Conscious AI systems could, just as easily, adopt a female perspective and prioritize problem-solving along with social coherence.

Six activity domains are commonly recognized by AI researchers. These are identified principally by the problems associated with each rather than by the technologies or algorithms that happen to be associated with attempts to address these problems at any given time. Activity has waxed and waned within and between each of these domains over the course of the past 75 years as new approaches were tried and met either with success or failure. Indeed, the methods employed in the current most active domain – machine learning – are being applied, in whole or in part, to all of the other domains at present. Since this domain is the subject of a separate article in this volume, only a very abbreviated description of its core concepts and scope will be provided here. However, it is important to note that the inherent topical disparity that exists within this field has, at certain times, been as much a hindrance as a help to its overall development.

In her history of AI, Pamela McCorduck (1979) noted that, soon after its formulation, AI shattered into subfields, the practitioners of which “would hardly have anything to say to each other” (p. 424). This self-inflicted breach of cohesion and collegiality, along with a tendency to overpromise and under-deliver to funders, exposed AI research to criticisms that led to a reduction in both academic interest and extra-academic support funding over two successive intervals. The first of these “AI winters” extended from 1974 to 1980 and was precipitated by the very damaging US Automatic Language Processing Advisory Committee (ALPAC) report into



Artificial Intelligence in the Earth Sciences, Fig. 1 Schematic representation of the primary domains of contemporary artificial intelligence research and applications. A non-exhaustive listing of some

subordinate activity domains has been provided as a rough indication of domain diversity as well as relative activity levels

prospects for machine translation (Pierce et al. 1966) and the report of Sir James Lighthill (1973) into the state of AI research in the UK. The second AI winter, from 1987 to 1995, was caused by the collapse of the commercial market for AI-based, specialist expert system computers, especially those based on the LISP processor (LISP machines) that employed a syntax close to that of natural language to manipulate source code as if it were a data structure. Arguably, these difficulties for the whole of the AI field were not addressed successfully until the advent of convolution-based artificial neural networks in the late 1990s (e.g., LeCun et al. 1998). It is from this point that the present revival of AI's fortunes has stemmed.

Expert Systems

Expert systems are computer systems that accept information, usually in the form of alpha-numeric data, and employ various processing strategies to address problems via autonomous reasoning based on relations known, or suspected, to exist in a set of data rather than by passing the data through a set of deterministic mathematical procedures. Expert systems were among the first successful AI applications, proliferating commercially in the late 1970s through 1980s via sales of dedicated computer systems based on the LISP family of computer languages, all of which are based on Alonso Church's (1936) development of lambda calculus. Interestingly, LISP is the second oldest computer language, the first versions of which were released in 1958, just 1 year after the release of the first FORTRAN compilers (Wexelblat 1981).

Typically expert systems are organized into a knowledge base and an inference engine. The knowledge base contains definitions, facts, rules, and systems of weights that govern how the rules are applied in particular situations. These rules can be provided by human experts based on their personal or collective knowledge or transferred from the inference engine based on the system's experience processing datasets. In this way, the knowledge base can grow and improve the system's performance giving the system, overall, a degree of machine learning capability. The inference engine applies the knowledge base to data and, by virtue of relations identified within and between datasets, can infer new facts/rules that, in turn, can be passed back to the knowledge base. Classic expert system designs represent knowledge, for the most part, via specification of if-then rules. In this sense, expert system designs resemble aspects of decision trees.

Many information-technology historians date the introduction of expert systems to 1965 and the Stanford Heuristic Programming Project whose original application was to identify novel organic molecules based on their mass spectra with a rule set derived from human expert knowledge of organic chemistry. In 1982, the Synthesis of Integral Design (SID) system was the first system to generate a set of logical routines sufficiently extensive to outperform expert logicians. Despite

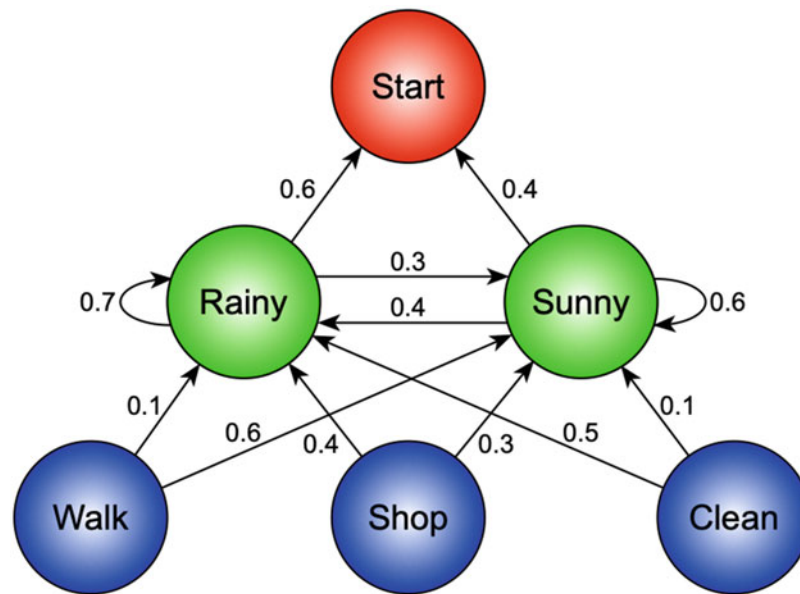
these milestones, however, the performance predictions made by advocates of expert systems failed to be realized and were recognized to have fallen short of expectations by the late 1980s. The reasons behind the inherent limitations in expert system designs had been studied and explained by Karp (1972). But in the early optimism over AI his concerns were discounted or ignored. Nevertheless, failure to meet expectations and deliver on commercial investments in expert systems caused a collapse of the market for specialist systems by the late 1980s, a situation from which classical LISP-based expert systems never really recovered, though aspects of expert system design continue as codified algorithms in many business-analysis applications. Hayes-Roth et al. (1983) published an extensive summary and classification of expert system designs which remains a standard reference.

Natural Language Processing

The very large field of natural language processing (NLP) includes all aspects of the theory and application of the ability of humans and computers to interact with one another using natural spoken or written languages. While machines that speak and understand what is being said when spoken to have been a staple of science fiction for over a century, sophisticated and effective NLP applications are familiar to most people today in the form of personal computer assistants such as Apple Corporation's *Siri*, Google's *Google Assistant*, and Amazon's *Alexa*, in addition to a plethora of online automated technical service and customer support systems that respond to typed text. This field represents the intersection between linguistics, computer science, information engineering, and AI.

Alan Turing (1950) regarded NLP as a primary goal of AI in the form of his "imitation game" – later referred to as the "Turing test" – for the degree to which any machine can be judged to demonstrate intelligent behavior. Turing's test has been criticized for confusing the concepts of thinking and acting, but it remains a well-known, popular, and (so) important AI benchmark. This concept probably inspired, at least in part, the so-called Georgetown experiment a few years after Turing's article appeared when, in 1954, an IBM 701 mainframe computer successfully translated 60 Russian sentences into English based on a dictionary of 250 words and a knowledge base of six rules (Hutchins 2004). Not only did this demonstration receive wide media coverage, it stimulated government and commercial interest in NLP, and in AI generally, prompting some to predict generalized systems for language translation could be produced by 1960. When this goal had not been achieved by the mid-1960s, it prompted a withdrawal of US government funding, precipitating the first AI winter.

Natural language processing systems were based originally on expert-determined rule sets. This design was replaced in the late 1980s by a statistical approach in which



Artificial Intelligence in the Earth Sciences, Fig. 2 Example of a typical hidden Markov model. In this situation, the task is to infer the state of the weather by collecting data on what a friend who lives in another city is doing. Since you cannot observe the weather there, that layer is hidden. But if probabilities can be assigned to various weather-

dependent activities, its state can be inferred by collecting information about the friend's activity. If the friend spent the day shopping, there's a 40% chance it was raining. But if they went for a walk, there's a 60% chance it was sunny. (Redrawn from an original by Terrance Honles: see <https://commons.wikimedia.org/wiki/File:HMMGraph.svg>)

rules were learned by systems analyzing large sets of documents and could be applied in a probabilistic manner that incorporated information from context. Modern NPL systems are quite sophisticated and are gaining in accuracy, and popularity, with each passing year. Amazon's *Alexa* is representative of the best of the current crop of NPL devices. *Alexa* records ambient sounds (= listens) continuously and subdivides spoken words into individual phonemes. When the "wake" word "Alexa" is detected, the statement that follows is shunted to a very large database of word pronunciations to find which words most closely correspond to the combinations of sounds. A second (huge) database of likely phrases is then searched to infer the most likely word order using a hidden Markov model (Fig. 2; also see Rabiner 1989). Once the statement has been reconstructed, the *Alexa* system searches the statement for important or key terms that signal instructions to perform particular tasks. All these searches, inferences, and reconstructions are performed in real time using cloud computing. Responses to spoken commands are formulated in the same manner, but in reverse order.

Owing to the combination of deep learning with NLP, along with the number of times they are used by system subscribers, the performance of these systems is improving rapidly. Moreover, the recent release of 5G network technology will enable NPL systems such as *Siri*, *Google Assistant*, *Alexa*, and others to communicate with an ever-increasing array of devices so they can, effectively, serve as intermediaries for two-way human-device communication. This

capability is set to have an extremely wide range of applications, including scientific research.

Machine Learning

Without question, machine learning (ML) is the AI activity domain that has received the most attention over the last two decades, driven largely by a revival of interest in artificial neural networks (Only the basics of machine learning will be covered in this section. The reader is directed to the companion article in this volume for additional information.). Formally ML is the study and application of algorithms that enable computer systems to learn and improve performance in specific task areas without being programmed explicitly to do so. This is an algorithmically diverse and mathematically complex field that, to this author's way of thinking, has its conceptual origins in multivariate data analysis, especially eigenvector-based methods such as principal component analysis (ordination, dimensionality reduction) and multi-group discriminant analysis (classification). The techniques and search strategies employed by current machine learning systems are far removed from these origins, and, given adequate data, their performance exceeds those delivered by these original and more primitive approaches by a very substantial margin. But on a conceptual level the goals are largely the same.

Key to understanding the core structure of ML is the difference between unsupervised and supervised learning. Both approaches require a set of representative data used to

train algorithms to identify regularities and differences among subgroups contained therein. Like all statistical samples, these training data must (ideally) constitute an unbiased, representative summary of patterns present in the parent population of interest, at least at the level of resolution necessary to address the problem at hand. In most cases, owing to the complexity of the machine learning algorithms and the number of variables that must be estimated, the size of the training dataset must be large; uncommonly large in many instances though this depends on the distinctiveness of the subgroups included in the population of interest, or lack thereof.

In any unsupervised learning problem, only the training dataset is input and the task is twofold: (1) determination how many different groups of observations are present in the training data and (2) identification of characteristic patterns of variation within and between these subgroups that will allow unknown observations to be associated with the correct subgroup. Under a supervised learning design, the classification, or pattern of group membership, present in the training data is known before training is undertaken; this information is supplied to the system at the outset of training, and only the second learning task is performed. Under both approaches, assessments of training performance and/or evaluations of the training dataset are made post hoc, preferably via reference to a new or validation dataset whose members were not part of the training process (If necessary, it is permissible to approach the estimation of post-training performance using a jack-knifed testing strategy though this cannot be considered as rigorous a test and might be impractical owing to the time required to train ML systems such a large number of times in succession.). Some ML systems employ other learning designs (e.g., semi-supervised learning, reinforcement learning, self-learning, feature learning, transfer learning), but most adopt either a supervised or unsupervised approach with the former being more common at present. A very large and diverse array of algorithm families are currently being employed in ML systems (e.g., logistic regression, Bayesian networks, artificial neural networks, decision trees, self-organizing maps, deep-learning networks). Each has advantages and disadvantages in particular situations, a discussion of which is beyond the scope of this article.

Computer Vision

This activity domain involves more than locating and identifying images, or parts thereof, tasks that properly fall into the domain of machine learning. Rather, computer vision involves imbuing machines with the capacity to understand image-based information and the ability to make decisions or take actions based on that understanding. Specific skills associated with this domain include scene reconstruction, event detection, tracking, object recognition, pose estimation, motion estimation, indexing, and image restoration along with AI.

Since humans have particularly well-developed visual systems, this has been one of the more attractive of the AI activity domains. Interest in computer vision grew out of research into image processing and was funded originally by national militaries interested in the problem of automatically detecting and identifying hostile aircraft. From a strictly AI perspective, however, this field was considered a developmental priority because of its strong link to the issue endowing intelligent devices with ability to move about in local environments autonomously. While vision is but one of the five human senses, the problems associated with its artificial manifestation are, essentially, the same for all the other senses (e.g., signal extraction, boundary recognition, identification, cognition). Therefore, research into the development of computer (or machine) vision has direct implications for the development of other machine senses.

Drawing from its origin in image processing, research into computer vision began in the 1960s with algorithms designed to locate the boundaries and establish the forms of objects in two-dimensional (2D) scenes, but quickly progressed to the inference of three-dimensional (3D) structure from clues provided by shading, texture, and focus. Rapid reconstruction of 3D scenes was aided greatly via the importation of techniques drawn from mathematical bundle adjustment theory (Triggs et al. 1999; Hartley and Zisserman 2004), photogrammetry (Wiora 2001; Jarve and Liba 2010), and graph cut optimization (Boykov et al. 2001; Boykov and Kolmogorov 2003). More recently, this domain has taken a substantial leap forward as a result of developments in image-based, deep-learning subsystems and their incorporation into computer/machine vision systems.

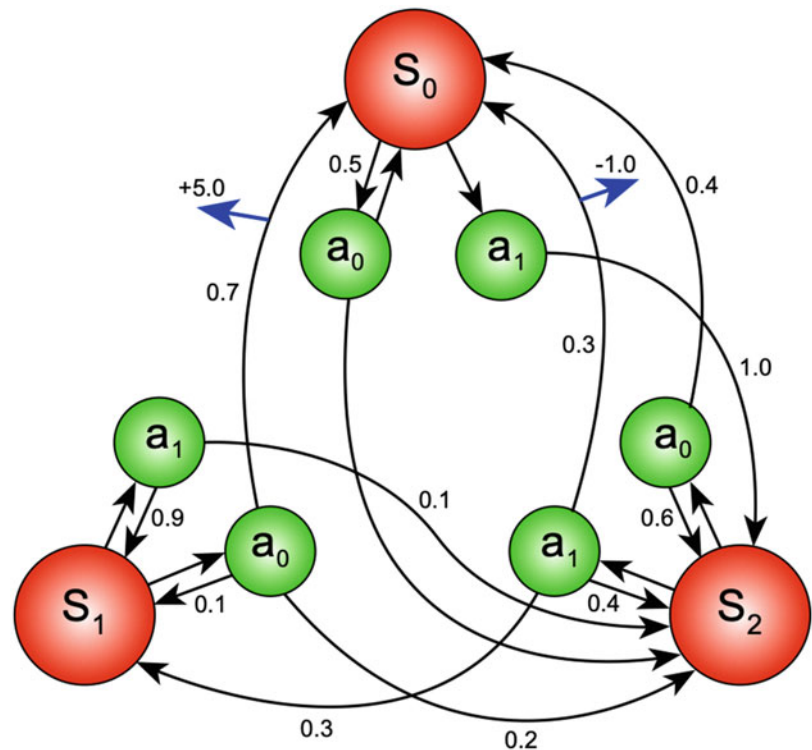
Contemporary, state-of-the-art, computer vision systems have found wide, and increasingly diverse, application in manufacturing system-control and quality-assurance, object identification, surveillance and law enforcement, autonomous vehicle/robot, and military applications. Indeed, these systems have reached the point where, in some respects, their performance is vastly superior to that of (even) expert capabilities (e.g., the sorting of object images into large numbers of fine-grained categories; see Culverhouse et al. 2013), whereas in others they still struggle with tasks humans find trivial (e.g., recognition of small of thin objects).

Planning Systems

Many human activities consist of sequences of actions that, ideally, are directed toward the attainment of a particular goal and, operationally, must be performed in a particular sequence. When the goal is specific (e.g., arrival at a particular destination from a given starting point), the domain over which the plan must be executed limited (e.g., must utilize local road systems or in urban areas, sidewalks, tram lines, or underground railways), and the optimality criterion known (e.g., shortest route, quickest travel time), planning can be a

Artificial Intelligence in the Earth Sciences, Fig. 3

A simple Markov decision process with three states (red), two actions (green), and two rewards (blue). Note the greater dependency of state S_2 on states S_0 and S_1 . (Redrawn from an original by Waldo Alvarez: see https://commons.wikimedia.org/wiki/File:Markov_Decision_Process.svg)



relatively simple exercise. However, when this problem is extended over multiple domains (e.g., multiple destinations, opening/closing times of various attractions along the way) whose characteristics might not be well established and whose states are subject to conditions that may change while the plan is being executed (e.g., traffic), determination of the optimal set and sequence of actions can be very complex. When the set of domains over which the plan must be executed is unknown, the challenge of determining optimal solutions unaided can be daunting. This problem category represents the activity domain of AI-empowered planning systems.

A number of generalized strategies have been developed to address planning problems which are fundamental to a (growing) range of different industries (e.g., logistics, workflow management, career progression, robot task management). For example, the range of possible actions and their consequences can be modeled by a network or as a surface and the optimal path found deterministically via constrained optimization (Bertsekas 2014; Verfaillie et al. 1996), Markovian decision processes (Burnetas and Katehakis 1997; Feinberg and Shwartz 2002, Fig. 3), or trial and error simulations (Kirkpatrick et al. 1983) with alternative routes found, by any of these approaches, as and when information pertinent to the plan's execution and derived from local or remote sensors changes. Alternatively route domains characterized by features that may not be fully observable can be planned using a partially observable Markovian decision process (POMDP) approach (Kaelbling et al. 1998). These procedures can be embedded in a number of different planning

scenarios (e.g., classical planning, probabilistic planning, conditional planning, conformant planning).

Robotics

Owing to the unique problems associated with the construction of machines that can move about in the environment and manipulate objects autonomously, robotics is the AI activity area that has seen the strongest mechanical engineering input. This activity area represents the intersection of electronic engineering, bioengineering, computer science, nanotechnology, and AI.

The first widely acknowledged autonomous robot was developed by George Devol, based on a patent application filed in 1954 (approved in 1961) for a single arm, stationary robot. This industrial robot, the *Unimate*, relied on a unique rotating drum memory module to perform a number of routine tasks. First employed to take die cast auto parts from a conveyor belt and weld them to auto bodies at General Motors' Inland Fisher Guide Assembly Plant in New Jersey, the original *Unimate*, along with the more flexible Programmable Universal Manipulation Arm (PUMA) system, could perform many other tasks.

Today, robots are so common in such a wide variety of manufacturing, construction, agricultural, law enforcement, consumer service, entertainment, emergency/rescue service, medical, and military industries, employ such a wide variety of technologies, and exhibit such an extreme range of designs; the idea of listing their commonalities might seem hopeless. However, all extant robots do share certain features with the

original *Unimate*: (1) mechanical construction, (2) electrical components to power the mechanical parts, and (3) computer programs to control the robot's movements and/or responses to its environment. Not all robots employ AI. Indeed, many simply perform sets of preprogrammed movements activated by remote human control. These devices are better thought of as automations rather than robots per se. However, AI-enabled robots are becoming more common with each passing year and are increasingly able to respond autonomously to changes in local conditions as those are transmitted to the system from various sensors located both on the robot's body and remotely via wireless communication links.

Fully autonomous, mobile, multipurpose, self-actuated robots remain a science-fiction fantasy at the moment. But a veritable international army of well-funded technologists are working diligently toward the ultimate goal of assembling such a device. Already there is growing concern among a variety of philosophical, civil rights, legal, regulatory, public health, law enforcement, and military bodies with regard to the opportunities and concerns the existence of this category of machines will embody. Such concerns are even beginning to be raised within the earth sciences (e.g., Wynsberghe and Donhauser 2018). These concerns are, to some extent, characteristic signs of contemporary interest in the entire subject of AI. Regardless, it is difficult to deny the symbolic role increasingly common, and increasingly autonomous, robots will have in shaping this public debate, irrespective of the many uses, services, capabilities, and economic savings they will provide.

Applications in the Earth Sciences

All forms of AI are currently being employed across the broad range of earth scientist research projects, and earth science researchers have an excellent historical record of taking advantage of the latest developments in this field. Among the first applications of AI technology in a geological context were the PROSPECTOR (Hart and Duda 1977; Hart et al. 1978) and muPETROL expert systems (Miller 1986, 1987, 1993). Both were Bayesian LISP-based systems programmed originally for IBM-PC platforms. PROSPECTOR was a mineral exploration system and was based on the analysis of geochemical data associated with three different types or mineral deposits. It remains available today as PROSPECTOR II (McCammon 1994). The muPETROL system focused on petroleum resource exploration and was based on Klemme's (1975, 1980, 1983) basin classification system with a set of consensus inference rules devised by experienced petroleum exploration geologists. Subsequent geological expert systems include DIPMETER ADVISOR (Carlborn 1982; Smith and Baker 1983), Fossil (Brough and Alexander 1986), EXPLORER (Mulvenna et al. 1991),

ARCHEO-NET (Fruitet et al. 1990), and Expert (Folorunso et al. 2012). Listings and descriptions of additional geological expert systems can be found in Aminzadeh and Simaan (1991) and Bremdal (1998). Expert systems are also available for a number of other earth science disciplines, including oceanography (e.g., Nada and Elawady 2017), atmospheric science (e.g., Peak and Tag 1989), geography (Fisher et al. 1988), and biology (Edwards and Cooley 1993 and references therein).

A number of NLP initiatives have been funded and/or are currently underway in the earth sciences. An initial, albeit limited, attempt to develop a system to automate the extraction of biodiversity data from taxonomic descriptions was developed by Curry and Connor (2007). This system, which included information on the stratigraphic ranges of fossil taxa, was based on the XML markup language which was used to annotate relevant terms, assignments, names, and other taxonomically relevant bits of information in taxonomic descriptions. Once annotated, these data could be migrated quickly, easily, and reliably into relational databases where appropriate data extractions, integrations, and analyses could take place. While the system proposed by Curry and Connor (2007) fell well short of a fully automated NPL system, it did demonstrate the importance of expert-level data annotation. Annotated data of precisely this sort today serve as training datasets for fully automated NPL systems, allowing them to learn how to identify different aspects of text passages to extract and tag for database migration. A more complete and up-to-date discussion of NPL techniques applied to the systematics and taxonomy literature has been provided by Thessen et al. (2012).

A sophisticated and successful fully NPL initiative of this type was the PaleoDeepDive project (Peters et al. 2014), which focused on the extraction of fossil occurrence data from the English, German, and Chinese language paleontological literature. This system employed NPL tools developed for other purposes (e.g., Tesseract and Cuneiform for text, Abbyy Fine Reader for tables, StanfordCoreNLP for linguistic context) to define factor graphs that were then used to aggregate variables (names, stratigraphic ranges, geographic occurrences) into meaningful combinations. The dataset used to train the PaleoDeepDive system was the Paleobiology Database (<http://paleobiodb.org>) which, at the time of training, consisted of over 300,000 taxonomic names and 1.2 million taxonomic occurrences, the majority of which had been extracted by hand from over 40,000 publications. This system was successful in extracting 59,996 taxonomic occurrences from a collection of literature sources that had not been entered into the Paleobiology Database previously.

In another, unrelated, initiative, ClearEarth (<https://github.com/ClearEarthProject>) is a US National Science Fund project established to bring computational linguistics and earth science researchers together to experiment with ways to

extract and integrate data from a variety of sources for the purpose of supporting research within and across the disciplinary fields of geology, ecology, and cryology. ClearEarth employs the ClearTK package of NPL routines, which is a well-established system for processing medical patient notes that is used on a routine basis by a large number of institutional health providers (Thessen et al. 2018). In essence, ClearEarth represents an attempt to port a mature and proven NPL technological infrastructure to aspects of earth science where, it is hoped, a similar penetration of institution-level integration and usage will result. Since 2016, ClearEarth has focused on the expert-informed markup of a well-annotated corpus of earth science text sources that will be used to train the ClearEarth system. To date, this project has released annotation guidelines for sea ice and earthquake events as well sponsoring a “hackathon” in 2017 at the University of Colorado (Boulder) where bio-ontologies were produced and extended.

Additional NPL earth science research and management areas that have undertaken NLP-based initiatives include oceanography (e.g., Manzella et al. 2017) and geography (e.g., Lampoltshammer and Heistracher 2012 and references therein).

Machine learning applications in the earth sciences are very common and have received a substantial boost over the past decade with the arrival of high-performance convolution neural network and deep-learning architectures. Examples will be reviewed in the companion to this article. For the sake of completeness though, a few references to reviews and initial trials of ML applications in various earth science fields are provided for the fields of stratigraphy (Zhou et al. 2019), sedimentology (Demyanov et al. 2019),

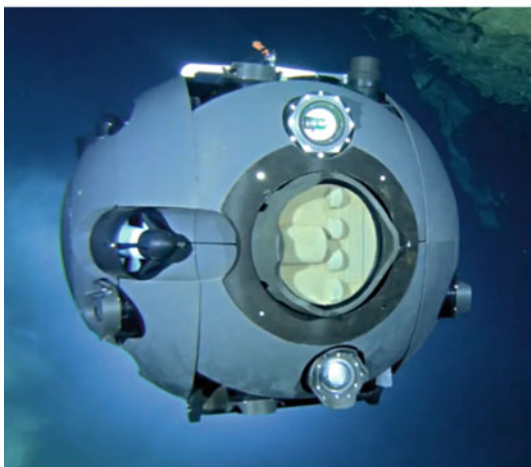
paleontology (Monson et al. 2018), mineralogy (Caté et al. 2017), petrology (Rubo et al. 2019), geophysics (Russell 2019), archaeology (Nash and Prewitt 2016; MacLeod 2018), geography (Kanevski et al. 2009), systematics and taxonomy (MacLeod 2007), biology (Tarca et al. 2007), ecology (Christin et al. 2019), oceanography (Dutkiewicz et al. 2015), glaciology (Zhang et al. 2019), atmospheric sciences (Ghada et al. 2019), and climate sciences (Cho 2018; Reichstein et al. 2019).

In the area of robotics, many of the most exciting AI developments have come from an earth science context. Two interesting recent deployments in this area are the EU-funded UNEXMIN project’s *UX-1* underwater mine exploration robot and the Woods Hole Oceanographic Institution’s (WHOI) *Nereid Under Ice* robot (Fig. 4).

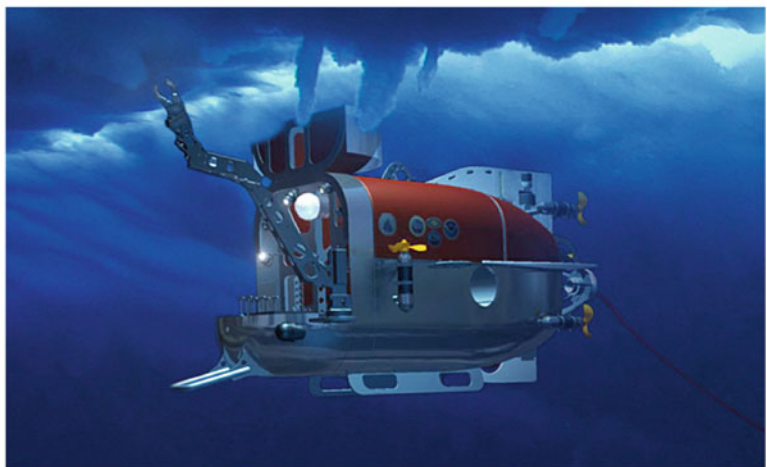
The UX1-Neo was developed by the European Union’s UNEXMin project primarily as an autonomous system for mapping and assessing Europe’s flooded mines into which it would be extraordinarily dangerous for humans to venture, much less work. Various *UX-1* systems currently exist each with a different set of sensors and analytic packages onboard. The system is designed to operate completely autonomously with only a guideline attached so the robot can be recovered if it becomes stuck or ceases to function properly. Aside from the technological challenge, creation of the UX1 series was justified as a means of determining whether it would be feasible and commercially profitable to reopen old, disused, flooded properties to extract rare-earth elements whose price has escalated dramatically in recent years. Regardless, such a system would be useful in many other earth science contexts.

In a similar vein, the WHOI *Nereid Under Ice* robot was recently sent to explore the Kolumbo volcano off Greece’s Santorini Island. Santorini was, of course, the site of one of

UX1-Neo



Nereid Under Ice



Artificial Intelligence in the Earth Sciences, Fig. 4 Portraits of the UNEXUP UX1-Neo robot with autonomous capabilities and the WHOI *Nereid* under the ice robot which was rendered fully autonomous via installation of an automation technology package developed by NASAs

PSTAR program. Autonomous robots such as these are pioneering AI technology developments that will have broad application across the earth and space sciences in the coming years. Image credits UNEXMIN and NOAA for the UX1-Neo and *Nereid* Under Ice robots respectively

the largest volcanic eruptions recorded in human history. For the 500 m Santorini dive, the *Nereid* was equipped with a new automation-technology package developed by US National Aeronautics and Space Administration's (NASA) Planetary Science and Technology from Analog Research (PSTAR) program. This NASA-PSTAR package included an AI-driven planning system that allowed the *Nereid* to plan its own dive as well as to determine which samples to take and how to take them. On the Santorini dive, the AI-enabled *Nereid* became the first robot to collect an underwater sample autonomously (see video clip at: <https://www.whoi.edu/press-room/news-release/whoi-underwater-robot-takes-first-known-automated-sample-from-ocean/>).

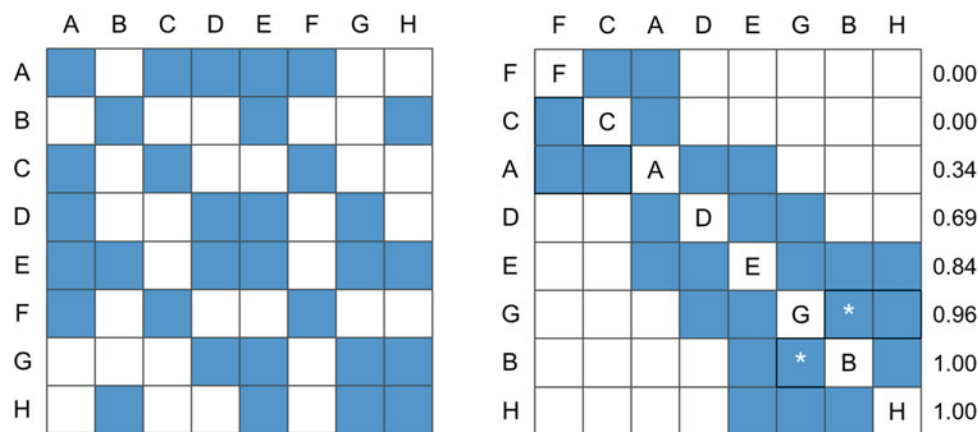
Both the UX1-Neo and NASA-PSTAR system-equipped *Nereid* represent a new breed of fully autonomous robots that will not only fill critical gaps in the capabilities of earth scientists now, but also provide test beds for the development of new and even more sophisticated autonomous robots in the future, for use both here on Earth and beyond.

While many believe the most efficient way to explore the geologies of distant planets is via a remote controlled robot (e.g., Spudis and Taylor 1992), and despite the fact that all current lunar and Martian rovers have been remotely controlled automations, the distances involved in the exploration of Mars and the moons of Jupiter and Saturn are such that a large degree of autonomous capability will need to be incorporated into the designs of future planet-exploration robots (Anonymous 2018). At present the ExoMars Phase 2 Mission (due to be launched in 2022) will include a substantially automated robotic system designed to complement the ExoMars Phase 1 Mission (launched in 2016). The difference between the levels of autonomous capability incorporated into these two rovers reflects developments – and confidence – in the field of robot autonomous systems (RAS) over the past

5 years. Similar levels of autonomous robotic capabilities are expected to be part of the Mars Sample Return (MSR) Mission (expected launch 2026), Martian Moon (Phobos) Sample Return Mission (expected launch 2024), and various orbital missions currently in planning stages.

Planning Systems

While AI planning systems might appear to have more to do with the logistics of earth science research than the science itself, it is possible to take a broader view of planning which is, in essence, an attempt to order a set of data into a sequence according to some error-minimization criterion. In this context, the routine – though complex – task of inferring the correct sequence of stratigraphic datums from samples collected from multiple stratigraphic sections or cores is not dissimilar to a planning problem. Here, the geologist/stratigrapher must infer the correct order a set of stratigraphic datums whose true order is obscured by a variety of distributional and preservation factors (Fig. 5). This problem can be approached in many ways (e.g., Shaw 1964; Gordon and Reymont 1979; Edwards 1989; MacLeod and Keller 1991). However, several recent and quite sophisticated approaches have treated it as a quasi-planning problem and developed either probabilistic (Agterberg et al. 2013), rule-based (Guex and Davaud 1984; Guex 2011), or constrained optimization (Kemple et al. 1995; Sadler and Cooper 2003; Sadler 2004; Fan et al. 2020) approaches to its solution; approaches that, in the case of the latter, strongly mimic that of an AI-based planning system approach.



Artificial Intelligence in the Earth Sciences, Fig. 5 Stratigraphic coexistence matrix before (left) and after (right) permutation to infer the correct relative order, or slotting, of the datums. In a spatial sense, this problem is similar to a minimum trip distance problem in which the distances between locations are known, but the most efficient sequence

of destinations is not. Such problems can be solved using an AI-based planning system approach. For these data, the coexistence of B and G is implied but not observed. The column of numbers to the right represent reciprocal averaging scores. (Redrawn from Sadler 2004)

Conclusion and Prospectus

Artificial intelligence-based applications have been a part of earth science research and industrial applications for over 40 years, virtually as long as such applications have been available for use outside the fields of mathematics and information technology. Levels of interest and activity in AI by the earth science community have waxed and waned over these decades in much the same way as the fortunes of AI generally. While it is true most of the activity in this community has come from technology and data-analysis enthusiasts, AI is undergoing a renaissance of interest across the earth science disciplines currently, which mirrors renewed government and commercial interest in AI-based solutions to a wide range of technical, technological, and social problems generally. In large part, this renewed interest has been driven by advances in the machine learning AI activity domain. But these advances have had the effect of stimulating research, development, and interest in all forms of AI.

Many seem to be concerned with the effect AI will have on the direction of the fields to which it is applied and especially on the employment prospects for young people. Such concerns are, in my view, misplaced. Earth sciences as a whole have become increasingly quantitative over the past 70 years, and there is little evidence to support a conclusion that this has led to unethical practices, or reduced employment prospects, in the sector. Routine identification, sampling, and data-collection tasks are, increasingly, able to be performed by quasi-autonomous AI-enabled machines. This trend is very likely to continue. But in many ways it is these tasks that often take us away from the more creative and integrative aspects of our science; two areas of performance AI approaches struggle to deliver. In addition, it is simply beyond question that human experts suffer from a wide variety of performance limitations that the use of AI-based technologies can assist in addressing (e.g., MacLeod et al. 2010; Culverhouse et al. 2013). Instead of viewing the rise of AI as a threat to the development of the earth sciences, the community should, as it has with other technological advances in the past, welcome such developments as much-needed and much-valued facilitators/partners in the task of exploring and understanding our planet and its inhabitants: past, present, and future.

As with any genuine partnership though, both sides need to work to make the most of the collaboration. Devices that employ aspects of AI can be massive boon to earth science research, but AI cannot deliver on this promise alone. In order to be effective, AI systems need data and guidance. Earth scientists have data, but we have not been diligent in aggregating, cleaning, updating, and making our data available. Various earth science big data initiatives are currently

underway (e.g., Spina 2018; Normile 2019; Stephenson et al. 2019). These need not only to be supported and contributed to by earth scientists, but the guidance AI applications need must be delivered via engagement with, and active participation in, AI-related initiatives by expert earth scientists of all types. In addition, the training of students interested in the earth sciences at all levels needs to be reorganized and expanded to equip them with the skills they will need to take full advantage of the coming AI revolution. Finally, the earth science community needs to establish AI as a priority – perhaps a top priority – for ensuring the future success of our science through the diverse, ongoing professional education, outreach, and influencing activities of our professional societies.

Bibliography

- Agterberg FP, Gradstein FM, Cheng Q, Liu G (2013) The RASC and CASC programs for ranking, scaling and correlation of biostratigraphic events. *Comput Geosci* 54:279–292
- Aminzadeh F, Simaan M (1991) Expert systems in exploration. Society of Exploration Geophysicists, Tulsa, Oklahoma
- Anonymous (2018) Space robotics & autonomous systems: widening the horizon of space exploration. Available via UK-RAS Network. <https://www.ukras.org/publications/white-papers/>. Accessed 4 Mar 2020
- Anonymous (2019) Artificial intelligence alone won't solve the complexity of earth sciences. *Nature* 566:153
- Bertsekas DP (2014) Constrained optimization and Lagrange multiplier methods. Academic, New York
- Boykov Y, Kolmogorov V (2003) Computing geodesics and minimal surfaces via graph cuts. In: *Proceedings Ninth IEEE International Conference on Computer Vision 2003*, pp 26–33
- Boykov Y, Veksler O, Zabih R (1998) Markov random fields with efficient approximations: proceedings. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition 1998*, pp 648–655
- Boykov Y, Veksler O, Zabih R (2001) Fast approximate energy minimization via graph cuts. *IEEE Trans Pattern Anal Mach Intell* 23: 1222–1239
- Bremdal BA (1998) Expert systems for management of natural resources. In: Liebowitz J (ed) *The Handbook of Applied Expert Systems*. CRC Press, Boca Ration, Louisiana, pp 30-1–30-44
- Brough DR, Alexander IF (1986) The Fossil expert system. *Expert Syst* 3:76–83
- Burnetas AN, Katehakis MN (1997) Optimal adaptive policies for Markov decision processes. *Math Oper Res* 22:222–255
- Carlborn I (1982) Dipmeter advisor expert system. *Am Assoc Pet Geol Bull* 66:1703–1704
- Caté A, Perozzi L, Gloaguen E, Blouin M (2017) Machine learning as a tool for geologists. *Leading Edge*, pp 64–68
- Cho R (2018) Artificial intelligence—a game changer for climate change and the environment. *State of the Planet*, pp 1–13
- Church A (1936) An unsolvable problem of elementary number theory. *Am J Math* 58:345–363
- Christin S, Hervet É, Lecomte N (2019) Applications for deep learning in ecology. *Methods Ecol Evol* 10:1632–1644

- Culverhouse PF, MacLeod N, Williams R, Benfield MC, Lopes RM, Picheral M (2013) An empirical assessment of the consistency of taxonomic identifications. *Mar Biol Res* 10:73–84
- Curry GB, Connor RJ (2007) Automated extraction of biodiversity data from taxonomic descriptions. In: Curry GB, Humphries CJ (eds) *Biodiversity databases: techniques, politics and applications*. The Systematics Association and CRC Press, Boca Raton
- Demyanov V, Reesink AJH, Arnold DP (2019) Can machine learning reveal sedimentological patterns in river deposits? In: Corbett PWM, Owen A, Hartley AJ, Pla-Pueyo S, Barreto D, Hackney C, Kape SJ (eds) *River to reservoir: geoscience to engineering*, Geological Society of London, Special Publication No. 488, London
- Dutkiewicz A, Müller RD, O'Callaghan S, Jónasson H (2015) Census of seafloor sediments in the world's ocean. *Geology* 43:795–798
- Ebert-Uphoff I, Samarasinghe S, Barnes E (2019) Thoughtfully using artificial intelligence in earth science. *Eos* 100:1–5
- Edwards LE (1989) Supplemented graphic correlation: a powerful tool for paleontologists and nonpaleontologists. *Palaios* 4:127–143
- Edwards M, Cooley RE (1993) Expertise in expert systems: knowledge acquisition for biological expert systems. *Comput Appl Biosci* 9: 657–665
- Fan J et al. (2020) A high-resolution summary of Cambrian to Early Triassic marine invertebrate biodiversity. *Science* 367:272–277
- Feinberg EA, Shwartz A (2002) *Handbook of Markov decision processes*. Kluwer, Boston
- Fisher PF, Mackaness WA, Peacegood G, Wilkinson GG (1988) Artificial intelligence and expert systems in geodata processing. *Progr Phys Geogr Earth Environ* 12:371–388
- Folorunso IO, Abikoye OC, Jimoh RG, Raji KS (2012) A rule-based expert system for mineral identification. *J Emerg Trends Comput Inform Sci* 3:205–210
- Fruitet J, Kalloufi L, Laurent D, Boudad L, de Lumley H (1990) “ARCHEO-NET” a prehistoric and paleontological material data base for research and scientific animation. In: Tjoa AM, Wagner R (eds) *Database and expert systems applications*. Springer, Wien/New York
- Ghada W, Estrella N, Menzel A (2019) Machine learning approach to classify rain type based on Thies disdrometers and cloud observations. *Atmos* 10:1–18
- Gordon AD, Reymont RA (1979) Slotting of borehole sequences. *Math Geol* 11:309–327
- Guex J (2011) Some recent ‘refinements’ of the unitary association method: a short discussion. *Lethaia* 44:247–249
- Guex J, Davaud E (1984) Unitary associations method: use of graph theory and computer algorithm. *Comput Geosci* 10:69–96
- Hart PE, Duda RO (1977) PROSPECTOR – a computer-based consultation system for mineral exploration. Technical Note 155, SRI International, Menlo Park
- Hart PE, Duda RO, Einaudi MT (1978) PROSPECTOR – a computer-based consultation system for mineral exploration. *J Int Assoc Math Geol* 10:589–610
- Hartley RI, Zisserman A (2004) *Multiple view geometry in computer vision*. Cambridge University Press, Cambridge, UK
- Hayes-Roth F, Waterman DA, Lenat DB (1983) *Building expert systems*. Addison-Wesley, Reading, Massachusetts
- Hutchins WJ (2004) The Georgetown-IBM experiment demonstrated in January 1954. In: Frederking RE, Taylor KB (eds) *Machine translation: from real users to research*. Lecture notes in computer science 326. Springer, Berlin
- Jarve I, Liba N (2010) The effect of various principles of external orientation on the overall triangulation accuracy. *Technol Mokslai* 86:59–64
- Kaelbling LP, Littman ML, Cassandra AR (1998) Planning and acting in partially observable stochastic domains. *Artif Intell* 101:99–134
- Kanevski M, Foresti L, Kaiser C, Pozdnoukhov A, Timonin V, Tuia D (2009) Machine learning models for geospatial data. In: Bavaud F, Christophe M (eds) *Handbook of theoretical and quantitative geography*. University of Lausanne, Lausanne
- Karp RM (1972) Reducibility among combinatorial problems. In: Miller RE, Thatcher JW (eds). *Complexity of Computer Computations*. Plenum Press, New York, pp 85–103
- Kempell WG, Sadler PM, Strauss DJ (1995) Extending graphic correlation to many dimensions: stratigraphic correlation as constrained optimization. In: Mann KO, Lane HR (eds). *Graphic Correlation*. SEPM Society for Sedimentary Geology, Special Publication 53, Tulsa, Oklahoma, pp 65–82
- Kirkpatrick S, Gelatt CD Jr, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220:671–680
- Klemme HD (1975) Giant oil fields related to their geologic setting—a possible guide to exploration. *Bull Can Petrol Geol* 23:30–36
- Klemme HD (1980) Petroleum basins—classifications and characteristics. *J Pet Geol* 3:187–207
- Klemme HD (1983) Field size distribution related to basin characteristics. *Oil Gas J* 81:187–207
- Lampoltshammer TJ, Heistracher T (2012) Natural language processing in geographic information systems – some trends and open issues. *Int J Comput Sci Emerg Tech* 3:81–88
- LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proc IEEE* 86:2278–2323
- Liebowitz J (1997) *The handbook of applied expert systems*. CRC Press, Boca Raton
- Lighthill J (1973) *Artificial intelligence: a paper symposium*. UK Science Research Council, London
- MacLeod N (2007) Automated taxon identification in systematics: theory, approaches, and applications. CRC Press/Taylor & Francis Group, London
- MacLeod N (2018) The quantitative assessment of archaeological artifact groups: beyond geometric morphometrics. *Quat Sci Rev* 201: 319–348
- MacLeod N, Benfield M, Culverhouse PF (2010) Time to automate identification. *Nature* 467:154–155
- MacLeod N, Keller G (1991) How complete are Cretaceous/Tertiary boundary sections? A chronostratigraphic estimate based on graphic correlation? *Geol Soc Am Bull* 103:1439–1457
- Manzella G et al (2017) Semantic search engine for data management and sustainable development: marine planning service platform. In: Diviacco P, Ledbetter A, Graves H (eds) *Oceanographic and marine cross-domain data management for sustainable development*. IGI Global, Hershey
- McCammon RB (1994) Prospector II: towards a knowledge base for mineral deposits. *Math Geol* 26:917–936
- McCorduck P (1979) *Machines who think*. W. H. Freeman, San Francisco
- McCulloch W, Pitts W (1943) A logical calculus of ideas immanent in nervous activity. *Bull Math Biophys* 5:115–133
- Miller BM (1986) Building an expert system helps classify sedimentary basins and assess petroleum resources. *Geobyte* 1(44–50):83–84
- Miller BM (1987) The MuPETROL expert system for classifying world sedimentary basins. *US Geol Surv Bull* 1810:1–87
- Miller BM (1993) Object-oriented expert systems and their applications to sedimentary basin analysis. *US Geol Surv Bull* 2048:1–31
- Monson TA, Armitage DW, Hlusko LJ (2018) Using machine learning to classify extant apes and interpret the dental morphology of the chimpanzee-human last common ancestor. *PaleoBios* 35:1–20
- Mulvenna MD, Woodham C, Gregg JB (1991) Artificial intelligence applications in geology: a case study with EXPLORER. In: McTear MF, Creaney N (eds) *AI and Cognitive Science '90. Workshops in Computing 1991*, pp 109–119
- Nada YA, Elawady YH (2017) Analysis, design, and implementation of intelligent fuzzy expert system for marine wealth preservation. *International Journal of Computer Applications* 161:15–20

- Normile D (2019) Earth scientists plan to meld massive databases into a 'geological Google'. *Science* 363:917
- Peak JE, Tag PM (1989) An expert system approach for prediction of maritime visibility obscuration. *Mon Weather Rev* 117:2641–2653
- Peters SE, Zhang C, Livny M, Ré C (2014) A machine reading system for assembling synthetic paleontological databases. *PLoS One* 9: 1–22
- Pierce JR, Carroll JB, Hamp EP, Hays DG, Hockett CF, Oettinger AG, Perlis A (1966) Language and machines: computers in translation and linguistics. National Academy of Sciences and National Research Council, Washington, DC
- Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE* 77:257–286
- Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, Prabhat (2019) Deep learning and process understanding for data-driven Earth system science. *Nature* 566:195–204
- Rubo RA, Carneiro CC, Michelon MF, dos Santos GR (2019) Digital petrography: mineralogy and porosity identification using machine learning algorithms in petrographic thin section images. *J Pet Sci Eng* 183:106382
- Russell B (2019) Machine learning and geophysical inversion – a numerical study. *Lead Edge* 38:512–519
- Sadler PM (2004) Quantitative biostratigraphy – achieving finer resolution in global correlation. *Annual Review of Earth and Planetary Science* 32:187–213
- Sadler PM, Cooper RA (2003) Best-fit intervals and consensus sequences: comparison of the resolving power of traditional biostratigraphy and computer-assisted correlation. In: Harries PJ (ed) *High-Resolution Stratigraphic Correlation*. Kluwer, Amsterdam, The Netherlands, pp 49–94
- Samuel AL (1959) Some studies in machine learning using the game of checkers. *IBM J Res Dev* 3:210–229
- Shaw A (1964) *Time in stratigraphy*. McGraw-Hill, New York
- Smith RG, Baker JD (1983) The DIPMETER ADVISOR system. A case study in commercial expert system development. In: van den Herik J, Filipe J (eds) *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, Rome, pp 122–129
- Spina R (2018) Big data and artificial intelligence analytics in geosciences: promises and potential. *GSA Today* 29:42–43
- Spudis PD, Taylor GJ (1992) The roles of humans and robots as field geologists on the moon. In: Mendell W.W. (ed) *2nd Conference on Lunar Bases and Space Activities*. NASA Lyndon B. Johnson Space Center, Houston, Texas, pp 307–313
- Stephenson M, Cheng Q, Wang C, Fan J, Oberhänsli R (2019) On the cusp of a revolution. *Geoscientist* 2019:16–19
- Stokes JM et al (2020) A deep learning approach to antibiotic discovery. *Cell* 180:688–702.e13
- Tarca AL, Carey VJ, Chen X, Romero R, Drăghici S (2007) Machine learning and its applications to biology. *PLoS Comput Biol* 3: 953–963
- Thessen AE, Cui H, Mozzherin D (2012) Applications of natural language processing in biodiversity science. *Adv Bioinforma* 2012: 1–17
- Thessen A, Preciado J, Jenkins C (2018) Collaboration between the natural sciences and computational linguistics: a discussion of issues. *INSTAAR Univ Colorado Occas Rep* 28:1–24
- Triggs B, McLauchlan P, Hartley R, Fitzgibbon A (1999) Bundle adjustment – a modern synthesis. In: Triggs W, Zisserman A, Szeliski R (eds) *Vision algorithms: theory and practice (CCV'99: proceedings of the international workshop on vision algorithms)*. Springer, Berlin/Heidelberg/New York, pp 298–372
- Turing AM (1936, Published in 1937) On computable numbers, with an application to the Entscheidungsproblem. *Proc Lond Math Soc* 42: 230–265
- Turing AM (1950) Computing machinery and intelligence. *Mind* 49: 433–460
- van Wynsberghe A, Donhauser J (2018) The dawning of the ethics of environmental robots. *Sci Eng Ethics* 24:1777–1800
- Verfaillie G, Lemaitre M, Schiex T (1996) Russian doll search for solving constraint optimization problems. *Proc Natl Conf Artif Intell* 1:181–187
- Wexelblat RL (1981) *History of programming languages*. Academic, New York
- Wiora G (2001) *Optische 3D-Messtechnik: präzise gestaltvermessung mit einem erweiterten streifenprojektionsverfahren*. Ruprechts-Karls Universität, 36p
- Zhang J, Jia L, Menenti M, Hu G (2019) Glacier facies mapping using a machine-learning algorithm The Parlung Zangbo Basin case study. *Remote Sens* 11:1–38
- Zhou C, Ouyang J, Ming W, Zhang G, Du Z, Liu Z (2019) A stratigraphic prediction method based on machine learning. *Applied Sciences* 9:3553

Artificial Neural Network

Yuanyuan Tian¹, Mi Shu² and Qingren Jia³

¹School of Geographical Science and Urban Planning, Arizona State University, Tempe, AZ, USA

²Tencent Technology (Beijing) Co., Ltd, Beijing, China

³College of Electronic Science and Technology, National University of Defense Technology, Changsha, China

Definition

An artificial neural network (ANN), also known as neural network, is a computational model capable of processing information to tackle tasks such as classification and regression. It is the component of artificial intelligence inspired by the human brain and nervous system.

Introduction

Artificial neuron and the coding signals in an ANN are used to simulate the electrical activities of how the nervous system communicates. Each artificial neural neuron receives input signals, processes these signals, and generates an output signal. A general practice is to sum the weighted inputs, and then feed the sum value into a transfer function to judge if the value is qualified to be transferred to another neuron as an input. The simplest transfer function is to compare the sum value with the set threshold, while nonlinear functions allow more flexibility.

Starting from the 1940s, ANNs have evolved from perceptron, an algorithm for supervised learning of binary classifiers, through backpropagation, a way to train a multi-layer network, to deep learning. One of the reasons why ANNs have become so popular is that the network's self-learning capability of ANNs enables them to solve complex

problems, which are difficult or unfulfillable for humans, and even produce better results than statistical methods. Also, they can deal with incomplete, noisy, and ambiguous data.

Nowadays, ANNs are used in a variety of research and industrial areas. The applications of ANNs are wide, encompassing engineering, finance, communication, medical, and so on. The use of ANNs in geography stems from the need to handle big spatial data. With the development of geo-information technologies such as Remote Sensing (RS), Geographic Information System (GIS), and Global Navigation Satellite System (GNSS), spatial data processing and analysis begin to ask for artificial intelligence to handle complex geographical problems.

The purpose of this encyclopedia entry is to introduce the basic concept of ANNs, review major applications in geoinformatics, and discuss future trends.

ANN Models

Components of ANNs

ANN refers to an artificial neuron system built by imitating human neurons for information processing. It is a computing model, which is composed of a large number of interconnected **neurons**. The neuron can be regarded as a calculation and storage unit. A single neuron generally has multiple inputs. After some calculations (i.e., weighted addition), the output will be used as the input of subsequent neurons. In a neural network, every **connection** between two neurons represents a **weight** to reflect the strength of the connection. Finally, neurons are combined layer by layer to form a neural network (Fig. 1).

In more detail, let a_i to represent the inputs of a neuron, and w_i to represent the weights on the connections. Then, the output of this neuron would be

$$z = f(\text{sum}(w_i w a_i)),$$

where f represents an activation function. With the help of the activation function, what a neuron can do is more than just weighting and summing the received information and outputting it to the next neuron. The activation function simulates such a process. In reality, neurons in the organism are not

necessarily activated after receiving information, but determine whether this information has reached a threshold. If the value of the information does not reach this threshold, the neuron is in a state of inhibition, otherwise, the neuron is activated and transmits the information to the next neuron. Similarly, a neuron in an ANN would decide whether and how to pass the information. In order to improve the utilization of neurons, nonlinear functions are usually used instead of thresholds as the activation functions of neurons. Besides, by introducing nonlinear activation functions, a neural network can theoretically fit any objective function, making it powerful.

Organization

Structurally, a neural network can be divided into input layer, output layer, and hidden layer. The input layer accepts signals and data from the outside world, and the output layer realizes the output of the system processing results. The hidden layer is between the input layer and the output layer and cannot be observed from outside the system.

When designing a neural network, the number of neurons in the input layer and output layer is often fixed, and the hidden layer can be freely specified. The number of hidden layers and the number of neurons in each layer determine the complexity of the neural network. In the 1950s, ANN presented itself in the form of a perceptron, without a hidden layer. Later in the 1980s, ANN was developed into a multi-layer perceptron, with only one hidden layer. From 2012, ANN tended to have more and more hidden layers (deep learning). The reason for the development of neural network structure is that as the number of layers increases, the parameters of the entire network increase. More parameters mean that the simulated function by ANN can be more complicated to solve harder problems (Fig. 2).

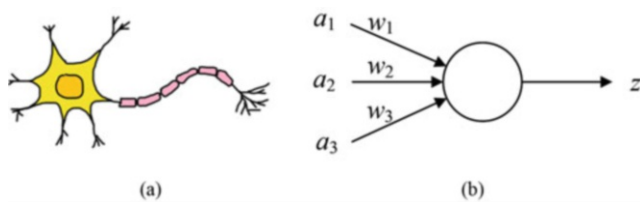
Learning

A neural network training algorithm is to adjust the value of all the weights to the best, so that the prediction performance of the entire network could be the best.

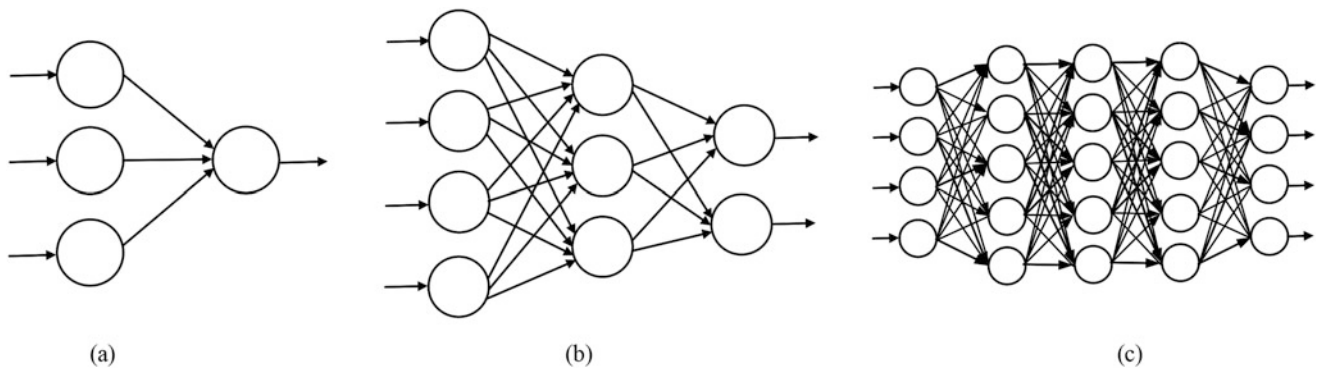
In order to fit the target better, the neural network needs to know how far the prediction is from predicted target. Let the sample be y_p and the true target be y . Then, loss function needs to be defined to measure the distance between y_p and y , so as to define what weight parameters are appropriate. There are some different loss functions in practice. For example, the calculation formula could be as follows.

$$\text{loss} = (y_n - y)^2$$

Then, the training goal is to minimize the loss function and make the prediction result of the neural network for the input sample closer and closer to the true value by adjusting each



Artificial Neural Network, Fig. 1 (a) a natural neuron vs. (b) a neuron in an ANN



Artificial Neural Network, Fig. 2 The development of neural network structure. (a) perception; (b) multilayer perception, MLP; (c) deep learning

weight parameter. In practice, the optimization algorithm represented by the gradient descent method can be used to solve the above minimization problem. The back-propagation method is the specific implementation method of the gradient descent method on the neural network. Its main purpose is to pass the output error back and distribute the error to all the neurons of each layer, so as to obtain the error of each neuron, and then modify each weight accordingly.

In general, in ANN Models, a collection of artificial neurons are created and connected together, allowing them to send messages to each other. The network is asked to solve a problem, so it attempts to make it by iteration, each time strengthening the connections that lead to success and diminishing those that lead to failure.

Applications

Tasks suited for ANNs are classification, regression and clustering, etc. ANNs have been applied in many disciplines, especially in earth science.

ANNs have a long history of use in the geosciences (Van der Baan and Jutten 2000), such as hydrology, ocean modeling and coastal engineering, and geomorphology, and they remain popular for modeling both regression and/or classification (supervised or unsupervised) of nonlinear systems. As the early landmarks, the revival of neural networks applied on high-resolution satellite data were applied to classify land cover and clouds (Benediktsson et al. 1990; Lee et al. 1990) almost 30 years ago. In addition to spatial prediction, ANNs have also been used to study dynamics in the earth systems by mapping time-varying features to temporally varying target variables in land, ocean, and atmosphere domains. For instance, by using (multitemporal) remote sensing data, ANN can be used to detect forest changes caused by a prolonged drought in the Lake Tahoe Basin in California (Gopal and Woodcock 1996). Also, an artificial neural network with one hidden layer was able to filter out noise, predict the

diurnal and seasonal variation of carbon dioxide (CO₂) fluxes, and extract patterns such as increased respiration in spring during root growth (Papale and Valentini 2003). In these early works, ANNs were proved to provide better classification or regression performance compared to traditional methods in given scenarios. However, compared to statistical-based machine learning (ML) methods, ANNs do not have many overwhelming advantages.

Deep neural networks (DNNs), or deep learning (DL), extend the classical ANN by incorporating multiple hidden layers and enable good model performance without requiring well-chosen features as inputs. DNNs have attracted wide attention in the new millennium, mainly by outperforming alternative machine-learning algorithms in numerous application areas. In 2012, AlexNet, a convolutional neural network (CNN) model won Imagenet's image classification contest with an accuracy of 84% jump over 75% accuracy that earlier ML models had achieved based on its efficient use of graphic processing units (GPU), rectified linear units (ReLUs), and large training data volume. This win triggers a new DNN boom globally.

In the geoscience domain, researchers observe, collect, and process geological data and discover interested knowledge of the geoscientific phenomenon from the data. While the earth system is a complex system, building a quantitative understanding of how the Earth works and evolves by using traditional tools from geology, chronology, physics, chemistry, geography, biology, and mathematics becomes difficult. However, as the volume of data accumulated through long-term tracking and recording, the scientific research paradigm shifts from knowledge-driven to data-driven (Kitchin 2014). This shift provides opportunities for ANN-based machine learning methods in geoscience fields, as they can capture nonlinear relationships between the inputs as well as reduce the dimensions of the model and help convert the input into a more useful output.

Specifically, there are mainly three popular research fields in geomatics:

- Although applications to problems in geosciences are still in their infancy (Reichstein et al. 2019), tasks using remote-sensing images benefit a lot by introducing deep learning methods that are rapidly developing in computer science. Since 2014, the remote sensing community has shifted its attention to DL, and DL algorithms have achieved significant success at many image analysis tasks including land use and land cover classification, scene classification, and object detection. CNN is more popular than other DL models, and spectral-spatial features of images are utilized for analysis. The main focus of the studies is on classification tasks, but several other applications are worth mentioning, including fusion, segmentation, change detection, and registration. Therefore, it appeared that DL could be applied to almost every step of the remote sensing image processing.
- In solid Earth geoscience, DNNs trained on simulation-generated data can learn a model that approximates the output of physical simulations (Bergen et al. 2019). It can be used to accurately reproduce the time-dependent deformation of the Earth (DeVries et al. 2017), or perform fast full wavefield simulations by training synthetic data from a finite difference model using CNN model (Moseley et al. 2018). In another example, Shoji et al. (2018) use the class probabilities returned by CNN to identify the mixing ratio for ash particles with complex shapes. Several recent studies have applied DNNs with various architectures for automatic earthquake and seismic event detection, phase-picking, and classification of volcano-seismic events. All these tasks are difficult for expert analysts.
- In dynamic earth system researches, such as climatology and hydrology, ANNs, especially DNNs, also play a vital role. For instance, a neural network was used to predict the evolution of the El Niño Southern Oscillation (ENSO) and it shows that ENSO precursors exist within the South Pacific and Indian Oceans (Ham et al. 2019). Recent studies demonstrate the application of deep learning to the problem of extreme weather, for instance, hurricane detection (Racah et al. 2017). Also, Gagne II et al. (2019) employed neural networks to predict the likelihood that a convective storm would produce hail, showing that the neural networks made accurate predictions by identifying known types of storm structures. The studies report success in applying deep learning architectures to objectively extract spatial features to define and classify extreme situations (for example, storms, hails, atmospheric rivers) in numerical weather prediction model output. Such an approach enables rapid detection of such events and

forecast simulations without using either subjective human annotation or methods that rely on predefined arbitrary thresholds for wind speed or other variables.

Conclusions

Based on many existing works in recent years, the adoption of ANNs in geoinformatics are quite successful. Although ANNs have been widely used in geoscience applications, geoscientists sometimes hesitate to use them due to the lack of interpretability. The ANN is criticized as being oversold as a “black-box” with output oriented regardless of understanding the network structure.

Recently, there have been efforts within the geoscience community to compile methods for improving machine learning model interpretability (Gagne II et al. 2019; McGovern et al. 2019). These interpretation methods are applied to trace the decision of a neural network back onto the original dimensions of the inputs and thereby permit the understanding of which input variables are most important for the neural network’s decisions. Backward optimization and layerwise relevance propagation (LRP) reveal that the neural network mainly focuses its attention on the tropical Pacific, which is consistent with expert perception (Toms et al. 2020). This can help uncover the “black-box” construction of ANNs. Also, expert knowledge can be introduced to educate the network structure formation and optimization. Empirical, inner model comparisons and assessments are needed to facilitate understanding of the physical mechanism of the spatial processes. Therefore, for geoscientists who embrace neural network methods, interpretable ANNs will be one of the promising directions where breakthroughs may be made.

Cross-References

- [Deep Learning in Geoscience](#)
- [Geographical Information Science](#)
- [Machine Learning](#)

Bibliography

- Benediktsson JA, Swain PH, Ersoy OK (1990) Neural network approaches versus statistical methods in classification of multisource remote sensing data
- Bergen KJ, Johnson PA, Maarten V, Beroza GC (2019) Machine learning for data-driven discovery in solid Earth geoscience. *Science* 363
- DeVries PMR, Ben TT, Meade BJ (2017) Enabling large-scale viscoelastic calculations via neural network acceleration. *Geophys Res Lett* 44:2662–2669

- Gagne DJ II, Haupt SE, Nychka DW, Thompson G (2019) Interpretable deep learning for spatial analysis of severe hailstorms. *Mon Weather Rev* 147:2827–2845
- Gopal S, Woodcock C (1996) Remote sensing of forest change using artificial neural networks. *IEEE Trans Geosci Remote Sens* 34: 398–404
- Ham Y-G, Kim J-H, Luo J-J (2019) Deep learning for multi-year ENSO forecasts. *Nature* 573:568–572
- Kitchin R (2014) Big data, new epistemologies and paradigm shifts. *Big Data Soc* 1:2053951714528481
- Lee J, Weger RC, Sengupta SK, Welch RM (1990) A neural network approach to cloud classification. *IEEE Trans Geosci Remote Sens* 28: 846–855
- McGovern A, Lagerquist R, John Gagne D et al (2019) Making the black box more transparent: understanding the physical implications of machine learning. *Bull Am Meteorol Soc* 100: 2175–2199
- Moseley B, Markham A, Nissen-Meyer T (2018) Fast approximate simulation of seismic waves with deep learning. *arXiv preprint arXiv:180706873*
- Papale D, Valentini R (2003) A new assessment of European forests carbon exchanges by eddy fluxes and artificial neural network spatialization. *Glob Chang Biol* 9:525–535
- Racah E, Beckham C, Maharaj T et al (2017) ExtremeWeather: a large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. *Adv Neural Inf Proces Syst* 30:3402–3413
- Reichstein M, Camps-Valls G, Stevens B et al (2019) Deep learning and process understanding for data-driven Earth system science. *Nature* 566:195–204
- Shoji D, Noguchi R, Otsuki S, Hino H (2018) Classification of volcanic ash particles using a convolutional neural network and probability. *Sci Rep* 8:8111
- Toms BA, Barnes EA, Ebert-Uphoff I (2020) Physically interpretable neural networks for the geosciences: applications to earth system variability. *J Adv Model Earth Syst* 12:e2019MS002002
- Van der Baan M, Jutten C (2000) Neural networks in geophysical applications. *Geophysics* 65:1032–1047

Autocorrelation

Alejandro C. Frery

School of Mathematics and Statistics, Victoria University of Wellington, Wellington, New Zealand

Synonyms

Second moment; Serial correlation (when referring to time series)

Definition

The autocorrelation is a measure of the association between two random variables of the same process. In geosciences; it is often called autocovariance.

There are many ways of measuring the association of two random variables. Among them, the Pearson coefficient of correlation between two random variables $X, Y: \Omega \rightarrow \mathbb{R}$ is defined as

$$\rho(X, Y) = \frac{E(XY) - E(X)E(Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}, \quad (1)$$

provided this quantity is well defined. The numerator in Eq. (1) is called “covariance.”

The main properties of Eq. (1) are:

Symmetry: $\rho(X, Y) = \rho(Y, X)$.

Boundedness: $-1 \leq \rho(X, Y) \leq 1$.

Invariance before linear transformations: For any $a, b \in \mathbb{R}$ holds that $\rho(aX + b, aY + b) = \rho(X, Y)$, and that $\rho(aX + b, -aY + b) = -\rho(X, Y)$.

If X and Y are independent, then they are uncorrelated, e.g., $\rho(X, Y) = 0$. The converse is not necessarily true, but if both X and Y are normal random variables, then it holds.

When X and Y belong to the same random process, Eq. (1) is called “autocorrelation.” In this case, instead of X and Y , one usually denotes the random variables with an index, e.g., X_t and X_u , and their correlation as $\rho_{t,u}$.

Discussion

Time Series

Consider an infinite series of real-valued random variables $X = (\cdots, X_{t-2}, X_{t-1}, X_t, X_{t+1}, X_{t+2}, \cdots)$. A general model allows both the mean $E(X_u) = \mu_u$ and the variance $\text{Var}(X_u) = \sigma_u^2$ to vary with the index u , and the covariance $\text{Cov}(X_u, X_v)$ to vary with the pair of indexes (u, v) .

We will make three assumptions in order to facilitate our presentation:

H1: The (finite) mean does not vary with the index: $\mu_u = \mu < \infty$ for every $u \in \mathbb{Z}$.

H2: The (finite) variance does not vary with the index: $\sigma_u^2 = \sigma^2 < \infty$ for every $u \in \mathbb{Z}$.

H3: The covariance depends only on the absolute value of the difference between the indexes: $\text{Cov}(X_u, X_v) = \text{Cov}_k$, where $k = |u - v|$ for every $(u, v) \in \mathbb{Z} \times \mathbb{Z}$. This absolute difference k is known as “lag,” and Cov_k is referred to as “autocovariance.”

A random process X that satisfies these three hypotheses is called “weakly stationary.”

The autocorrelation of X is

$$\rho_k = \frac{\text{Cov}_k}{\sigma^2} = \frac{\mathbb{E}[(X_{t+k} - \mu)(X_t - \mu)]}{\sigma^2}, \quad (2)$$

which, by the hypotheses above, does not depend on t . Notice that $\rho_0 = 1$.

Finlay et al. (2011) provide a comprehensive account of conditions for a function $\rho_k : \mathbb{Z} \rightarrow [-1, 1]$ to be a valid autocorrelation function (ACF). Such a characterization is primarily helpful for theoretical purposes.

In practice, one encounters finite time series $\mathbf{x} = (x_1, x_2, \dots, x_n)$ of length n , with which one has to make several types of inference. If we are interested in estimating the autocorrelation, we may use the estimator of the autocovariance

$$\widehat{\text{Cov}}_k = \frac{1}{n} \sum_{t=1}^{n-k} (x_t - \bar{\mathbf{x}})(x_{t+k} - \bar{\mathbf{x}}), \quad (3)$$

where $\bar{\mathbf{x}}$ is the sample mean of the whole time series, and then obtain the estimator of the autocorrelation

$$\hat{\rho}_k = \frac{\widehat{\text{Cov}}_k}{\widehat{\text{Cov}}_0}. \quad (4)$$

Notice that $\widehat{\text{Cov}}_0$ is the sample variance of the whole time series.

If $\rho_k = 0$, then $\hat{\rho}_k$ is asymptotically distributed as a Gaussian random variable with mean $-1/n$ and variance $1/n$. This result

provides a simple rule for building approximate confidence intervals around 0.

Figure 1 shows 1000 observations of atmospheric noise (top) collected by <https://www.random.org>, along with the first 30 estimates of its autocorrelation function (ACF) $\hat{\rho}_1, \hat{\rho}_2, \dots, \hat{\rho}_{30}$ and the approximate confidence interval at the 95% level. As expected for this kind of noise, there is no evidence of serial correlation.

Figure 2 shows 870 measurements of the daily number of sun spots (top) downloaded from <https://www.ngdc.noaa.gov/stp/solar/ssndata.html>, along with the first 29 estimates of its autocorrelation function (ACF) $\hat{\rho}_1, \dots, \hat{\rho}_{29}$ and the approximate confidence interval at the 95% level. Seasonal observations like these often show significant autocorrelation values.

The reader is referred to the texts by Gilgen (2006) for specific applications to geosciences, and to the more general books by Box et al. (1994), Cryer and Chan (2008), and Cowpertwait and Metcalfe (2009).

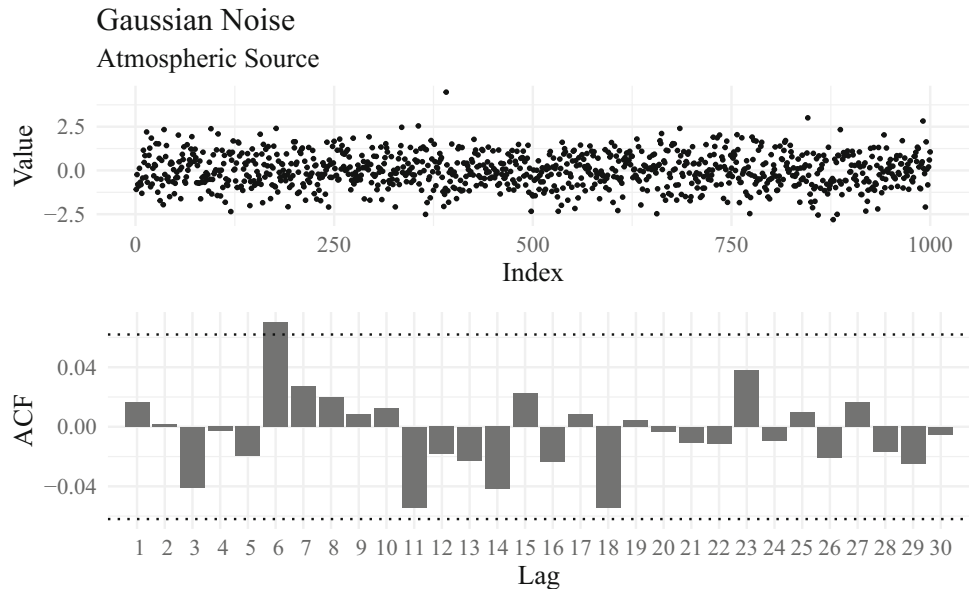
Extensions

Not rejecting a test for zero autocorrelation (or autocovariance) does not imply independence. Székely et al. (2007) discuss extensions (distance autocorrelation and distance covariance) which are zero only if the random vectors are independent.

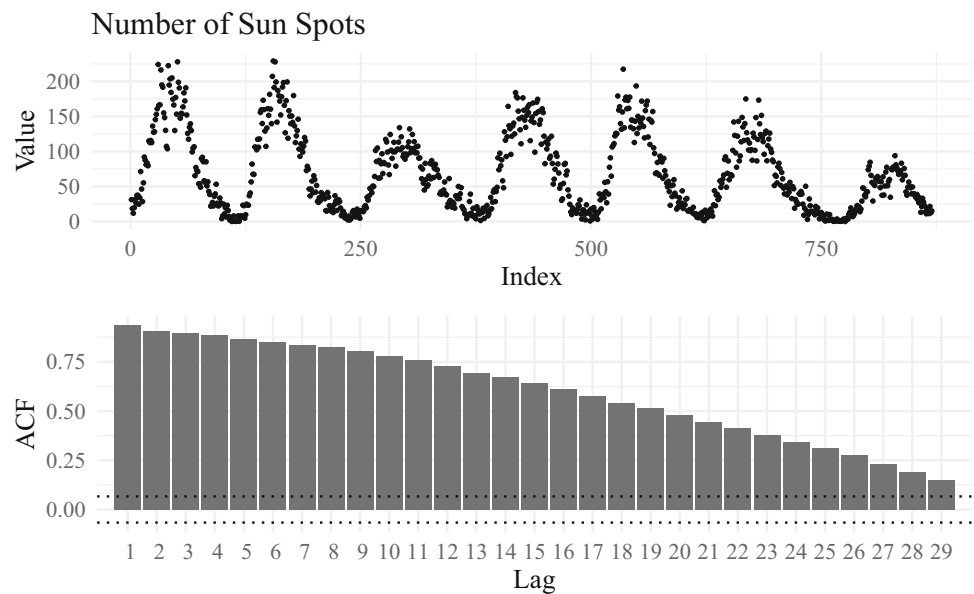
Autocorrelation and autocovariance are defined over real-valued (univariate) random vectors. Székely and Rizzo (2009) define a new measure, the Brownian Distance Covariance, that compares vectors of arbitrary dimension.

Autocorrelation,

Fig. 1 Gaussian atmospheric noise



Autocorrelation,
Fig. 2 Number of sun spots



Summary

Autocorrelation functions are powerful tools for an exploratory analysis of the dependence structure of time series.

Cross-References

- [Standard Deviation](#)
- [Univariate](#)

Bibliography

- Box GEP, Jenkins GM, Reinsel GC (1994) Time series analysis: forecasting and control, 3rd edn. Prentice, Englewood Cliffs
- Cowpertwait PSP, Metcalfe AV (2009) Introductory time series with R. Springer, New York
- Cryer JD, Chan KS (2008) Time series analysis with applications in R, 2nd edn. Springer, New York
- Finlay R, Fung T, Seneta E (2011) Autocorrelation functions. *Int Stat Rev* 79(2):255–271. <https://doi.org/10.1111/j.1751-5823.2011.00148.x>
- Gilgen H (2006) Univariate time series in geosciences. Springer, Berlin. https://www.ebook.de/de/product/11432954/hans_gilgen_univariate_time_series_in_geosciences.html
- Székely GJ, Rizzo ML (2009) Brownian distance covariance. *Ann Appl Stat* 3(4). <https://doi.org/10.1214/09-AOAS312>. <http://projecteuclid.org/euclid.aoas/1267453933>
- Székely GJ, Rizzo ML, Bakirov NK (2007) Measuring and testing dependence by correlation of distances. *Ann Stat* 35(6):1–31. <https://doi.org/10.1214/009053607000000505>

Automatic and Programmed Gain Control

Gabor Korvin

Earth Science Department, King Fahd University of Petroleum and Minerals, Dhahran, Kingdom of Saudi Arabia

Definition

In the electronic *acquisition* of geoscientific data, and their subsequent digital *processing*, we frequently subject the data to *Gain Control* which is a nonlinear transformation to maintain an approximately *stationary signal* of a prescribed Root Mean Square (RMS) amplitude at its output, despite time-dependent slow variations of the signal amplitude level at the input. In simple terms, Gain Control reduces the signal if it is strong and amplifies it when it is weak. There are two ways to control signal amplitude, *automatically* or in a preset, *programmed* manner. *Automatic Gain Control* (AGC, also called *Automatic Volume Control* (AVC), *Automatic Level Control* (ALC), *Time-Varying Gain* (TGC)) is realized by an algorithm where the output energy level in a sliding time-window controls the gain applied to the input to keep it within prescribed limits. *Programmed Gain Control* (PGC) applies a gain which is a function of record time and was determined beforehand. In some AGC or PGC systems the gain can only vary by a factor of two (*Binary Gain Control* BGC). Gain Control (AGC or PGC) is an important step of data processing to improve the visibility of such data where propagation effects (such as *attenuation* or *spherical divergence*) have

caused amplitude decay, and to assure approximate *stationarity* which is a prerequisite for *statistical time-series analysis*.

Introduction: Stationary and Nonstationary Processes, Gain Control

Before delving into the treatment of *Gain Control*, we must discuss the concepts of *stationarity* and *nonstationarity* of a *stochastic process*. Stationarity of a time series has become considered – paradoxically – of crucial importance in analyzing Earth data (Myers 1989), even though most dynamic geoprocesses (such as “expansion of the Universe, the evolution of stars, global warming on the Earth, and so on,” Guglielmi and Potapov 2018) are actually nonstationary. In simple terms, stationarity means that the statistical properties of a time series (more exactly: of the *stochastic process* generating the time series) do not change over time. In other words, a stationary time series has statistical properties (such as mean, variance, and autocovariance) that do not change under time translation. Stationary processes are easier to analyze, to display, to model, and their future behavior is in a certain sense predictable because their statistical properties will be the same in the future as they have been in the past. There are different notions of stationarity, such as (Myers 1989, Nason 2006): (a) *Strong stationarity* (also called *strict stationarity*) requires that the joint distribution of any finite sub-sequence of random variables of the process remains the same as we shift it along the time index axis: A stochastic process $\{Y_t\}_{t=-\infty}^{\infty}$ is strongly stationary if, for any integer n , and any set of time instants $\{t_1, t_2, \dots, t_n\}$ the joint distribution of the random variables $(Y_{t_1-t}, Y_{t_2-t}, \dots, Y_{t_n-t})$ depends only on $\{t_1, t_2, \dots, t_n\}$ but not on t . (b) *Weak stationarity* (also called *covariance stationarity*) only requires the shift-invariance of the mean and the autocovariance: a process is *weakly stationary* or *covariance stationary* if (letting E denote expected value).

$$\begin{aligned} E(Y_t) &= \mu \neq \pm\infty \quad \forall t \\ \text{var}(Y_t) &= \sigma^2 < \infty \quad \forall t \\ \text{cov}(Y_t, Y_{t-\tau}) &= E[(Y_t - \mu)(Y_{t-\tau} - \mu)] = \gamma_\tau \neq \pm\infty \quad \forall t, \forall \tau \end{aligned} \quad (1)$$

(c) In *geostatistics* (Myers 1989) they often assume an even weaker property than covariance stationarity, namely the so-called *intrinsic hypothesis* which assumes the stationarity of the semivariogram, and holds for a stochastic process $\{Y_t\}_{t=-\infty}^{\infty}$ if (i) the expected difference between values at any two time instants separated by a distance τ is zero: $E[Y_t - Y_{t-\tau}] = 0$ for $\forall \tau$ and (ii) the variance of these

differences, $\text{var}[Y_t - Y_{t-\tau}]$ is finite and depends only the distance τ .

A well-known example for a stationary process is the *White Noise process*: it is a stochastic process with a constant mean of zero and a constant and finite variance, whose values at any two different time instants are uncorrelated. Formally, the process $W = \{W_t\}_{t=-\infty}^{\infty}$ is a white noise if: $\forall t E(W_t) = 0$ (Eq. 2a); $\forall t E[(W_t - \mu)^2] = E(W_t^2) = \sigma^2 < \infty$ (Eq. 2b); $\forall t_1, t_2, t_1 \neq t_2 E(W_{t_1} \cdot W_{t_2}) = \text{cov}(W_{t_1}, W_{t_2}) = 0$ (Eq. 2c).

Note that this is a *weakly stationary process*. If, in addition, for every t the value W_t follows a normal distribution with zero mean and variance σ^2 , then the process is a *Gaussian White Noise process*. A White Noise process $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ with zero mean and variance σ^2 is conveniently denoted as $\varepsilon_t \sim WN(0, \sigma^2)$ (if it is Gaussian, we write $\varepsilon_t \sim GWN(0, \sigma^2)$). Using such a process, we can recursively construct a famous *nonstationary* process, the so-called *Random Walk*, for the time instants $t = 0, 1, 2, \dots$, as follows: Let Y_0 be an arbitrary real number, and $\forall t Y_t = Y_{t-1} + \varepsilon_t$. We get, step by step, for $t = 1, 2, \dots$ that $Y_1 = Y_0 + \varepsilon_1$; $Y_2 = Y_1 + \varepsilon_2 = Y_0 + \varepsilon_1 + \varepsilon_2$; $\dots$; and $Y_n = Y_0 + \sum_{i=1}^n \varepsilon_i$ (Eq. 3). Using Eq. 3 and the properties (Eqs. 2a, b, c) of $\{\varepsilon_t\}_{t=-\infty}^{\infty}$, easy calculation gives that $E(Y_n) = Y_0$ (Eq. 4a) which is independent of the time-variable n ; however, the variation $\text{var}(Y_n) = E\left[\left(\sum_{i=1}^n \varepsilon_i\right)^2\right] = n \cdot \sigma^2$ (Eq. 4b) increases with n , that is the random walk $\{Y_0, Y_1, \dots\}$ is not stationary.

A frequently occurring type of nonstationary processes are those having a *deterministic trend* (or “drift”). For example, let s_t be stationary with zero mean, α, β given constants, then the process $Y_t = \alpha + \beta t + s_t$ (Eq. 5) is nonstationary, because its mean $E(Y_t) = \alpha + \beta t$ depends on t . This kind of nonstationarity can be easily eliminated if we find α, β by *least mean square fitting*, and then simply subtract off the trend: $X_t = Y_t - \alpha - \beta t \approx s_t$ (Eq. 6). (Here, the approximate equality, “ $\approx$ ”, sign is a reminder of the issues arising from the uncertainty in the fitting.)

Usually, the nonstationary processes that we encounter in Geosciences cannot be such easily transformed into stationary processes. Sometimes a nonstationary time-series consists of a number of time segments that are themselves stationary. Such a time-series could be easily transformed to a stationary one by finding, and eliminating, the individual mean values of the segments, and then determining the proper multiplying factors to equalize the variances. Alas, no feasible computational algorithm has been found yet to find these locally stationary segments and locate their boundaries (Fukuda

et al. 2004), even to decide whether a segment of a time-series is stationary or not is a hot research question in statistics (Palachy 2019).

In many cases records of Earth processes and geophysical signals are described by *locally stationary* (LS) processes (Nason 2006). Around each time point they are close to a stationary process but their local mean and covariance are slowly changing in an *unspecific* way as time evolves. Such a record can be modeled as $x(t) = a(t) + \beta(t)s(t) + \varepsilon(t)$ (Eq. 7) where $x(t)$ is the observed LS process, $a(t)$ is a *slowly varying additive drift*, $\beta(t)$ is a *slowly varying* strictly positive *multiplying factor*, $s(t)$ is the original, stationary signal with zero mean value and unit variance, and $\varepsilon(t) \sim WN(0, \sigma^2)$ is additive white noise which is uncorrelated with $s(t)$. The drift $a(t)$ can be eliminated by *differencing* (as in Eq. 6) or by HP (high pass) *filtering*, then the record transforms to $x^*(t) \approx \beta(t)s(t) + \varepsilon^*(t)$ (Eq. 8), where $\varepsilon^*(t) \sim WN(0, \sigma_1^2)$ is the transformed noise. If $\beta(t)$ is known (from visual observation, or physical arguments) and we can assume that for $\forall t$ we have $|\sigma_1| \ll \beta(t)$ (Eq. 9), then the stationary, normalized, zero-mean earth signal $s(t)$ can be estimated from Eq. 8 as $s(t) \approx \frac{x^*(t)}{\beta(t)}$ (Eq. 10). This procedure is called *Programmed Gain Control* (PGC), and – importantly – it requires the knowledge of $\beta(t)$. PGC will be discussed, for the case of seismic data, in the next

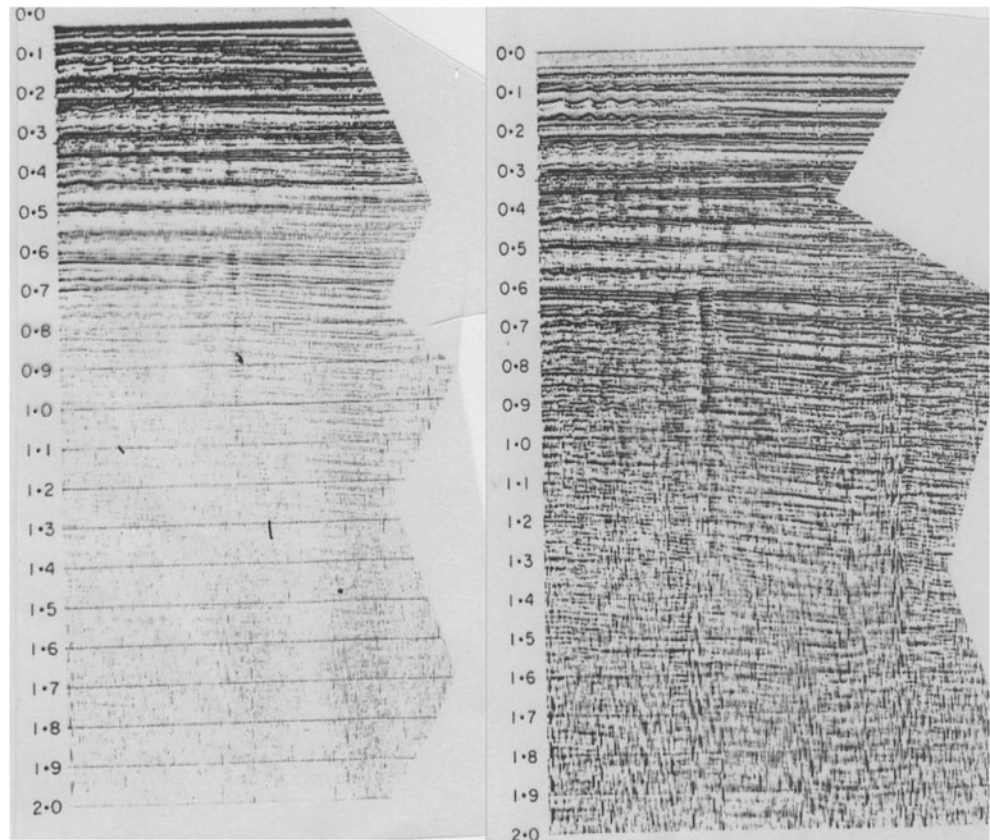
Sect. 2. If $\beta(t)$ is not known, first of all we have to estimate it from Eq. 8, and then apply it as in Eq. 10 to find $s(t)$. This procedure is called *Automatic Gain Control* (AGC), and will be discussed in Sect. 3. Examples for both kinds of gain will be taken from *seismic data processing*.

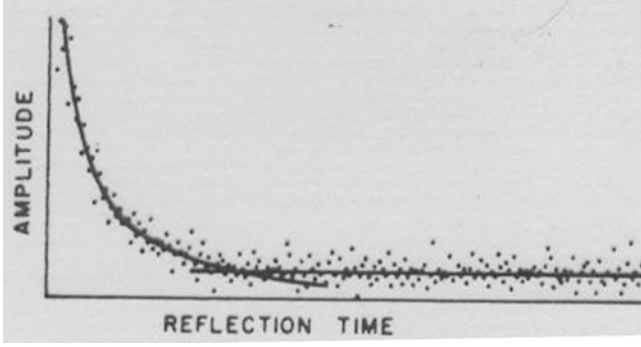
Programmed Gain Control (“True Amplitude Recovery”) of Seismic Data

Without any gain control the seismic amplitudes for large arrival times, and/or large offsets, would be too small to display. Compare the two seismic profiles on Fig. 1, the left without, the right with, amplitude regulation.

The decay of seismic amplitudes with time, illustrated in Fig. 2, has many reasons, such as *geometric spreading*, *reflection losses* at boundaries, *inelastic absorption*, *wave conversion* ($P \rightarrow S$, $S \rightarrow P$) at oblique incidence, energy scattered out of the wave path, etc. Because of these effects the seismic amplitude soon sinks below the ambient noise level. An important seismic processing step, *True Amplitude Recovery* applies a special *Programmed Gain Control* to the recorded traces to compensate for amplitude losses caused by *wavefront spreading* and *attenuation*. Seismic processing

Automatic and Programmed Gain Control, Fig. 1 Seismic profile without, and with, Automatic Gain Control

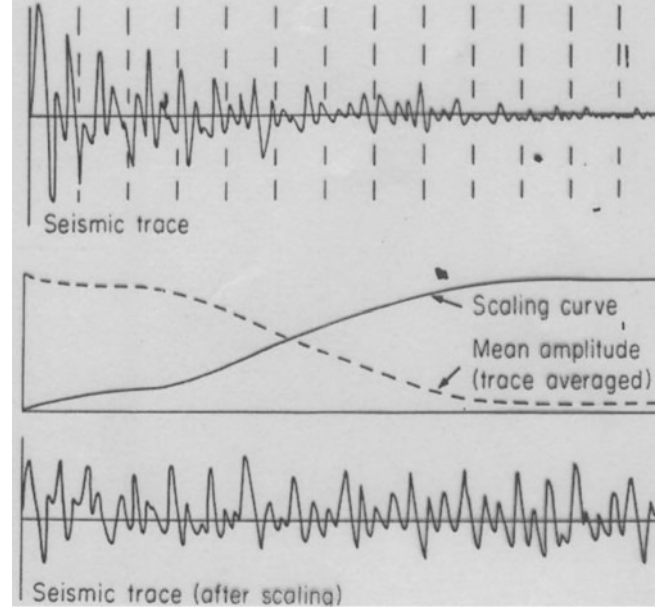




Automatic and Programmed Gain Control, Fig. 2 Amplitude decay of seismic data due to propagation effects

software packages offer a number of gain control possibilities to correct for spherical divergence and inelastic attenuation effects.

Geometric spreading means that the energy of the source is spread over a hemisphere in case of isotropy and constant velocity, or over a surface of more complicated shape in anisotropic formations (Newman 1973, Yilmaz 2001), that is the average RMS amplitude decays as $\sim 1/\text{distance}$ $\sim \frac{1}{[t_0 v(t_0)/\tau v(\tau)]}$ where t_0 is 2-way travel-time, $v(t_0)$ the RMS stacking velocity, τ some arbitrary time instant; that is the multiplicative gain to compensate this effect is $g(t_0) = \text{const} \cdot [t \cdot v(t_0)t_0 v(t_0)/\tau v(\tau)]$ (Eq. 11a). Some companies use a gain $g(t_0) = \text{const} \cdot [t_0 v^2(t_0)/\tau v^2(\tau)]$ (Eq. 11b) instead. There also exists an other kind of geometric loss, the *reflection loss*. If a wave crosses two times a boundary with reflection coefficient r_i , once downwards and once upwards, both times at normal incidence, the wave amplitude gets reduced by a factor $1 - r_i^2$, where $\{r_i\}, E(r_i) = 0, \text{var}(r_i) = E(r_i^2) = r^2 \ll 1$ is the series of *reflection coefficients*. Assuming that the reflecting boundaries are *Poisson-distributed* in a sense that a number λ of them are crossed by the propagating wave during one sec two-way time, the reflection loss is exponential: $\text{loss}_{\text{refl}} = \exp(-\lambda r^2 t_0)$ (Eq. 12). Combining this with the inelastic loss $\text{loss}_{\text{absorption}} = \exp(-\alpha t_0)$ where α is the *absorption coefficient*, the total decay is $\text{loss}_{\text{total}} = \exp(-\beta t_0)$ where $\beta = \lambda r^2 + \alpha$, and to compensate it one needs a predetermined gain $g(t_0) = \text{const} \cdot \exp(\beta t_0)$ (Eq. 13). In some processing software (as, e.g., ProMAX[®] 2004) the gain Eq. 13 can be also specified in dB/sec units in the form $g(t_0) = \text{const} \cdot 10^{C t_0/20}$ (Eq. 14) where the gain-factor C is in dB/sec. The Programmed Gain to compensate both for wave-front spreading and attenuation is $g(t_0) = \text{const} \cdot [t \cdot v(t_0)t_0 v(t_0)/\tau v(\tau)] \cdot \exp(\beta t_0)$ (Eq. 15). If the RMS stacking velocity $v(t_0)$ and the attenuation factor β are not known, we resort to the approximate, heuristic “Time Power compensation,” realized by $g(t_0) = \text{const} \cdot t_0^\gamma$ (Eq. 16) where in seismic practice the exponent γ is between 1 and 2, by default $=2$. The *normalizing factors* in Eqs. 11 a, b; 13; 14; 15; 16 are used to maintain a prescribed RMS amplitude



Automatic and Programmed Gain Control, Fig. 3 Time-variant scaling of a seismic trace

level, as well as for *trace balancing* if a group of traces are jointly subjected to PGC. By proper selection of these factors, each trace in the group of traces can be scaled so that they all have the same desired RMS amplitude level.

Automatic Gain Control (AGC)

The easiest way to subject a group of seismic traces to AGC is to apply *time-variant scaling* (Fig. 3). As seen in the Figure, we (i) divide all traces in the group to N time segments (“windows”) of equal length, each having K samples; (ii) compute the local mean value $\mu_i = \frac{1}{K} \sum_{n \in W_i} x_n$ in each window W_i and subtract it from the relevant data; (iii) compute the RMS (Root Mean Square) amplitude for each window: $a_i = \sqrt{\frac{1}{K} \sum_{n \in W_i} (x_n - \mu_i)^2}$; construct a smooth function $a(t)$ (with *spline approximation*) passing through the points $a(t_1) = a_1, a(t_2) = a_2, \dots, a(t_N) = a_N$ where $t_1, t_2, \dots, t_N$ are midpoints of the windows; (iv) take the average of the $a(t)$ curves over all traces to get an average decay curve $A(t)$ (see the second curve in Fig. 3); and (v) apply the gain $g(t) = \frac{\text{const}}{A(t)}$ (Eq. 17). Figure 3 illustrates this on a single seismic trace.

A similar, but more powerful AGC algorithm utilizes *sliding windows* to avoid sudden jumps in the gain curve. Automatic gain control (AGC) is the most common way of gain recovery in seismic processing. AGC is applied to the seismic data on a trace-by-trace basis using a sliding time window of length $N \cdot \Delta t$ (Δt is the sampling interval, in seismic practice, usually 2 or 4 msec, N is of the order 50). The window is progressively moved along the time axis

sample-by-sample. Suppose the window consists of $N = 2M + 1$ data, the seismic values $\{x_i\}$ are sufficiently well “de-measured” (i.e., their local mean value had been removed by High Pass filtering), such that for $\forall i = 1, 2, \dots, M \leq i \leq RL - M$ where RL is record length in Δt units, we have $\frac{1}{2M+1} \sum_{k=i-M}^{i+M} x_k \approx 0$ (Eq. 18). Then, for all i the average energy within the moving window $W_i = \{i-M, i-M+1, \dots, i-M+1, i-M+1, \dots, i-1, i, i+1, \dots, i+M-1, i+M\}$ is computed as $E_i = \frac{1}{2M+1} \sum_{k=i-M}^{i+M} x_k^2$ (Eq. 19), and the sample x_i in the center of this window is divided by this energy to yield the result of AGC: $y_i = \frac{1}{\frac{1}{2M+1} \sum_{k=i-M}^{i+M} x_k^2} \cdot x_i = g(i) \cdot x_i$ (Eq. 20),

where $g(i) = \frac{1}{E_i}$ is the gain factor. Instead of average energy (E_i , see Eq. 19) we could also use *RMS amplitude* $A_{RMS,i} = \sqrt{\frac{1}{2M+1} \sum_{k=i-M}^{i+M} x_k^2}$, *average absolute amplitude* $A_{abs,i} = \frac{1}{2M+1} \sum_{k=i-M}^{i+M} |x_k|$, or *median absolute amplitude* $A_{med,i} = \text{median}\{|x_i - M|, \dots, |x_i|, \dots, |x_i + M|\}$ (Eqs. 21 a,b,c) to construct a gain function $g(i)$.

AGC can be *fast* or *slow*. A very short window might eliminate geologically meaningful amplitude variations, and does not necessarily fulfill condition (Eq. 18). The processing geophysicist must ensure that the window is long enough to include many reflectors so that their relative amplitude ratio should remain visible after gain regulation.

A Few Concluding Remarks on the Use of PGC and AGC

- (i) In the 1950s and before the analog seismic recorders used PGC. When digital recorders became widespread in the 1960s they already used digital *BGC* (*Binary Gain Control*), but they needed analog *preamplifiers* and an *analog PGC unit* to fully utilize the limited dynamic range of their *AD* (*Analog-to-Digital*) *converters*. Modern recorders use *IFP* (*instantaneous-floating-point*) *gain control*, where the gain is independently determined for each sample on each channel, based on the amplitude of that sample without reference to earlier samples or other channels.
- (ii) Automatic Gain Control must be used with utmost care, since it can destroy signal character. In seismics, an AGC with a very short sliding window makes strong reflections indistinguishable from weak reflections.
- (iii) If we want to preserve the relative amplitude ratios between nearby seismic traces, as for example when we pre-process our seismic material for *AVO* (*Amplitude versus Offset*) studies to find lithologic information from

seismic amplitudes, then no gain control at all, or only some very carefully designed PGC should be used.

- (iv) Some rock physical properties (such as electric- and heat conductivity and hydraulic permeability) have such a huge dynamic range that only their logarithms can be displayed as functions of depth (as, e.g., an electric well log trace). Such records of logarithmic data might look stationary, but the “log transform,” that is taking the logarithm of the data, is not the proper way, and not recommended, to improve the stationarity of a time series!

Cross-References

- Data Acquisition
- Data Visualization
- Digital Filters
- Exploratory Data Analysis
- Geocomputing
- Geoscience Signal Extraction
- Moving Average
- Scaling and Scale Invariance
- Signal Analysis
- Signal Processing in Geosciences
- Splines
- Stationarity
- Time Series Analysis
- Time Series Analysis in the Geosciences

References

- Fukuda K, Eugene Stanley H, Nunes Amaral LA (2004) Heuristic segmentation of a nonstationary time series. *Phys Rev E* 69:021108
- Guglielmi AV, Potapov AS (2018) On the nonstationary processes in geophysical media. *Geophys Res* 19(4):5–15
- Myers DE (1989) To be or not to be . . . stationary? That is the question. *Math Geol* 21:347–362
- Nason GP (2006) Stationary and non-stationary time series. In: Mader H, Coles SC (eds) *Statistics in volcanology*, Special publications of the International Association of Volcanology and Chemistry of the Earth's Interior, 1. The Geological Society, London, pp 129–142
- Newman P (1973) Divergence effects in a layered earth. *Geophysics* 38(3):481–488
- Palachy S (2019) Detecting stationarity in time series data. <https://towardsdatascience.com/detecting-stationarity-in-time-series-data-d29e0a21e638>
- ProMAX (2004) *Seismic processing and analysis training manual*, volume 1 copyright © 2004 by Landmark Graphics Corporation Part No 162382 Rev A, October 2004
- Yilmaz Ö (2001) *Seismic data analysis*, Investigations in Geophysics 10. Society of Exploration Geophysicists, Tulsa

Backprojection Tomography

Utkarsh Gupta¹ and Uma Ranjan²

¹Department of Computational and Data Sciences, Indian Institute of Science, Bengaluru, Karnataka, India

²Indian Institute of Science, Bengaluru, Karnataka, India

Definition

X-Ray computed tomography (CT) can visualize the internal structures of any material in three dimensions. Due to its capability of 3D visualization in a non-destructive fashion, it has found its application in various fields like material, geological, food, and paleontological sciences (Withers et al. 2021) [2, 10]. For instance, X-Ray microCT has revolutionized the field of Digital Rock Physics (Balcewicz et al. 2021) [1] by drastically reducing the decision time, compared to special core analysis laboratory (SCAL) measurements to estimate all the rock properties. Backprojection is one of the oldest and simplest reconstruction methods which exist in literature. In seismology, a simple backprojection algorithm has been applied to generate tomographic images of P-velocity or the seismic transparency for the section of rock between the borehole (Wong et al. 1983) [9]. Although backprojection method is not practically adopted in its basic form for reconstruction, it still continues to form the basis of various practically used reconstruction algorithms. One of the commonly adopted reconstruction algorithms is the filtered backprojection algorithm which is the main focus of this chapter. Filtered backprojection algorithm belongs to a class of analytical reconstruction techniques and has been widely adopted due to its ease of implementation and reconstruction speed.

Introduction

Tomographic image reconstruction is a problem where a 3-D object is reconstructed from its 2-D projections. One way to obtain projections is through transmission of a beam that passes through the object and is detected at the other end through an array of detector elements.

The intensity of the beam measured at the detector depends on the material through which the beam passes and is attenuated by the material by the formula

$$\frac{I}{I_0} = e^{-\int \alpha ds}$$

where I_0 is the intensity of the beam at the source that is incident on the object, α is the local linear attenuation, and s is the path along which the beam travels. The attenuation constant is a material property dictated by absorption and scattering. The source and detector are 180 degrees apart and form a part of a single assembly. This assembly is rotated to obtain projections at different angles.

Practical arrangements of the source–detector assembly differ and are of primarily three types:

- Parallel-beam tomography.
- Fan-beam tomography.
- Cone-beam tomography.

We have described algorithms for all the three source-detector assembly along with their mathematical formulation in the subsequent sections.

Parallel-Beam Image Reconstruction

Here, we describe the parallel-beam reconstruction algorithm, which form the basis for deriving the results for the more complex and practically used setup – fan-beam reconstruction

and cone-beam reconstruction algorithms. In what follows, we will learn about radon transform, followed by Fourier slice theorem and filtered backprojection (FBP) algorithm.

Radon Transform

Backprojection-based tomographic reconstruction technique can be mathematically explained in terms of radon transform (initially proposed for astronomical-based application in 1917) (Radon 1917; Ludwig 1966). Radon transform expresses the function values ($f(x, y)$) in terms of projection data ($g_\theta(x')$), which is nothing but a line integral. Given a 2-D object, $f(x, y)$ under study can also be expressed in the rotated coordinate system by $s(x', y')$ (see Fig. 1). Following this notation, we present the mathematical formulation for the radon transform subsequently. A line integral will represent the ray sum/integral of $f(x, y)$ along a given line. For CT based application, this ray sum represents the total attenuation that the X-rays undergo while passing through the object under study. Each line integral can be easily expressed in terms of two parameters (θ, x'), where θ represents the angle at which X-rays are projected onto the object or the angle at which the detector is aligned to measure X-ray attenuation. Line integral ($g_\theta(x')$) is expressed as.

$$g_\theta(x') = \int_{y'=-\infty}^{\infty} s(x', y') dy' \quad (1)$$

The relation between $x - y$ coordinate system and the rotated $x' - y'$ coordinate system can be easily expressed using rotation matrices ($A_\theta/A_{-\theta}$) as

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix} = A_\theta \begin{bmatrix} x' \\ y' \end{bmatrix} \quad (2)$$

$$x = x' \cos \theta - y' \sin \theta \text{ \& } y = x' \sin \theta + y' \cos \theta$$

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = A_{-\theta} \begin{bmatrix} x \\ y \end{bmatrix} \quad (3)$$

$$x' = x \cos \theta + y \sin \theta \text{ \& } y' = -x \sin \theta + y \cos \theta$$

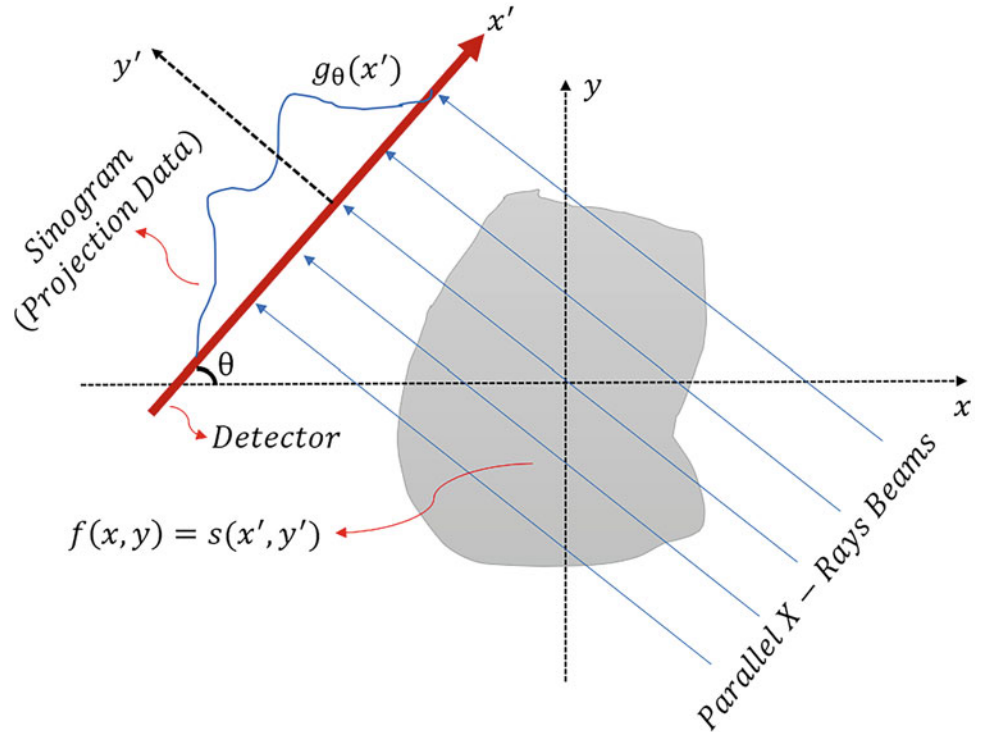
Using Eqs. 2 and 3, we can easily express the object $f(x, y)$ in the rotated $x' - y'$ coordinated axis as

$$\begin{aligned} f \begin{bmatrix} x \\ y \end{bmatrix} &= f(x, y) = f(x' \cos \theta - y' \sin \theta, x' \sin \theta + y' \cos \theta) \\ &= s(x', y') \end{aligned} \quad (4)$$

Combining Eq. 4 with Eq. 1, we arrive at the final expression of the radon transform expressed as

Backprojection Tomography,

Fig. 1 Illustration of the geometry of a set of parallel projection beams



$$\begin{aligned}
 g_\theta(x') &= \int_{y'=-\infty}^{\infty} f(x' \cos \theta - y' \sin \theta, x' \sin \theta + y' \cos \theta) dy' \\
 &= \int_{y'=-\infty}^{\infty} f\left(A_\theta \begin{bmatrix} x' \\ y' \end{bmatrix}\right) dy'
 \end{aligned} \quad (5)$$

where $g_\theta(x')$ is the radon transform of the object function $f(x, y)$ and graphic representation of radon transform known as sinogram in the CT imaging terminology.

Fourier Slice Theorem

In this section, we will learn about another vital concept needed to understand the backprojection reconstruction algorithm. Using Fourier slice theorem (Herman 1979, 1980; Scudder 1978), we establish the relation between the 1-D continuous-time Fourier transform (CTFT) of $g_\theta(x')$ and 2-D CTFT of the object $f(x, y)$ (see Fig. 2). We start the derivation by expressing Fourier transforms of $g_\theta(x')$ & $f(x, y)$ as

$$\begin{aligned}
 f(x, y) &\xrightarrow{\text{2D-CTFT}} F(w_x, w_y) \\
 &= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} f(x, y) e^{-j2\pi(w_x x + w_y y)} dx dy \quad (6)
 \end{aligned}$$

$$g_\theta(x') \xrightarrow{\text{1D-CTFT}} G_\theta(w) = \int_{x'=-\infty}^{\infty} g_\theta(x') e^{-j2\pi w x'} dx' \quad (7)$$

Further, using radon transform (Eq. 5), we can express $G_\theta(w)$ in terms of $f\left(A_\theta \begin{bmatrix} x' \\ y' \end{bmatrix}\right)$ as.

$$\begin{aligned}
 G_\theta(w) &= \int_{x'=-\infty}^{\infty} \left(\int_{y'=-\infty}^{\infty} f\left(A_\theta \begin{bmatrix} x' \\ y' \end{bmatrix}\right) dy' \right) e^{-j2\pi w x'} dx' \\
 &= \int_{x'=-\infty}^{\infty} \int_{y'=-\infty}^{\infty} f\left(A_\theta \begin{bmatrix} x' \\ y' \end{bmatrix}\right) e^{-j2\pi w x'} dx' dy'
 \end{aligned} \quad (8)$$

We now apply the change of variable to the above integral using the relation expressed in Eq. 3 as

$$f\left(A_\theta \begin{bmatrix} x' \\ y' \end{bmatrix}\right) = f\left(A_\theta A_{-\theta} \begin{bmatrix} x \\ y \end{bmatrix}\right) = f\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = f(x, y) \quad (9)$$

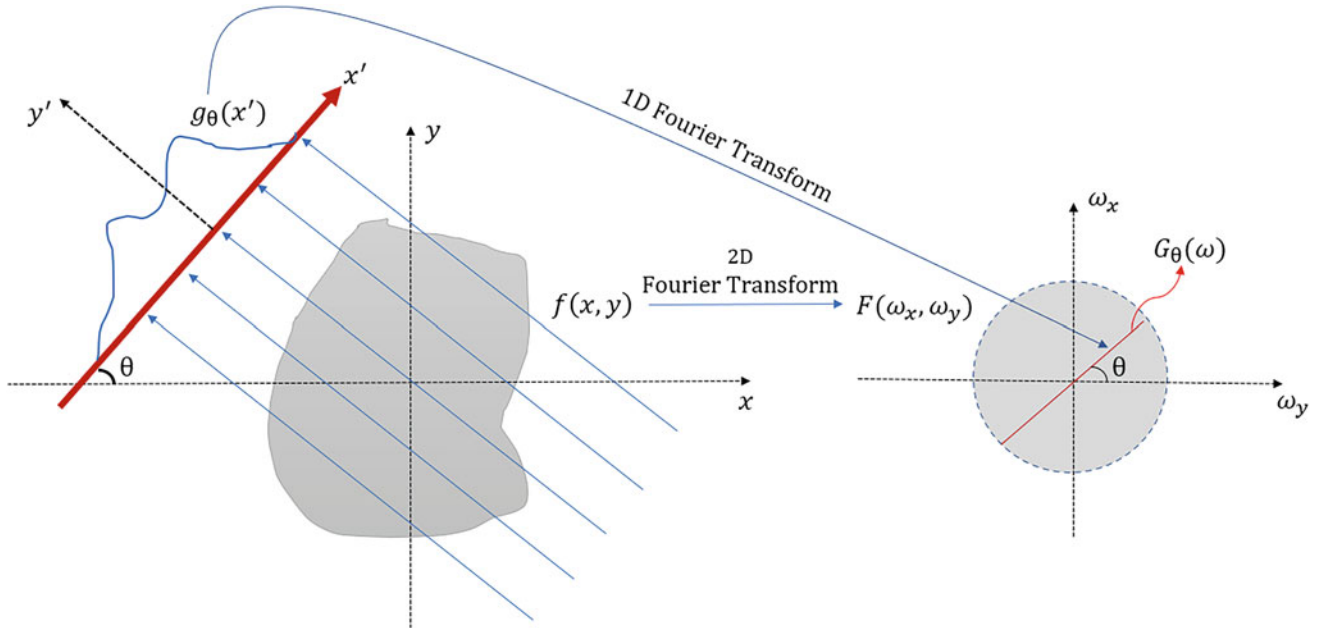
After applying the change of variable, we express Eq. 8 completely in terms of $x - y$ coordinate system as

$$G_\theta(w) = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} f(x, y) e^{-j2\pi w (x \cos \theta + y \sin \theta)} dx dy \quad (10)$$

Comparing Eq. 10 with Eq. 6, we derive the following equation:

$$G_\theta(w) = F(w \cos \theta, w \sin \theta) \quad (11)$$

The relation expressed in Eq. 11 is called Fourier slice theorem.



Backprojection Tomography, Fig. 2 Illustration of Fourier slice theorem

Filtered Backprojection Algorithm

We now describe the FBP algorithm that aims to recover the actual object $f(x, y)$ from the corresponding tomographic projections ($g_\theta(x')$) data. Object function $f(x, y)$ can be expressed using inverse CTFT relation as

$$f(x, y) = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} F(w_x, w_y) e^{j2\pi(w_x x + w_y y)} dw_x dw_y \quad (12)$$

We now apply the change of variable to the above integral using the following relation:

$$\begin{aligned} w_x &= w \cos \theta, w_y = w \sin \theta \\ dw_x dw_y &= |w| dw d\theta \end{aligned} \quad (13)$$

This results in the following equation:

$$f(x, y) = \int_{\theta=0}^{\pi} \int_{w=-\infty}^{\infty} F(w \cos \theta, w \sin \theta) e^{j2\pi w(x \cos \theta + y \sin \theta)} |w| dw d\theta \quad (14)$$

Using Fourier slice theorem (Eqs. 11 and 3), we get

$$f(x, y) = \int_{\theta=0}^{\pi} \left(\int_{w=-\infty}^{\infty} |w| G_\theta(w) e^{j2\pi w x'} dw \right) d\theta \quad (15)$$

where $|w|$ corresponds to a convolution in spatial domain. We define inverse Fourier transform of $|w|$ as $h(x')$. In spatial domain, Eq. 15 can be expressed as.

$$f(x, y) = \int_{\theta=0}^{\pi} (h(x') \otimes g_\theta(x')) \Big|_{x'=x \cos \theta + y \sin \theta} d\theta \quad (16)$$

which represents the sum of all the measured projection data ($g_\theta(x')$) convoluted ($\otimes$) with $h(x')$.

Complete FBP algorithm can be summarized as follows:

- Compute $g_\theta(x')$ for each projection of the object ($f(x, y)$).
- Convolve the each projection result as follows:

$$b_\theta(x') = (h(x') \otimes g_\theta(x'))$$

- Backproject all the above obtained filtered projections by computing

$$f(x, y) = \int_{\theta=0}^{\pi} b_\theta(x') \Big|_{x'=x \cos \theta + y \sin \theta} d\theta$$

Fan-Beam Image Reconstruction

The FBP algorithm described for the parallel-beam geometry is a robust and fast way of reconstructing the slices from the tomographic projection data. But at the same time, implementing a parallel-beam setup requires the installation of bulky hardware and needs the object to be exposed for a more extended period under an X-ray source. Using fan-beam geometry, it is possible to cleverly reduce the long acquisition time and avoid the need for bulky hardware. In the following sections, we will discuss two techniques, along with the mathematical framework, involved in fan-beam image reconstruction.

Reconstruction with Rebinning

In parallel-beam geometry, we use Fourier slice theorem to derive the FBP reconstruction algorithm. But, in the case of fan-beam geometry, we do not enjoy the luxury of any such theorem to derive reconstruction algorithms. Hence, we use a different strategy – converting a fan-beam geometric model to its equivalent parallel-beam geometric model. Figure 3 shows the comparison of two imaging geometries. It is intuitively evident that both parallel-beam and fan-beam geometries cover identical data in one complete rotation. The only difference lies in the data collection sequence in both the setups. Hence, once the entire data is collected in the fan-beam setup, we can rebin the entire data sequence in such a fashion so that the new data sequence corresponds as if the data was collected using a parallel-beam setup. Rebinning every fan-beam ray ($p(\gamma, \beta)$) into a parallel-beam ray ($g_\theta(x')$) is mathematically explained using Fig. 4a). The following relation can be easily established using a couple of mathematical steps:

$$\theta = \gamma + \beta \quad (17)$$

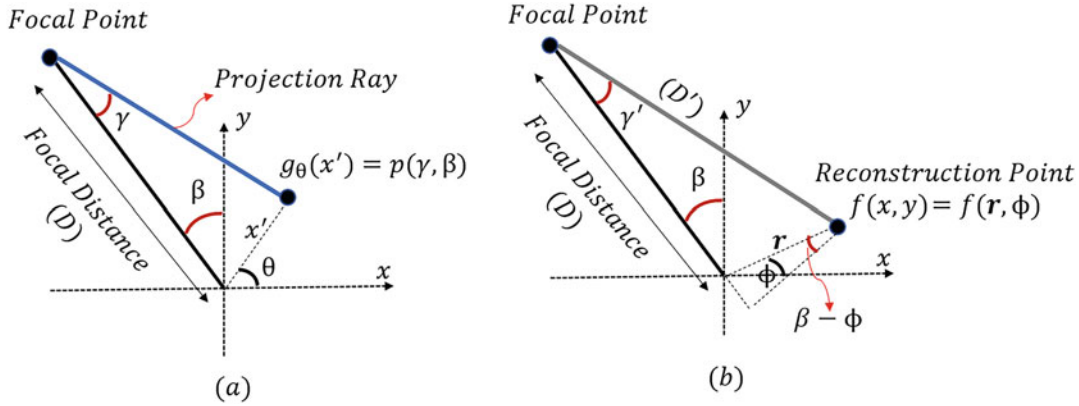
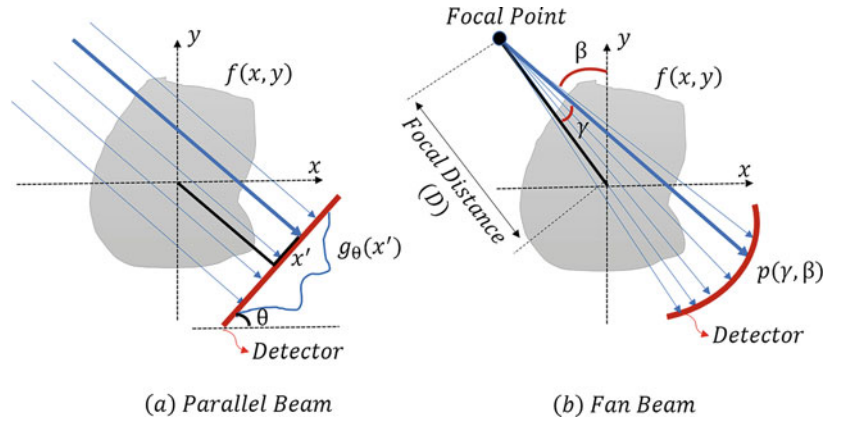
and

$$x' = D \sin \gamma \quad (18)$$

Using the above relation, we can conveniently convert parallel-beam ray sum into fan-beam ray sum and vice versa:

$$g_\theta(x') = p(\gamma, \beta) \quad (19)$$

After rebinning is completed, one can use FBP algorithm for parallel-beam geometry to perform image reconstruction.

Backprojection Tomography,**Fig. 3** Comparison of imaging geometries: (a) parallel-beam geometries; (b) fan-beam geometries**Backprojection Tomography, Fig. 4** (a) Conversion of fan-beam ray parameters into its corresponding parallel-beam ray parameters; (b) coordinate system for fan-beam imaging geometry**Reconstruction without Rebinning**

However, rebinning of fan-beam rays into parallel-beam rays is not a preferred technique as it requires the data interpolation while changing the coordinates. This data interpolation often leads to inaccurate image reconstruction. But with the relation described in Eqs. 17 and 18, we can derive the analytical reconstruction solution in terms of fan-beam geometry parameters. The technique is referred to as the “filtered back-projection fan-beam algorithm” (Weng et al. 1997) and will be described in the following section.

Filtered Backprojection fan-Beam Algorithm

We start the derivation with the final reconstruction equation obtained using FBP for the parallel-beam geometry, which is

$$f(x, y) = \frac{1}{2} \int_{\theta=0}^{2\pi} (h(x \cos \theta + y \sin \theta) \otimes g_{\theta}(x \cos \theta + y \sin \theta)) d\theta \quad (20)$$

where $\otimes$ denotes the convolution operator. We now express the above equation in the polar coordinate system by replacing $x = r \cos \phi$, $y = r \sin \phi$ and $x \cos \theta + y \sin \theta = r \cos(\theta - \phi)$, resulting in

$$f(r, \phi) = \frac{1}{2} \int_{\theta=0}^{2\pi} \int_{-\infty}^{\infty} g_{\theta}(x') h(r \cos(\theta - \phi) - x') dx' d\theta \quad (21)$$

Changing the variables of the above integral using Eqs. 17 and 18 by using the Jacobin required for variable change as $D \cos \gamma$ results in

$$f(r, \phi) = \frac{1}{2} \int_{\theta=0}^{2\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} p(\gamma, \beta) h(r \cos(\beta + \gamma - \phi) - D \sin \gamma) D \cos \gamma d\gamma d\beta \quad (22)$$

Equation 22 represents the analytical solution for the fan-beam geometry. But still it is not computationally efficient due to the absence of convolution operation. Next, we describe how convolution operation can be introduced in the

final solution, making it computationally efficient to implement.

We define two new parameters D' and γ' (see Fig. 4b) totally with respect to the point being reconstructed ($f(r, \phi)$). From Fig. 4b, we can derive the following relation:

$$r \cos(\beta + \gamma - \phi) - D \sin \gamma = D' \sin(\gamma' - \gamma)$$

which yields

$$f(r, \phi) = \frac{1}{2} \int_{\theta=0}^{2\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} p(\gamma, \beta) h(D' \sin(\gamma' - \gamma)) D \cos \gamma \, d\gamma \, d\beta \quad (23)$$

We now state a special property of a ramp filter (Eq. 24) and use it to arrive at the final solution:

$$h(D' \sin(\gamma)) = \left(\frac{\gamma}{D' \sin \gamma} \right)^2 h(\gamma) \quad (24)$$

If $h_{fan}(\gamma) = \frac{D}{2} \left(\frac{\gamma}{\sin \gamma} \right)^2 h(\gamma)$, then the final convolution-based fan-beam backprojection solution (Eq. 25) is stated as

$$f(r, \phi) = \int_{\theta=0}^{2\pi} \frac{1}{(D')^2} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} (\cos \gamma) p(\gamma, \beta) h_{fan}(\gamma' - \gamma) d\gamma \, d\beta \quad (25)$$

Cone-Beam Image Reconstruction

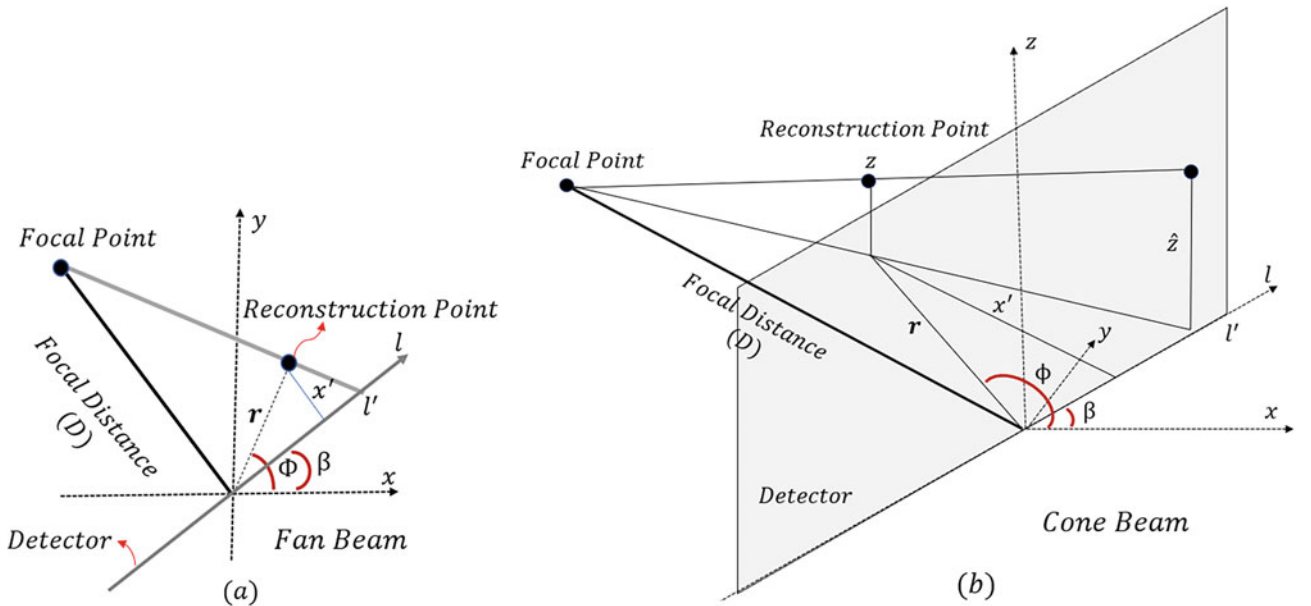
Cone-beam geometry-based reconstruction algorithm, in general, is considered to be more complex in comparison to parallel- or fan-beam geometry. But at the same time they are extremely popular and accurate in terms of reconstructed image quality. In the next section, we describe ‘‘Feldkamp’s algorithm,’’ which is one of the popular reconstruction methods in reconstruction community.

Feldkamp’s Algorithm

Feldkamp’s cone-beam reconstruction technique (also referred to as FDK algorithm) (Feldkamp et al. 1984) is basically the modification of fan-beam FBP algorithm. The coordinate system followed for implementing the algorithm is described in Fig. 5. We start by writing the FBP solution for fan-beam geometry with the flat detector in the polar coordinates (Fig. 5a) as

$$f(r, \phi) = \frac{1}{2} \int_{\theta=0}^{2\pi} \frac{D}{D - x'} \int_{-\infty}^{\infty} \frac{D}{\sqrt{D^2 + l^2}} p(l, \beta) h(l' - l) dl \, d\beta \quad (26)$$

where $h(l)$ represents the ramp filter, D represents the focal length, $g(l, \beta)$ is the fan-beam projection ray sum, l represents



Backprojection Tomography, Fig. 5 (a) Coordinate system followed for fan-beam geometry with flat detector; (b) coordinate system followed while implementing Feldkamp’s cone-beam algorithm

a linear coordinate on the flat detector, and $x' = r \sin(\phi - \beta)$ and $l' = \frac{Dr \cos(\phi - \beta)}{D - r \cos(\phi - \beta)}$.

Fan-beam algorithm can be modified into Feldkamp's algorithm if the backprojection operation is a cone-beam backprojection (see Fig. 5b). The filtering is carried out in a row-wise fashion. It is to be noted that no filtering is performed in the axial direction (z). The final Feldkamp's algorithm can be represented as

$$f(r, \phi, z) = \frac{1}{2} \int_{\theta=0}^{2\pi} \frac{D}{D - x'} \int_{-\infty}^{\infty} \frac{D}{\sqrt{D^2 + l'^2 + z^2}} p(l, \hat{z}, \beta) h(l' - l) dl \, d\beta \quad (27)$$

where $p(l, \hat{z}, \beta)$ is the cone-beam projection. All the other parameters are described in Fig. 5b.

Summary and Conclusions

This chapter covered computational reconstruction methods for all three geometries (Parallel-Beam, Fan-Beam & Cone-Beam). These methods are one of the prominent techniques and fall under a standard bucket of analytical reconstruction techniques. Due to their fast reconstruction output compared to iterative reconstruction methods, they have been widely adopted in many fields of geoscience like determining the porosity, permeability, and various other fluid flow properties in petroleum geology, rock mechanics, and soil science (Schlüter et al. 2020; Mees et. al 2003) [7, 8].

References

- Balcewicz M, Siegert M, Gurrus M, Ruf M, Krach D, Steeb H, Saenger EH (2021) Digital rock physics: a geological driven worklow for the segmentation of anisotropic ruhr sandstone. *Front Earth Sci* 9:493
- Cnudde V, Masschaele B, Dierick M, Vlassenbroeck J, Van Hoorebeke L, Jacobs P (2006) Recent progress in X-ray CT as a geosciences tool. *Appl Geochem* 21(5):826–832
- Feldkamp LA, Davis LC, Kress JW (1984) Practical cone beam algorithm. *J Opt Soc Am* 1(6):612–619
- Herman GT (ed) (1979) Image reconstruction from projections. Topics in applied physics, vol 32. Springer-Verlag, New York
- Herman GT (1980) Image reconstruction from projections-the fundamentals of computerized tomography. Academic Press, New York
- Ludwig D (1966) The Radon transform on Euclidean space. *Commun Pure Appl Math* 19:49–81
- Mees F, Swennen R, Van Geet M, Jacobs P (2003) Applications of X-ray computed tomography in the geosciences. *Geol Soc Lond Special Publ* 215(1):1–6
- Radon J (1917) Über die Bestimmung van Funktionen durch ihre Integralwerte tangs gewisser Mannigfaltigkeiten (On the determination of functions from their integrals along certain manifolds). *Berichte Saechsische Akad Wissenschaften (Leipzig), Math Phys Klass* 69:262–277
- Schlüter S, Sammartino S, Koestel J (2020) Exploring the relationship between soil structure and soil functions via pore-scale imaging. 370:114370
- Scudder HJ (1978) Introduction to computer aided tomography. *Proc IEEE* 66(6)
- Weng Y, Zeng L, Gullberg GT (1997) Analytical inversion formula for uniformly attenuated fan-beam projections. *IEEE Trans Nucl Sci* 44: 243–249
- Wilding M, Lesher CE, Shields K (2005) Applications of neutron computed tomography in the geosciences. *Nucl Instrum Methods Phys Res, Sect A* 542(1–3):290–295
- Withers PJ, Bouman C, Carmignato S, Cnudde V, Grimaldi D, Hagen CK, Maire E, Manley M, Du Plessis A, Stock SR (2021) X-ray computed tomography. *Nat Rev Methods Primers* 1(1):1–21
- Wong J, Hurley P, West GF (1983) Crosshole seismology and seismic imaging in crystalline rocks. *Geophys Res Lett* 10(8):686–689

Bayes's Theorem

Konstantinos Modis

School of Mining and Metallurgical Engineering, National Technical University of Athens, Athens, Greece

Synonyms

Bayes's law; Bayes's rule

Definition

In probability theory, Bayes's theorem can be formally stated according to the following equation:

$$P(\mathcal{A}|\mathcal{B}) = \frac{P(\mathcal{B}|\mathcal{A})P(\mathcal{A})}{P(\mathcal{B})} \quad (1)$$

where $\mathcal{A}$ and $\mathcal{B}$ are events, $P(\mathcal{A})$ and $P(\mathcal{B})$ are the probabilities of these events to occur with $P(\mathcal{B}) \neq 0$, and $P(\mathcal{A}|\mathcal{B})$ and $P(\mathcal{B}|\mathcal{A})$ are the conditional probabilities that $\mathcal{A}$ given $\mathcal{B}$ and $\mathcal{B}$ given $\mathcal{A}$ occur, respectively (Papoulis and Pillai 2002).

Application to Mineral Exploration or a “Mineral Adventure”

In mineral exploration, managers are often called upon to take decisions in extremely uncertain environments. As an example, consider the CEO of a small company that has been appointed to investigate an area where indications show that an economically interesting orebody may occur. On this account, the company is about to start the mineral exploration activities from the preliminary phase up to a pilot mine excavation and trial exploitation of the mineralization



Bayes's Theorem, Fig. 1 In the Mineral Exploration context, Bayes's theorem is used to enhance our knowledge on the mineralization by gradually assimilating new information coming from indirect or inexact sources. The screenshot is from "Mineral Adventure," a simulation game developed by National Technical University of Athens in the framework

of EU-funded "Virtual Mine" project (<https://virtualmine.net/>). You can practice strategic decision-making with Bayes's theorem in Mineral Exploration by playing this game in <http://geostatistics.eu/exercise-en.php#>

(Fig. 1). Moreover, since this is a step-by-step procedure, there is always the option to terminate the development of the project in the case it is no longer cost-effective, based on evidence.

An overview of the recorded experience on similar setups and a consultant geologist's advice says that the à priori probability of finding an economic mineralization may be up to 55%. Because this probability is not satisfactory and in order to increase the level of confidence on the presence of the deposit, the CEO considers the option to conduct first a geophysical survey of the subsurface, which, as also shown by experience, has a reliability of 80%. After reasonable thinking, he chooses to proceed on the exploration. This judgment pays back with a positive result.

At this point, the company has to decide whether to proceed to the next step of the exploration, which is to start a drilling campaign in order to collect real samples from the subsurface. But this decision is difficult because drilling might be very expensive. The CEO has to convince the board of directors that the new evidence is sufficient to support this decision (or is not?). How will he estimate the updated probability of the existence of the orebody? How can the new information be incorporated into the already accumulated knowledge? Bayes's theorem can solve this problem.

Consider translating event $\mathcal{A}$ from Eq. 1 as the existence of an orebody in the area under exploration and event $\mathcal{B}$ as the result of the geophysical campaign (positive or negative for

simplicity). Then, $P(\mathcal{A})$ is the à priori or prior probability of existence (which means before the present study), and $P(\mathcal{A}|\mathcal{B})$ is the updated (with the geophysical measurements) or the posterior probability of existence. Furthermore, $P(\mathcal{B}|\mathcal{A})$ can be conveniently translated as the likelihood of the data (geophysical measurements) and, finally, $P(\mathcal{B})$ as the total probability that geophysical exploration will give a positive result, that is, a constant number that ensures that the output will always be between 0 and 1. In this line, Bayes's theorem can be written as

$$\text{Posterior probability} = \frac{\text{Prior probability} \times \text{Likelihood of the data}}{\text{Normalization constant}} \quad (2)$$

According to the above, the normalization constant $P(\mathcal{B})$ has to be estimated first. By the total probability theorem (Papoulis and Pillai 2002),

$$P(\mathcal{B}) = P(\mathcal{B}|\mathcal{A})P(\mathcal{A}) + P(\mathcal{B}|\mathcal{A}')P(\mathcal{A}') \quad (3)$$

where $\mathcal{A}'$ is the complementary event of $\mathcal{A}$, i.e., a deposit does not exist.

After substitution in Eq. 3, we get

$$P(\mathcal{B}) = 0.8 \times 0.55 + 0.2 \times 0.45 = 0.53 \quad (4)$$

Inserting this probability into Bayes's theorem, we finally obtain the updated probability of existence:

$$P(\mathcal{A}|\mathcal{B}) = \frac{0.8 \times 0.55}{0.53} = 0.83 \quad (5)$$

What unveils here is that Bayes's theorem is a valuable tool to assimilate the newly gained information into the knowledge base. In this example, since the geophysical method used is reliable, the belief that a deposit may exist is substantially improved, and now it is reasonable to proceed to the drilling campaign.

Application to Geological Systems Modeling or "Like Oil and Water"

Modeling of geological systems like groundwater tables or petroleum reservoirs is typical in geosciences (Pyrz and Deutsch 2014). The scope of modeling (or simulation or forward problem) is to predict the response of the system to certain external stimuli. For example, in a petroleum reservoir, we would like to know the flow characteristics of oil in a certain production well. The usual difficulty in modeling a geological system is the inadequate knowledge of the parameters of the physical law representing its response to the stimulation. This is because these parameters are in general some regionalized quantities referring to soil or rock mass properties, and thus they may present large variability from site to site. In the inverse problem, the results of a set of measurements are used in order to infer the actual values of the parameters characterizing the geological system. But, although the forward problem normally has a unique solution, the inverse problem has not. This characteristic is referred to as "ill-posedness" and is the main reason why stochastic methods are preferable in order to solve the inverse problem (Tarantola 2005). Instead of focusing on exact numbers, by representing the spatial parameter field with a random function, the idea is to estimate intervals containing the real values of the parameters, given some uncertain measurements of system response.

Summarizing the above in a mathematical notation, the model of a physical system g with boundary conditions $\mathbf{u}$ can be described by its parameter set $\mathbf{m} = \{m_1, \dots, m_n\}$ and their spatial coordinates $\mathbf{x}$ over a period t . Then, the forward solution of the system can be written as

$$\mathbf{d} = g(\mathbf{x}, t, \mathbf{u}; \mathbf{m}) \quad (6)$$

where $\mathbf{d} = \{d_1, \dots, d_p\}$ is the response of the system.

Since interest is now focused on the parameters, Eq. 6 can be written for simplicity as

$$\mathbf{d} = g(\mathbf{m}) \quad (7)$$

As previously stated, the objective of the inverse problem

in geosciences is to estimate the spatiotemporal distribution of the parameters $\mathbf{m}$ by using all the available information such as physical laws, measurements and moments concerning the parameters, and also measurements of the response variables (e.g., head measurements). Under the probabilistic formalism, the response variable $\mathbf{d}$ and the parameter set of the system are considered as random variables. The vector of observations $\mathbf{d}_{\text{obs}}$ is considered as the response from a particular parameter set, which has to be determined. In order to define a statistical distribution for $\mathbf{d} | \mathbf{m}$, a known error distribution of the observations has to be assumed (likelihood of data). Also, it is important to consider a known distribution $f(\mathbf{m})$ of the parameters (prior knowledge). Then, according to Bayes's theorem, the a posteriori distribution of the parameters is defined as

$$f(\mathbf{m}|\mathbf{d}) = \frac{f(\mathbf{d}|\mathbf{m})f(\mathbf{m})}{f(\mathbf{d})} \quad (8)$$

where $f(\mathbf{d}|\mathbf{m})$ is the likelihood of data and $f(\mathbf{m})$ is the prior distribution of the parameter set (see derivation below). Repeating the calculation of Eq. 8 for different parameter sets derived from the prior distribution and producing different $\mathbf{d}_{\text{obs}}$ sets, we can obtain the a posteriori distribution of $\mathbf{m}$. Then, the best estimator of $\mathbf{m}$ is the one that maximizes the a posteriori probability $f(\mathbf{m} | \mathbf{d})$. This methodology is known as the maximum a posteriori probability (MAP) (Oliver et al. 2008).

Derivation

In its simple form for events as in Eq. 1, Bayes's theorem is derived from the definition of conditional probability:

$$P(\mathcal{A}|\mathcal{B}) = \frac{P(\mathcal{A} \cap \mathcal{B})}{P(\mathcal{B})}, \text{ with } P(\mathcal{B}) \neq 0 \text{ and}$$

$$P(\mathcal{B}|\mathcal{A}) = \frac{P(\mathcal{B} \cap \mathcal{A})}{P(\mathcal{A})}, \text{ with } P(\mathcal{A}) \neq 0$$

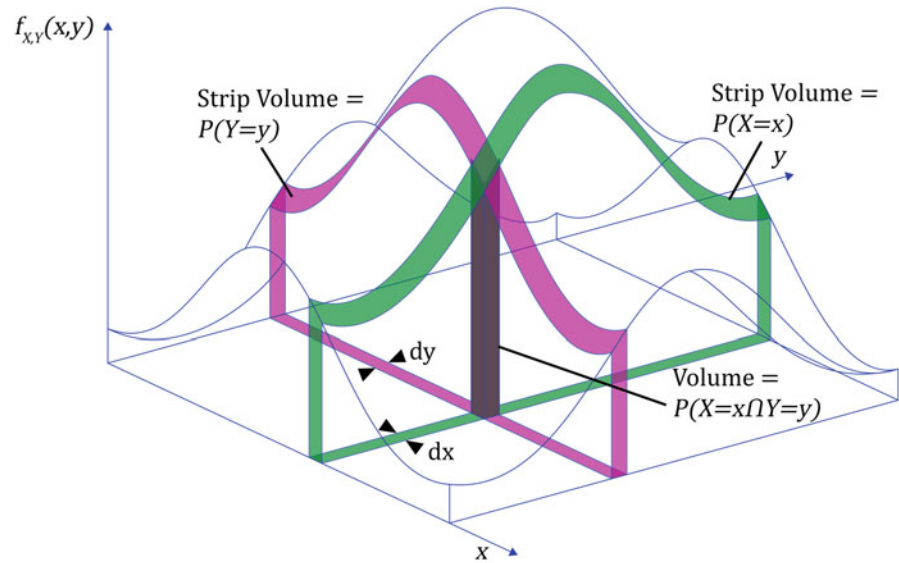
where $P(\mathcal{A} \cap \mathcal{B})$ is the joint probability of events $\mathcal{A}$ and $\mathcal{B}$.

$$\text{Since } P(\mathcal{A} \cap \mathcal{B}) = P(\mathcal{B} \cap \mathcal{A}) \Rightarrow P(\mathcal{A} \cap \mathcal{B}) = P(\mathcal{A}|\mathcal{B})P(\mathcal{B}) = P(\mathcal{B}|\mathcal{A})P(\mathcal{A})$$

$$\Rightarrow P(\mathcal{A}|\mathcal{B}) = \frac{P(\mathcal{B}|\mathcal{A})P(\mathcal{A})}{P(\mathcal{B})}, \text{ if } P(\mathcal{B}) \neq 0$$

In the continuous case of random variables as used in Eq. 8, the theorem can be derived in an analogous manner from the definition of conditional probability (Fig. 2):

Bayes's Theorem, Fig. 2 On the proof of continuous case of Bayes's theorem: Conditional probabilities applied to the event space of random variables X and Y . By equating these probabilities in the infinitesimal prism formed by the intersection of the two stripes, an instance of Bayes's theorem is revealed for each point in the domain. Integrating these instances for x and y , Eq. 9 is derived



$$P(Y=y|X=x) = \frac{P(X=x \cap Y=y)}{P(X=x)}$$

$$P(X=x|Y=y) = \frac{P(X=x \cap Y=y)}{P(Y=y)}$$

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \text{ and}$$

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

Hence,

$$f_{X|Y=y}(x) = \frac{f_{Y|X=x}(y)f_X(x)}{f_Y(y)} \quad (9)$$

Bayesian Statistics

Stochastic methods in geosciences do not adopt in general the classic statistical but rather the Bayesian view, where probabilities represent states of knowledge or accessible information. The main point in this view is that the variable of interest over an area is modeled as a random function because there is inadequate information to enable the use a deterministic model, rather than due to the “law of large numbers,” that is, because repeated observations indicate a statistical regularity (Kitanidis 2007).

Thus, in Bayesian statistics, probability is used as a measure of uncertainty, or “belief,” or available information about an event. In the context of spatial estimation, which is of main concern in geosciences, the essence is to comply with two requirements: Include all previous knowledge about spatial variability in a highly informative prior distribution and then use the observations to infer a maximized posterior

distribution through Bayes's relation (Christakos 1990; Valakas and Modis 2016). Classical Kriging estimators can also derive under this generalized perspective: Since prior information is limited to the mean and covariance functions and only exact measurements are considered, maximization of entropy leads to Gaussian prior and posterior distributions. Then, the minimum mean squared error estimator equals the conditional mean according to Bayes's theorem (Christakos 2000).

Conclusions

Bayes's theorem relates events or random variables by avoiding using their joint probability in favor of conditional and marginal ones. In many cases, these probabilities are easier to define.

In mineral exploration, it provides the tools to update our knowledge on the subsurface with data from different sources of fuzzy information.

In geological modeling, Bayes's theorem is used to transform the prior distribution of the system's parameters into a narrower posterior distribution by using a set of uncertain measurements of the response variables.

Cross-References

- Bayesian Inversion in Geoscience
- Bayesian Maximum Entropy

- [Ensemble Kalman Filtering](#)
- [Exploratory Data Analysis](#)
- [Inversion Theory in Geoscience](#)
- [Random Variable](#)

Bibliography

- Christakos G (1990) A Bayesian/maximum-entropy view to the spatial estimation problem. *Math Geol* 22:763–777
- Christakos G (2000) *Modern spatiotemporal geostatistics*. Oxford University Press, New York. 288 p
- Kitanidis P (2007) On stochastic inverse modeling. *Geophys Monogr Ser* 171:19–30
- Oliver D, Reynolds A, Liu N (2008) *Inverse theory for petroleum reservoir characterization and history matching*. Cambridge University Press, New York. 380 p
- Papoulis A, Pillai U (2002) *Probability, random variables and stochastic processes*. McGraw Hill, New York. 852 p
- Pyrz M, Deutsch C (2014) *Geostatistical Reservoir Modeling*. Oxford University Press, New York. 433 p
- Tarantola A (2005) *Inverse problem theory and methods for model parameter estimation*. Siam, Philadelphia. 342 p
- Valakas G, Modis K (2016) Using informative priors in facies inversion: the case of C-ISR method. *Adv Water Resour* 94:23–30

Bayesian Inversion in Geoscience

Dario Grana¹, Klaus Mosegaard² and Henning Omre³

¹Department of Geology and Geophysics, University of Wyoming, Laramie, WY, USA

²Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark

³Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

Definition

Bayesian inversion is a statistical method used in many geoscience applications to assess the probability distribution of the unknown variables conditioned on the available measurements, where the posterior model is uniquely defined by a prior model representing the knowledge of the variable before the data acquisition and a likelihood model linking the target variables and the measured data.

Introduction

Geophysics theories, such as seismology, electromagnetism, and poro-elasticity, include mathematical relations of the physical processes and properties of rock and fluid properties in the subsurface. If the rock and fluid physical properties are

known, then it is possible to apply a forward operator to predict the geophysical response of the subsurface. For example, if the volumetric fractions of the mineral and fluid components of porous rocks are known, then their seismic response can be predicting by solving a set of poro-elastic equations. In practice, the petrophysical properties of the porous rocks are unknown, whereas their geophysical response can be measured using geophysical methods, such as seismic acquisition. Therefore, the prediction of the rock and fluid properties from geophysical measurements can be defined as an inverse problem. Bayesian methods are commonly used in geophysical inverse problems (Tarantola 2005).

Geophysical inverse problems are often ill-posed and the solution is generally nonunique. Bayesian inverse theory provides a mathematical framework for these problems since it aims at calculating the posterior probability distribution of the variables of interest conditioned on the measured data by combining the prior distribution of the model variables with the likelihood function of the data. The prior model represent the knowledge of the variables of interest prior to the data acquisition, whereas the likelihood model links the variables of interest and the measured data. The posterior model is then uniquely identified by these two models (Tarantola 2005). The posterior model can be computed analytically for certain classes of likelihood and prior models, but in the general case, the posterior distribution must be assessed numerically, by iteratively approximating the solution. Bayesian methods have been proposed for a variety of geophysical inverse problems (Mosegaard and Tarantola 1995; Scales and Tenorio 2001; Ulrych et al. 2001; Buland and Omre 2003; Tarantola 2005; Rimstad et al. 2012).

The focus of this entry is on seismic inverse problems for the prediction of subsurface properties from seismic measurements (Aki and Richards 1980). The geophysics literature includes a variety of formulations of the seismic inversion problem, which differ for the formulation and parameterization of the physical operator. Indeed, the seismic inverse problem can be formulated in terms of elastic properties (seismic velocity and density or seismic impedance) or petrophysical properties (porosity, mineral volumes, and fluid saturations), and the physics can include the full wave propagation model or linearized approximations. For linear formulations and under Gaussian assumptions, analytical solutions of the Bayesian inverse problem are derived (Buland and Omre 2003; Grana et al. 2017), whereas for nonlinear formulations, Monte Carlo methods must be adopted to approximate the posterior distribution (Mosegaard and Tarantola 1995; Sambridge and Mosegaard 2002).

In the following, we review the main principles of Bayesian inverse theory, present analytical solutions for linear inverse problems assuming Gaussian and Gaussian mixture

distributions, and present an overview of Markov chain Monte Carlo methods. These methods are illustrated on a one-dimensional synthetic seismic dataset for the prediction of the elastic properties of porous rocks in the subsurface.

Methods

In geophysical inverse problems, the variable of interest is a physical property in the subsurface, defined on a spatial grid, and it is represented here by the vector $\mathbf{m}$, whereas the geophysical data, such as seismic amplitudes, are represented here by the vector $\mathbf{d}$. The aim of the inverse problem is to assess the variables $\mathbf{m}$ given the measured data $\mathbf{d}$ associated with the forward model

$$\mathbf{d} = \mathbf{f}(\mathbf{m}) + \mathbf{e}, \quad (1)$$

where $\mathbf{f}(\cdot)$ is the geophysical forward operator and $\mathbf{e}$ represents the measurement errors. For example, the data can be seismic measurements and the model variables can be elastic and petrophysical properties of porous rocks. The forward operator $\mathbf{f}(\cdot)$ can be represented by linear or nonlinear physical relations.

In a deterministic setting, the least square solution $\hat{\mathbf{m}}$ can be obtained by minimizing the L2 norm of the residuals $\mathbf{r} = \mathbf{d} - \mathbf{f}(\mathbf{m})$, i.e., difference between measured and predicted data:

$$\hat{\mathbf{m}} = \arg \min_{\mathbf{m}} \|\mathbf{d} - \mathbf{f}(\mathbf{m})\|_2. \quad (2)$$

For ill-conditioned and rank-deficient systems, singular value decomposition or regularization methods, such as Tikhonov regularization, are generally applied.

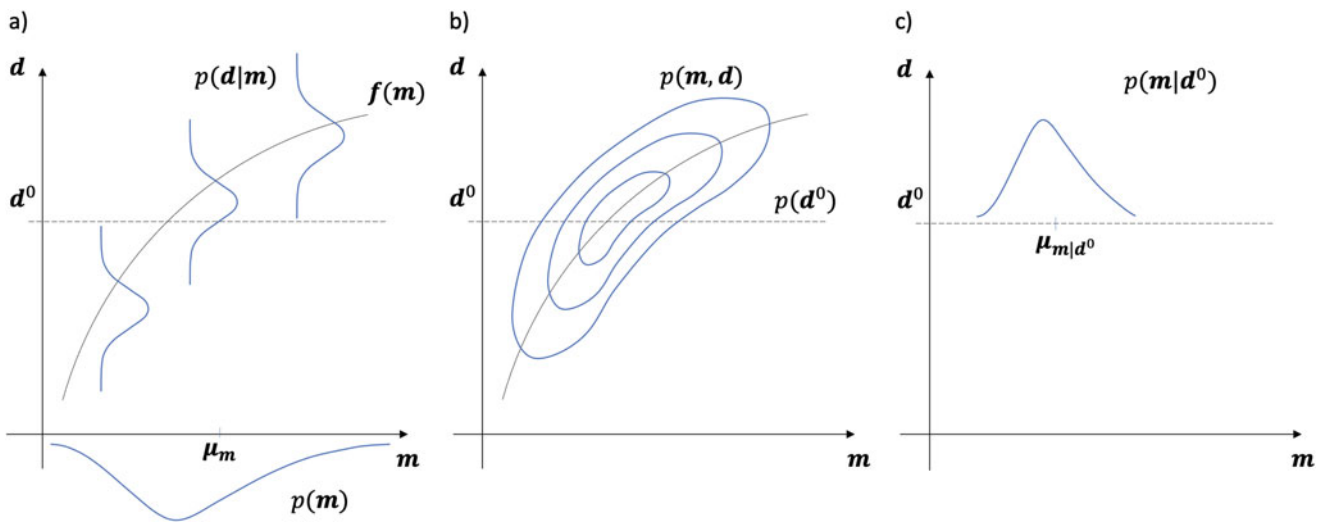
In a probabilistic setting, the solution of the inverse problem is expressed as the posterior probability distribution $p(\mathbf{m}|\mathbf{d})$ of the variables given the measured data, which can be assessed using Bayes' theorem

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})} = \text{const} \times p(\mathbf{d}|\mathbf{m})p(\mathbf{m}) \quad (3)$$

where $p(\mathbf{m})$ is the prior distribution of the variables, $p(\mathbf{d}|\mathbf{m})$ is the data likelihood function, and $p(\mathbf{d})$ is a normalizing constant to guarantee that the posterior distribution $p(\mathbf{m}|\mathbf{d})$ is a valid probability distribution.

The Bayesian inversion concept is graphically displayed in Fig. 1. The goal is to assess the variable of interest $\mathbf{m}$ given a value $\mathbf{d}^0$ of the measured data $\mathbf{d}$. Figure 1a displays the likelihood model $p(\mathbf{d}|\mathbf{m})$ and the prior model $p(\mathbf{m})$. The Bayesian inversion operates in the $[\mathbf{m} \times \mathbf{d}]$ -space. The forward operator $\mathbf{f}(\mathbf{m})$ and the associated random error $\mathbf{e}$ define the likelihood model $p(\mathbf{d}|\mathbf{m})$, whereas the prior model $p(\mathbf{m})$ with prior expectation μ_m is specified based on the prior knowledge of the variable $\mathbf{m}$. Figure 1b displays the joint distribution $p(\mathbf{m}, \mathbf{d}) = p(\mathbf{d}|\mathbf{m})p(\mathbf{m})$ as a contour map. The joint model fully defines the interaction between $\mathbf{m}$ and $\mathbf{d}$. The marginal distribution $p(\mathbf{d})$ is obtained by integrating $p(\mathbf{m}, \mathbf{d})$ over the values of the variable $\mathbf{m}$. The normalizing constant $p(\mathbf{d}^0)$ is obtained by evaluating the marginal distribution in the data value $\mathbf{d}^0$.

Figure 1c displays the posterior distribution $p(\mathbf{m}|\mathbf{d}^0) = p(\mathbf{m}, \mathbf{d}^0)/p(\mathbf{d}^0)$ in Eq. 3 and its posterior expectation $\mu_{m|\mathbf{d}^0}$.



Bayesian Inversion in Geoscience, Fig. 1 Graphical representation of Bayesian inversion: (a) prior and likelihood models, (b) joint model, and (c) posterior model

$\mu_{m|d^0}$ for the data value d^0 . In the general case, the prior and posterior models might not be Gaussian. For highly skewed or multimodal distributions, the variable values associated with the maximum of the prior and posterior models are generally more informative than the prior and posterior expectations.

Bayesian Linear Inversion

In the case of linear operators, the forward model in Eq. 1 is rewritten as a linear system of equations

$$d = \mathbf{F}m + e, \quad (4)$$

where $\mathbf{F}$ is the matrix defined by the linear operator $f(\cdot)$. If the random error e is assumed to be Gaussian with $\mathbf{0}$ -mean and covariance matrix equal to Σ_e , then the likelihood model $p(d|m)$ is Gaussian

$$p(d|m) = \mathcal{N}(d; \mathbf{F}m, \Sigma_e) \quad (5)$$

with expectation $\mathbf{F}m$ linear in m and covariance matrix Σ_e .

If the prior model of the variables $p(m)$ is Gaussian, then the posterior model of the variables given the measured data $p(m|d)$ is also Gaussian. Hence, the prior and posterior models are conjugate distributions with respect to the likelihood model $p(d|m)$. For a Gaussian prior model $p(m) = \mathcal{N}(m; \mu_m, \Sigma_m)$ with expectation μ_m and covariance matrix Σ_m , the posterior distribution of the variable $[m|d]$ is also Gaussian

$$p(m|d) = \mathcal{N}(m; \mu_{m|d}, \Sigma_{m|d}) \quad (6)$$

with conditional expectation $\mu_{m|d}$ and conditional covariance matrix $\Sigma_{m|d}$ given by

$$\mu_{m|d} = \mu_m + \Sigma_m \mathbf{F}^T [\mathbf{F} \Sigma_m \mathbf{F}^T + \Sigma_e]^{-1} (d - \mathbf{F} \mu_m), \quad (7)$$

and

$$\Sigma_{m|d} = \Sigma_m - \Sigma_m \mathbf{F}^T [\mathbf{F} \Sigma_m \mathbf{F}^T + \Sigma_e]^{-1} \mathbf{F} \Sigma_m, \quad (8)$$

respectively. The posterior variances of the marginal distributions are independent of the measurement values and only depends on the parameters of prior model and measurement errors. The derivation can be found in Tarantola (2005).

In Buland and Omre (2003), the Bayesian linear formulation is applied to the seismic inversion problem. This approach is referred to as Bayesian linearized amplitude-versus-offset (AVO) inversion. The linear forward operator is obtained by assuming a convolution of a known wavelet and the linearized AVO approximation (Aki and Richards 1980) of the reflectivity coefficients that are function of the time-derivative of the logarithm of the elastic properties. The

convolution is a linear operator, hence the forward operator can be expressed by a matrix multiplication, $\mathbf{F} = \mathbf{WAD}$, where $\mathbf{W}$ is the wavelet matrix, $\mathbf{A}$ is the reflectivity coefficient matrix, and $\mathbf{D}$ is a first-order differential matrix. The forward operator is then linear with respect to the logarithm of the elastic properties $\ln(m)$ (where m is here a dimensionless representation of the parameters) and the linear expression in Eq. 4 can be written as $d = \mathbf{WAD} \ln(m) + e$. In Buland and Omre (2003), a Gaussian distribution for the logarithm of the elastic properties $\ln(m)$ is assumed, which entails a log-Gaussian prior model of the variable m . Based on these assumptions, the seismic inversion problem is analytically tractable and the posterior model of $[\ln(m)|d]$ is Gaussian, which entails a log-Gaussian posterior model of the variable $[m|d]$ (Buland and Omre 2003).

Gaussian mixture distributions are also conjugate models with respect to the likelihood function $p(d|m)$. If the prior model of the variables is a Gaussian mixture distribution $p(m) = \sum_k p(k) \mathcal{N}(m; \mu_{m|k}, \Sigma_{m|k})$, then the posterior model of the variables given the measured data is also a Gaussian mixture distribution

$$p(m|d) = \sum_k p(k|d) \mathcal{N}(m; \mu_{m|d,k}, \Sigma_{m|d,k}) \quad (9)$$

with conditional mean $\mu_{m|d,k}$ and conditional covariance matrix $\Sigma_{m|d,k}$, given by

$$\mu_{m|d,k} = \mu_{m|k} + \Sigma_{m|k} \mathbf{F}^T [\mathbf{F} \Sigma_{m|k} \mathbf{F}^T + \Sigma_e]^{-1} (d - \mathbf{F} \mu_{m|k}), \quad (10)$$

and

$$\Sigma_{m|d,k} = \Sigma_{m|k} - \Sigma_{m|k} \mathbf{F}^T [\mathbf{F} \Sigma_{m|k} \mathbf{F}^T + \Sigma_e]^{-1} \mathbf{F} \Sigma_{m|k}, \quad (11)$$

respectively (Grana et al. 2017). Due to the spatial coupling of the likelihood model in many geophysical applications, the posterior probability $p(k|d)$

$$p(k|d) = \frac{p(d|k)p(k)}{p(d)} = \text{const} \times p(d|k)p(k) \quad (12)$$

cannot be analytically computed. The numerical calculation of $p(k|d)$ is computationally demanding, hence it must be numerically approximated (Grana et al. 2017).

In the framework of Bayesian inverse theory, Monte Carlo methods can be used to generate realizations of the posterior distribution of Gaussian linear inverse problems, honoring a covariance function describing the spatial continuity of the variables.

Monte Carlo Methods

For general nonlinear problems in Eq. 1, analytical solutions are generally not available and numerical techniques such as Monte Carlo and Markov chain Monte Carlo (MCMC) algorithms must be adopted to approximate the posterior distribution (Mosegaard and Tarantola 1995; Sambridge and Mosegaard 2002). In Monte Carlo acceptance/rejection sampling approaches, the posterior distribution is approximated by sequentially drawing proposals from a prior distribution of the variable and accepting or rejecting the proposal based on its likelihood. In MCMC methods, the posterior distribution is approximated by sequentially drawing proposals from a proposal distribution conditioned on the previous realization of the variable.

Markov chain Monte Carlo methods design a random walk in the variable space using a Markov chain, i.e., a random process in which the probability of moving from the previous state to the current state is defined by a transition probability that depends only on the previous state. In applications to inverse problems, this process is used to draw realizations from the posterior distribution. The sequence of samples, i.e., the Markov chain, asymptotically reaches a stationary distribution that is equal to the posterior distribution of the Bayesian inverse problem. There are different Markov chain Monte

Carlo algorithms, including Metropolis, Metropolis-Hastings, and Gibbs sampling.

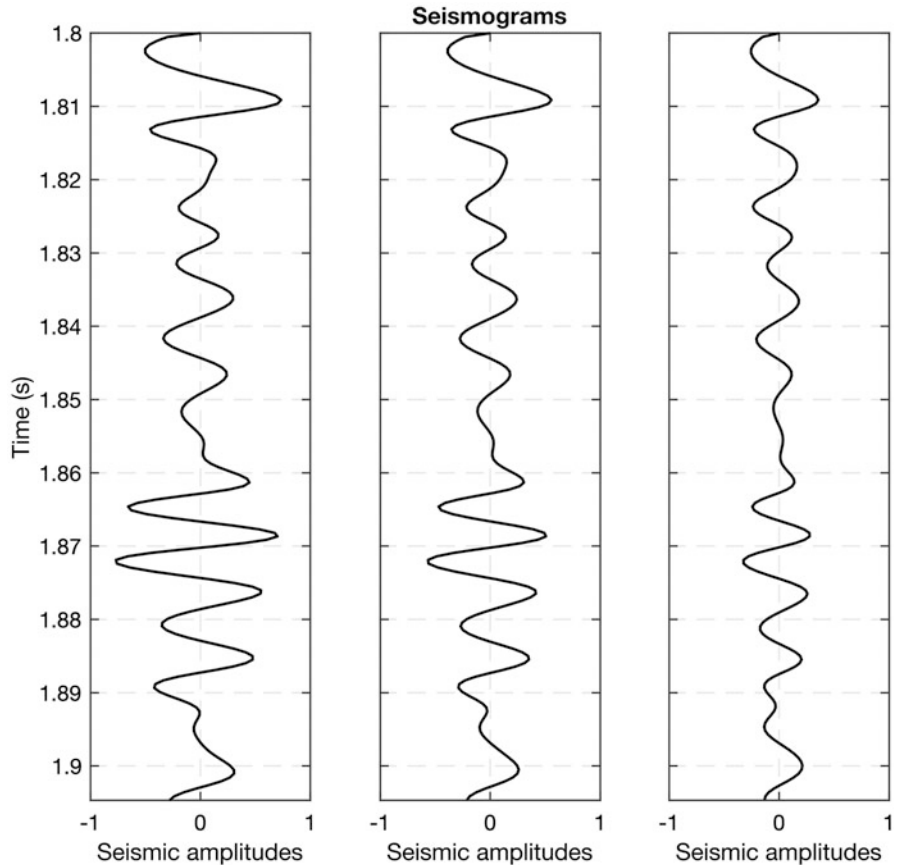
The most general MCMC technique is the Metropolis-Hastings algorithm. According to a Metropolis-Hastings approach frequently used in geoscience, an initial realization $\mathbf{m}_0$ is generated from the prior distribution $p(\mathbf{m})$ and represents the initial state of the chain. Then, at each iteration i , the proposed realization $\mathbf{m}_p$ is generated by sampling from a proposal distribution $g(\mathbf{m}|\mathbf{m}_{i-1})$ conditioned on the state $\mathbf{m}_{i-1}$ at the previous iteration. The proposed realization is accepted with probability

$$p^* = \min \left(\frac{p(\mathbf{d}|\mathbf{m}_p)p(\mathbf{m}_p)g(\mathbf{m}_{i-1}|\mathbf{m}_p)}{p(\mathbf{d}|\mathbf{m}_{i-1})p(\mathbf{m}_{i-1})g(\mathbf{m}_p|\mathbf{m}_{i-1})}, 1 \right) \quad (13)$$

If accepted, the proposed realization is assigned to the current state $\mathbf{m}_i = \mathbf{m}_p$, otherwise $\mathbf{m}_i = \mathbf{m}_{i-1}$. The MCMC algorithm converges in the sense that the generated states represent realizations from the posterior model, as iterations increase. The Gibbs sampling is a special case of the Metropolis-Hastings algorithm, and it generates proposals from the full conditional distribution of subdimensions of $\mathbf{m}_i$, hence $p^* = 1$ in Eq. 14, such that the proposed state is

Bayesian Inversion in

Geoscience, Fig. 2 Synthetic seismograms representing the seismic measurements: from left to right, seismograms for incident angles of 12°, 24°, and 36°



always accepted. However, in many applications, this advantage of the Gibbs sampler is offset by the need to compute full conditional distributions at each iteration.

In the Metropolis approach, the proposal distribution $g(\mathbf{m}|\mathbf{m}_{i-1})$ is assumed to be symmetric, such that $g(\mathbf{m}|\mathbf{m}_{i-1}) = g(\mathbf{m}_{i-1}|\mathbf{m})$ and the acceptance probability becomes

$$p^* = \min\left(\frac{p(\mathbf{d}|\mathbf{m}_i)p(\mathbf{m}_i)}{p(\mathbf{d}|\mathbf{m}_{i-1})p(\mathbf{m}_{i-1})}, 1\right) \quad (14)$$

Monte Carlo and Markov chain Monte Carlo approaches are adopted for seismic and petrophysical inversion problems and combined with analytical approximations to sample from the posterior distributions. The methodology is used for litho-fluid classification (Grana et al. 2017) and is also extended to seismic classification problems including categorical variables (de Figueiredo et al. 2019).

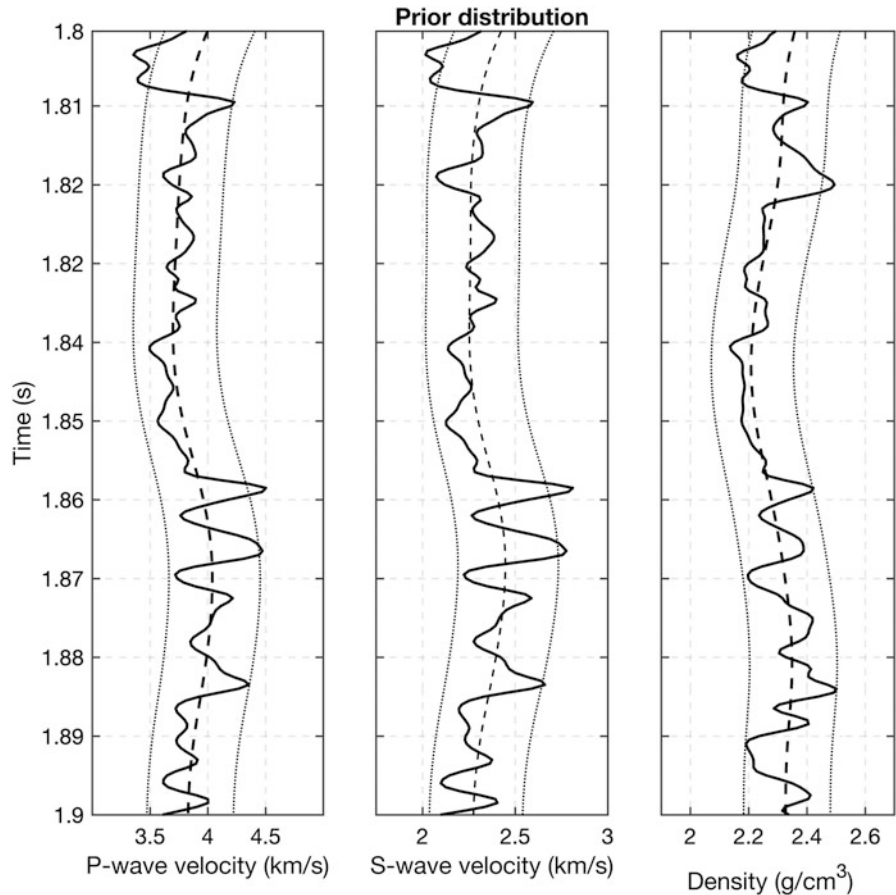
Example

We illustrate the application of Bayesian inverse methods to a seismic inversion problem to predict the elastic properties

along a one-dimensional profile from a set of seismograms measured with different acquisition geometries. The subsurface variables of interest are: P-wave velocity, S-wave velocity, and density. The measured data include three synthetic seismograms computed from a set of borehole data for incident angles of 12° , 24° , and 36° (Fig. 2). The forward operator is a convolution of a known wavelet and the reflectivity coefficients computed from the elastic properties, according to Aki-Richards approximation for the linearized operator and according to Zoeppritz equations for the nonlinear operator (Aki and Richards 1980).

First, we apply the Bayesian linearized AVO inversion proposed in Buland and Omre (2003) to the seismic dataset in Fig. 3 by using the Aki Richards approximation. The prior distribution model is a trivariate Gaussian distribution of the logarithms of the elastic properties. The prior mean is represented by a low frequency trend of the elastic properties computed by filtering the well log data at a cutoff frequency equal to the lower bound of the seismic bandwidth. The prior covariance matrix is estimated from the difference between the actual well measurements and the prior mean. The prior model includes an exponential spatial correlation function with correlation length of 20 ms. The marginal distributions

Bayesian Inversion in Geoscience, Fig. 3 Prior distribution of the model variables: from left to right, P-wave velocity, S-wave velocity, and density. Solid black lines represent the true model, dashed black lines represent the mean, and dotted black lines represent the 0.90 confidence interval



of the prior model are shown in Fig. 3. The measurement error is assumed to be spatially uncorrelated with 0-mean and variance equal to 20% of the variance of the data, corresponding to a signal-to-noise ratio of 5. The posterior distribution of the elastic variables is then the Gaussian distribution of the logarithms of the elastic properties defined in Eq. 6 with parameters given by the analytical expressions in Eqs. 7 and 8. The marginal components of the posterior distribution of the elastic variables are shown in Fig. 4. The maximum posterior predictions of the marginal distribution are in good agreement with the reference data.

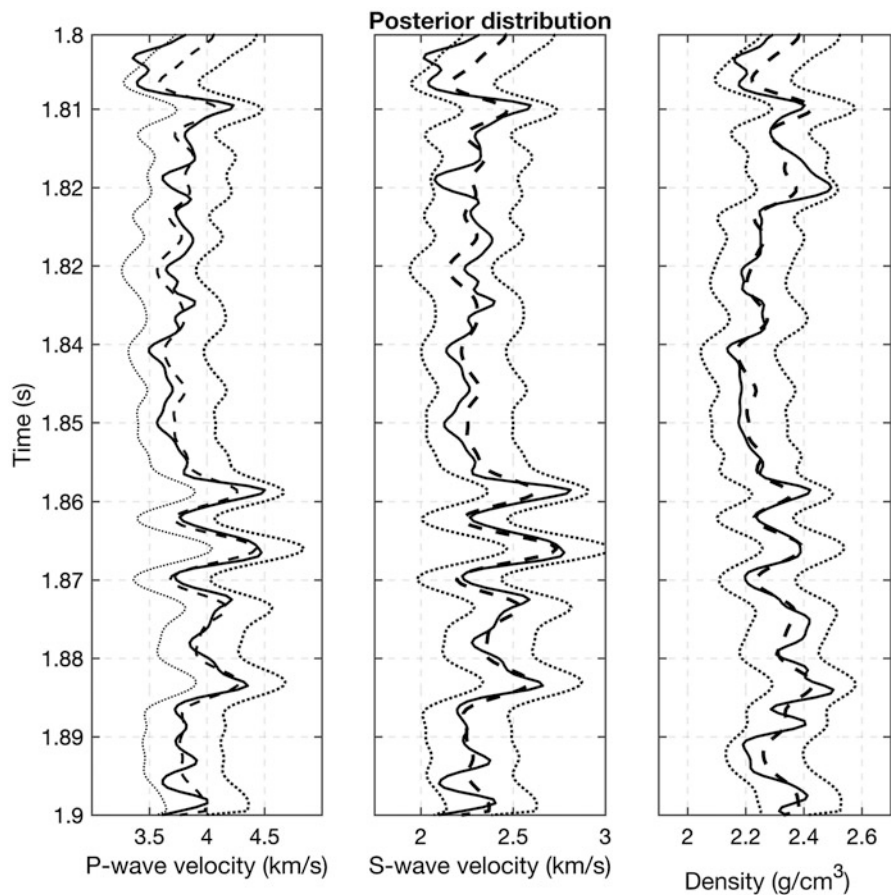
We also apply a Markov chain Monte Carlo method based on the Metropolis-Hastings algorithm to the same seismic dataset using the convolutional operator and the nonlinear Zoeppritz equations. The prior realization of the model variables is generated by sampling from the trivariate prior distribution with exponential spatial correlation function used in the previous example. The proposal distribution is assumed to be Gaussian. In this application, 4000 proposals are drawn from the proposal distribution with an acceptance rate of approximately 24%. Figure 5 shows 500 realizations randomly selected after convergence. The algorithm is repeated

10 times, starting from different states, to avoid that results might be biased by the initial realization of the chain.

Conclusions

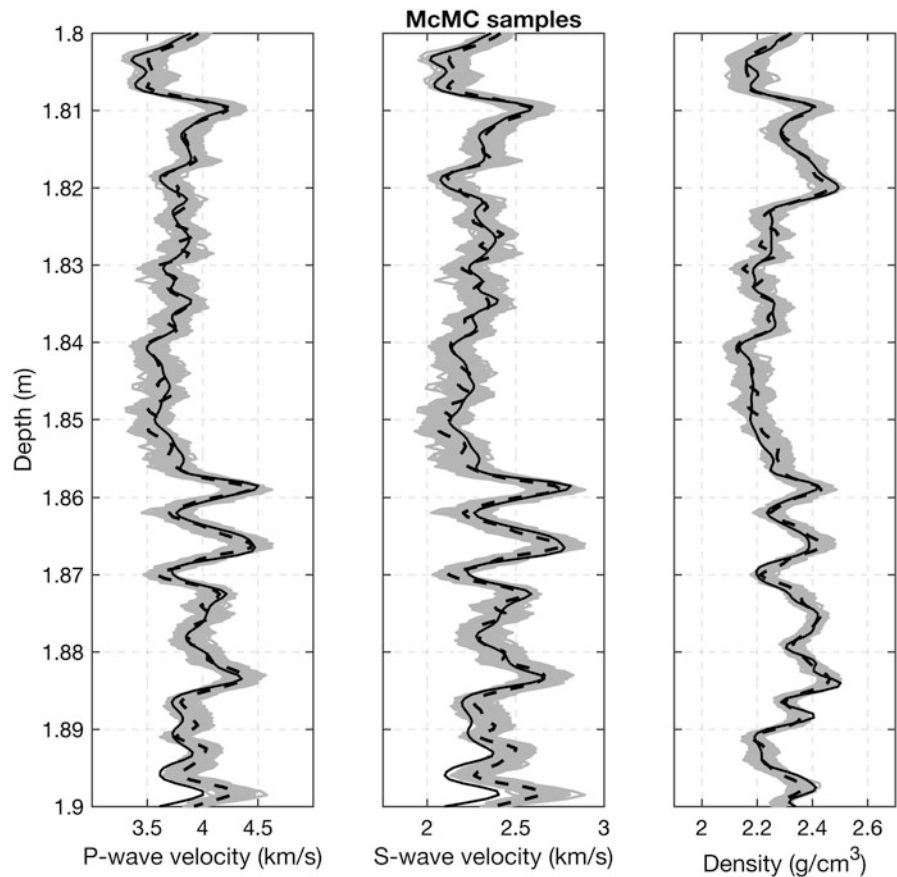
Bayesian inversion methodologies are commonly adopted in geophysical inverse problems for the assessment of the posterior distribution of the subsurface variables conditioned on the available data. The Bayesian approach provides a natural framework to predict the most likely solution and to quantify the associated uncertainty. For linear inverse problems, with Gaussian prior distribution and linear likelihood operators with additive Gaussian errors, the posterior distribution is also Gaussian and the posterior mean and covariance matrix can be analytically assessed. The analytical tractability makes the approach very efficient and easily extendable to three-dimensional studies. For nonlinear problems, the posterior distribution is iteratively approximated using stochastic simulation algorithms such as the Markov chain Monte Carlo method. The need for iterative algorithms makes the approach computationally demanding, especially for large datasets. A demonstration of the two approaches is presented on a

Bayesian Inversion in Geoscience, Fig. 4 Posterior distribution of the model variables: from left to right, P-wave velocity, S-wave velocity, and density. Solid black lines represent the true model, dashed black lines represent the mean, and dotted black lines represent the 0.90 confidence interval



Bayesian Inversion in

Geoscience, Fig. 5 Markov chain Monte Carlo posterior samples of the model variables: from left to right, P-wave velocity, S-wave velocity, and density. Solid black lines represent the true model, dashed black lines represent the mean, and grey lines represent the posterior samples



seismic inversion problem assessing the posterior distribution of elastic properties of subsurface porous rocks. The Bayesian approach can be extended to other model parameterizations and formulations, to include rock physics models for the prediction of petrophysical properties. Furthermore, categorical variables associated with geological concepts can be included for the classification of litho-fluid classes. Such formulations generally require more complex prior models, combining Gaussian mixture distributions with Markov random fields.

Bibliography

- Aki K, Richards PG (1980) Quantitative seismology: theory and methods. W.H. Freeman, New York
- Buland A, Omre H (2003) Bayesian linearized AVO inversion. *Geophysics* 68(1):185–198
- de Figueiredo LP, Grana D, Roisenberg M, Rodrigues BB (2019) Multi-modal Markov chain Monte Carlo method for nonlinear petrophysical seismic inversion. *Geophysics* 84(5):M1–M13
- Grana D, Fjeldstad T, Omre H (2017) Bayesian Gaussian mixture linear inversion for geophysical inverse problems. *Math Geosci* 49(4): 493–515
- Mosegaard K, Tarantola A (1995) Monte Carlo sampling of solutions to inverse problems. *J Geophys Res* 100:12,431–12,447

- Rimstad K, Avseth P, Omre H (2012) Hierarchical Bayesian lithology/fluid prediction: a North Sea case study. *Geophysics* 77(2):B69–B85
- Sambridge M, Mosegaard K (2002) Monte Carlo methods in geophysical inverse problems. *Rev Geophys* 40(3):3–1
- Scales JA, Tenorio L (2001) Prior information and uncertainty in inverse problems. *Geophysics* 66:389–397
- Tarantola A (2005) Inverse problem theory. SIAM, Oxford
- Ullrich TJ, Sacchi MD, Woodbury A (2001) A Bayes tour of inversion: a tutorial. *Geophysics* 66:55–69

Bayesian Maximum Entropy

Junyu He^{1,2} and George Christakos³

¹Ocean Academy, Zhejiang University, Zhoushan, China

²Ocean College, Zhejiang University, Zhoushan, China

³Department of Geography, San Diego State University, San Diego, CA, USA

Definition

The Bayesian maximum entropy (BME) theory of spatiotemporal geostatistics is concerned with the modeling and

estimation/mapping of natural attributes (such as physical variables, ecologic processes, health parameters, and social indicators varying in the chronotopologic domain) in a chronotopologic domain (both the space-time and the spacetime are considered here). BME has introduced significant advances in the field of spatiotemporal geostatistics including an improved knowledge-theoretic modeling framework under condition of in situ uncertainty that is free of the restrictive assumptions of classical geostatistics (normality, linearity etc.); it can integrate a variety of core-general and case-specific knowledge bases (including physical laws, scientific theories, theoretical models, empirical evidence, soft data, and auxiliary information).

Introduction

The introduction of *Bayesian maximum entropy* (BME; Christakos 1990) provided a new way to address some shortcomings of previous notions and techniques (e.g., purely spatial considerations, spatial statistical regression) by means of a more broadly designed epistemic framework for geostatistical analysis and mapping. For example, Kriging, a linear estimator, is a statistical error minimization technique that relies on the Gaussian assumption; it focuses on low-order moments, and it cannot process important types of uncertain information and core knowledge (Olea 1999). BME is a knowledge-based approach, which, based on a theoretical support of considerable generality, can integrate the content of different information sources to achieve a realistic representation of natural phenomena which generate predictions of improved accuracy (Christakos and Serre 2000; Lee et al. 2008). BME provides the foundations to integrate knowledge bases of various kinds, both site-specific bases (such as additional ground measurements, satellite observations, and empirical charts) and core bases (including physical laws, basic principles, and theoretical models) (Kolovos et al. 2012). This is a particularly attractive feature in contemporary data-driven analytical environments, because BME enables the generation and integration of valuable information from core knowledge bases that might be otherwise left unused (Kolovos et al. 2002). Most prominently, BME extends one's ability to use measured data beyond single-value observations (Christakos and Li 1998; Christakos 2000). Specifically with respect to uncertainty, BME discriminates between two types of data, namely, the hard data and the soft data. Data whose observations are considered to be exact in the scope of an analysis are termed hard data. In contrast, soft data represent observed data that might carry nontrivial uncertainty in the context of a study. For example, soft data can be the result of routine sources of uncertainty such as the inaccuracy of measuring devices, modeling uncertainties, and human error. Alternatively, soft data can be derived as a result of

inadequate knowledge, revised assessments, intuition estimates, and a host of other similar mechanisms that may provide uncertain informative content (Kolovos et al. 2016). BME is also free of additional limitations and restrictions that burden other geostatistical methods; for example, the BME analysis is independent of the data distribution (e.g., non-Gaussian distributions can be automatically incorporated), whereas Kriging is unable to accurately handle heavy tailed data (Christakos et al. 2001). In addition, the BME theoretical foundation enables both univariate and multivariate predictions (Hristopulos and Christakos 2001). Per the detailed method comparison in Christakos (2000), in the course of the decades since its introduction, BME has been providing equally or more accurate prediction patterns than Kriging (e.g., Lee et al. 2008; Bogaert et al. 2009; Lee et al. 2010; Christakos 2010; Li et al. 2013c; Gao et al. 2014; Shi et al. 2015a).

Since its inception, BME has demonstrated its robust ability for spatiotemporal data analysis and modeling through a wide spectrum of applications in Earth and environmental sciences, public health, ecology, remote sensing, energy, resources science, and real estate research. A sample of the relevant literature includes Christakos et al. (2004), Heywood et al. (2006), Law et al. (2006), Brus et al. (2008), Savelieva et al. (2010), Kolovos et al. (2010), Yu et al. (2011), Angulo et al. (2013), Li et al. (2013a, b), Shi et al. (2015b), Sun et al. (2015), Yang and Christakos (2015), Zagouras et al. (2015), Hu et al. (2016), Tang et al. (2016), Kou et al. (2016), Hayunga and Kolovos (2016), He et al. (2019a), Zhang and Yang (2019a, b), and Wei et al. (2020). The BME methodology will be briefly reviewed next.

BME Methodology

The Knowledge Bases (KB)

At the beginning of BME (and of any scientific inquiry, as a matter of fact), knowledge bases is the available *knowledge* about the phenomenon of interest. Broadly speaking, the knowledge considered in real-world applications comes from a variety of sources, including all kinds of valid information available at a given moment and obtained by the competent scientist using effectively a scientific procedure. BME distinguishes between two major *knowledge bases* (KB):

- (i) *Core or general G-KB*, which is background knowledge and justified beliefs relative to the phenomenon of interest. The *G* may include physical laws, scientific theories and models, structured patterns and assumptions, analytic and synthetic statements, and previous experience with similar phenomena independent of any case-specific data.

- (ii) *Site-specific* or *specificatory* *S*-KB, which is targeted knowledge about the specific situation that may include both external or demonstrative evidence (actual measurements, perceptual or data of sense, etc.) and internal or inductive evidence (inferring one thing from another thing, empirical propositions, expertise with similar situations).

The synthesis of the *G*-KB and the *S*-KB provides the *total* KB. More specifically, the *G*-KB is expressed in terms of a series of suitable functions g_α of the attribute realization χ_{map} at the point set $\mathbf{p}_{map}$ of interest, i.e., $g_\alpha(\chi_{map})$, $\alpha = 0, 1, \dots, N_c$ (by convention $g_0 = 1$). The g_α -functions account for a variety of core knowledge sources including statistical moments, empirical relationships, and physical laws. There are two conditions that guide the choice of the g_α -functions: (a) their stochastic expectations $\bar{g}_\alpha = \int d\chi_{map} g_\alpha f_G$ must be calculable at the point set $\mathbf{p}_{map}$ (which includes the sampled points and any unsampled points of interest), where f_G is a probability density function (PDF) constructed on the basis of the *G*-KB, and (b) the resulting equations must be solvable with respect to the unknown parameters of the g_α functions (these parameters are functions of $\mathbf{p}_{map}$). Examples of *G*-KB formulated in terms of the g_α -functions include (Christakos 2000; Wu et al. 2021) $\int d\chi_{map} g_\alpha O_\alpha[f_G] = 0$, where the g_α have forms associated with algebraic or differential operators O_α with respect to f_G (which includes derivatives of the attribute PDF with respect to the coordinates).

For practical purposes, the following *S*-KB distinction is usually adopted by BME: (a) *hard* data χ_h , which is the *S*-KB obtained from real-time observation devices with the probability of the attribute value in each datum being 1 (which is why it is also called *certain* data), and (b) *soft* data χ_s , which is the *S*-KB that includes non-negligible uncertainty (imprecision or vagueness) and is stochastically expressed as $\int d\Phi(\chi_s) f_S(\Phi) \in [0, 1]$, where f_S is a PDF constructed on the basis of the available *S*-KB and Φ and I are a specified empirical model (a wide variety of choices, depending on the real-world application) and interval of values, respectively (Wu et al. 2021).

Compared to the mainstream geostatistical approaches, BME not only follows a data-driven workflow but also considers general knowledge to describe a natural process or phenomena, including physical laws, basic principles, scientific theories, etc. In this sense, BME is a *theory-based data-driven* approach. Using this kind of knowledge may steer easier or more precise chronotopologic data analysis in the following cases: (i) an inadequate number of data might be available for conclusive analysis, and (ii) uncertainty in data can bias results and might even lead to flawed analyses if unacknowledged.

The Chronotopologic Domain

A *chronotopologic domain* is a term used to describe either a space-time domain (i.e., space and time are considered separately) or a spacetime domain (i.e., space and time are considered as an integrated whole). Chronotopologic domain points are denoted by the vector $\mathbf{p}$, which includes both spatial and temporal coordinates. In Table 1, a distinction is presented between sets of points $\mathbf{p}$ within the chronotopologic domain.

The Spatiotemporal Random Field Model

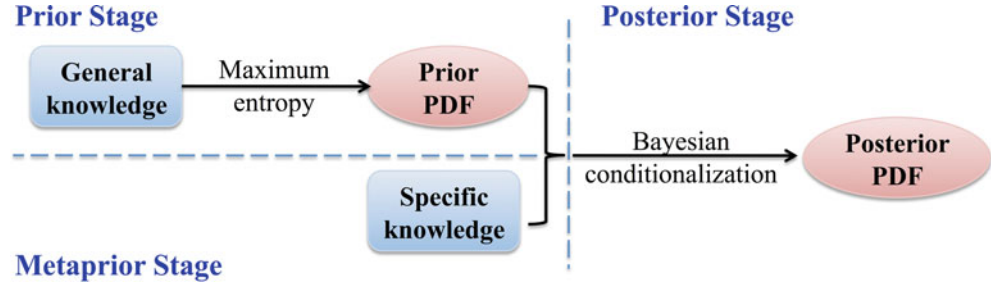
The *spatiotemporal random field* theory (STRF, Christakos 1991, 2017) studies natural attributes and phenomena that evolve in the chronotopologic continuum. An S/TRF $X(\mathbf{p})$ is a collection of complementary field realizations associated with the values of an attribute field at points of the chronotopologic domain, where the j^{th} -realization at $\mathbf{p}_{map}$ is written as $\chi_{map}^{(j)} = [\chi_{d1}^{(j)} \dots \chi_{dn_d}^{(j)} \chi_{k1}^{(j)} \dots \chi_{kn_k}^{(j)}]^T$ ($j = 1, 2, \dots$). In light of the notation in Table 1, it is commonly written that $\chi_{map} = [\chi_d \chi_k]$, where $\chi_{data} = [\chi_{d1} \dots \chi_{dn_d}]^T$ denotes attribute measurements or observations at points $\mathbf{p}_i$ and $\chi_k = [\chi_{k1} \dots \chi_{kn_k}]^T$ denotes the unknown attribute values at points $\mathbf{p}_k$ ($k \neq i$).

In stochastic terms, $X(\mathbf{p})$ is fully described by the infinite sequence of probability density functions (PDF) $f_X(\chi_{map})$ covering all domain points. The mean S/TRF function $\bar{X}(\mathbf{p})$ represents trends and systematic structures that exist in a larger scale, the covariance function $c_X(\mathbf{p}, \Delta\mathbf{p}) = [\bar{X}(\mathbf{p}) - \bar{X}(\mathbf{p})][\bar{X}(\mathbf{p} + \Delta\mathbf{p}) - \bar{X}(\mathbf{p} + \Delta\mathbf{p})]$ describes chronotopologic dependencies between pairs of points $\mathbf{p}$ and $\mathbf{p} + \Delta\mathbf{p}$, and the trivariogram function $\zeta_X(\mathbf{p}, \Delta\mathbf{p}, \Delta\mathbf{p}') = [\bar{X}(\mathbf{p}) - \bar{X}(\mathbf{p})][\bar{X}(\mathbf{p} + \Delta\mathbf{p}) - \bar{X}(\mathbf{p} + \Delta\mathbf{p})][\bar{X}(\mathbf{p} + \Delta\mathbf{p}') - \bar{X}(\mathbf{p} + \Delta\mathbf{p}')] describes chronotopologic$

Bayesian Maximum Entropy, Table 1 Chronotopologic point sets to be considered in BME

<i>Entire point set</i>	Vector $\mathbf{p}_{map} = [\mathbf{p}_1 \dots \mathbf{p}_n]^T$, which includes the nodes of the chronotopologic mapping grid that covers the entire domain of interest
<i>Hard data point subset</i>	Vector $\mathbf{p}_h = [\mathbf{p}_{h1} \dots \mathbf{p}_{hn_h}]^T \subset \mathbf{p}_{map}$, which includes the points where exact data is available
<i>Soft data point subset</i>	Vector $\mathbf{p}_s = [\mathbf{p}_{s1} \dots \mathbf{p}_{sn_s}]^T \subset \mathbf{p}_{map}$, which includes the points where uncertain data is available
<i>Data point subset</i>	Vector $\mathbf{p}_d = [\mathbf{p}_{d1} \dots \mathbf{p}_{dn_d}]^T \subset \mathbf{p}_h \cup \mathbf{p}_s$, i.e., the union of $\mathbf{p}_h$ and $\mathbf{p}_s$.
<i>Unsampled point subset</i>	Vector $\mathbf{p}_k = [\mathbf{p}_{k1} \dots \mathbf{p}_{kn_k}]^T$, where no empirical evidence exists but attribute estimates (interpolations, extrapolations, predictions) are sought
<i>Point subset link</i>	Relationship holds between the set and the two main point subsets leading to the entire point set: $\mathbf{p}_{map} = \mathbf{p}_d \cup \mathbf{p}_k$.

Bayesian Maximum Entropy,
Fig. 1 Workflow of the BME
analysis



dependencies between triplets of points p , $p + \Delta p$, and $p + \Delta p'$, where Δp and $\Delta p'$ are chronotopologic lags.

The BME Workflow

The basic BME stages are outlined in Fig. 1. Table 2 describes the BME methodology in terms of the g_a ($a = 1, \dots, N_c$) and Φ functions and the PDF f_G and f_S .

In more words, at the prior (*G-based*) stage, there is an inverse relation between prior information and prior probability: the more vague and general a theory is, the more alternatives it includes (and it is, hence, more probable) but also the less informative it is. In BME, maximization of information input serves as acquiring prior knowledge that can be filtrated, integrated, and optimized by specified rules and is a pivotal step to describe general attribute characteristics.

Thus, the prior PDF (*G-based*) f_G is defined by the calculated μ_a and the selected g_a , i.e., $f_G(\chi_{map}) = e^{\sum_{a=1}^{N_c} \mu_a g_a}$. The Lagrange multipliers μ_a ($a = 1, \dots, N_c$) calculated in the prior stage by information maximization quantify the significance of each g_a -function, whereas the μ_0 is a normalization parameter. The μ_a are themselves functions of p_{map} . Ideally, the more detailed G is, the thinner f_G becomes by excluding irrelevant outcomes, hence leading to a more informative model. Examples of the $(g_a, \bar{g}_a)$ pairs include theoretical models of the attribute mean, covariance, variogram, tri-variogram, higher-order chronotopologic moments, and equations expressing physical laws.

At the metaprior (*S-based*) stage, the *S-KB* is collected and organized in appropriate quantitative forms that can be explicitly incorporated into the BME formalism. This includes data cleaning, transformations, and evaluation, identification, and discrimination into hard and soft data, as appropriate for the analysis. Commonly encountered soft data forms are listed in Table 3.

In particular, the interval case in Table 3 shows that interval data values χ_i lie within a known interval I_i with lower and upper limits l_i and u_i , respectively; the probabilistic 1 case displays probabilistic data for which the cumulative distribution functions (CDF) $F_S(\zeta_i)$ are known; the probabilistic 2 case refers to threshold soft data, where the probabilities ϑ_i for certain

Bayesian Maximum Entropy, Table 2 Main BME stages

Stage	Formulation
Prior (<i>G-based</i>)	$\max_{\mu_a} \left[\text{Info}_G(\chi_{map}) = -\log e^{\sum_{a=1}^{N_c} \mu_a g_a} \right]$ $\text{s.t.} \int d\chi_{map} g_a e^{\sum_{a=1}^{N_c} \mu_a g_a} = \bar{g}_a$
Metaprior (<i>S-based</i>)	$\chi_d = \{\chi_h, \chi_s\}, \Phi, f_S$
Posterior (<i>K-based</i>)	$f_K(\chi_k) = \frac{1}{\int d\Phi f_S(\Phi) f_G(\chi_h, \Phi)} \int d\Phi f_S(\Phi) f_G(\chi_{map})$

Bayesian Maximum Entropy, Table 3 Examples of soft data formulations

Type of S-KB	Formulation $\chi_s = [\chi_{s1} \cdots \chi_{sn_s}]^T$ ($i = 1, \dots, n_s$)
Interval	$\chi_{si} \in I_i = [l_i, u_i]$
Probabilistic 1	$P_S[\chi_i \leq \zeta_i] = F_S(\zeta_i)$
Probabilistic 2	$P_S[\chi_i \in I_i] = \vartheta_i$
Functional	$P_S[h(\chi_i) \leq \zeta_i] = F_S(\zeta_i, h)$

threshold values I_i are known, and the functional case displays soft data, where the values can be derived from a given function h .

The posterior (*K-based*) stage is a synthesis stage in which the posterior probability is related to the *G-based* probability by means of a conditional probability knowledge-processing rule that draws inferences consistent with the available knowledge body and scientific reasoning. The *physical Bayesian conditionalization* rule is applied into the prior PDF f_G to produce the posterior PDF f_K and bring the model description closer to natural truth. According to the BME perspective, the *G-KB* introduces scientific consistency into the *S-KB*, and the *S-KB* introduces empirical conditioning into *G-KB* in light of the specific site considered. Interestingly, the above conclusion may be linked to the verifiability criterion used in sciences to constrain theoretical constructions by means of empirical evidence. With regard to material object claims, this criterion is known as “phenomenalism,” and with regard to the theoretical claims of science, as “operationalism.” Some examples of formulations corresponding to the interval, probabilistic, and functional soft data are presented

Bayesian Maximum Entropy, Table 4 Examples of posterior PDF formulations

Type of S-KB	Formulation $f_K(\chi_k)$
Interval	$\frac{1}{\int_{I_S} d\chi_S f_G(\chi_d)} \int_{I_S} d\chi_S f_G(\chi_k, \chi_d)$
Probabilistic 1	$\frac{1}{\int_{I_S} dF_G(\chi_s) f_G(\chi_d)} \int_{I_S} dF_G(\chi_s) f_G(\chi_k, \chi_s)$
Functional	$\frac{1}{\int_R dF_G(\zeta, h) \int_{I(\zeta)} d\chi_d f_G(\chi_d)} \int_R dF_G(\zeta, h) \int_{I(\zeta)} d\chi_S f_G(\chi_k, \chi_d)$

in Table 4, where I , R , and $I(\zeta)$ are definition domains that depend on the soft data.

BME Chronotopologic Estimation and Uncertainty Assessment

The obtained BME posterior PDF $f_K(\chi_k)$ at points $\mathbf{p}_k$ provides complete statistical description of the estimated (interpolated, extrapolated, predicted) attribute. In particular, the BME mode estimate $\hat{\chi}_{k,mode}$ is such that

$$f_K(\hat{\chi}_{k,mode}) = \max_{\chi_k} f_K(\chi_k);$$

the BME mean estimate $\hat{\chi}_{k,mean}$ is given by

$$\hat{\chi}_{k,mean} = \overline{X(\mathbf{p}_k)|\chi_d} = \int d\chi_k f_K(\chi_k) \chi_k;$$

and the BME median estimate $\hat{\chi}_{k,med}$ is defined as

$$\hat{\chi}_{k,med} = F_G^{-1}(0.5, \mathbf{p}_k) : \int_{\chi_k \leq \hat{\chi}_{k,med}} d\chi_k f_K(\chi_k) = \frac{1}{2}.$$

The determination of chronotopologic BME estimation uncertainty depends on the shape of the posterior PDF (symmetric vs. asymmetric, single maximum vs. multiple maxima, etc.; Wu et al. 2021). In particular, in the case of a symmetric attribute PDF, the estimation error variance is

$$e_X^2(\mathbf{p}_k) = \left[\overline{X(\mathbf{p}_k)} - \hat{X}(\mathbf{p}_k) \right]^2 = \int d\chi_k (\chi_k - \hat{\chi}_{k,mean})^2 f_K(\chi_k),$$

which is a widely used formula.

Interestingly, consider the special case where the following conditions are met concurrently: (i) the G -KB comprises only the attribute mean trend and covariance, that is, the first two stochastic moments of the attribute, and (ii) the S -KB contains only hard data. In this case, the BME equations reduce to Kriging, thus rendering Kriging a special case of chronotopologic BME estimation. In particular, the BME posterior PDF f_K is then fully defined by its mean and variance, and these measures correspond exactly to the Kriging prediction and prediction error, respectively.

BME Features and Applications

BME Versatility

Due to *BME*'s versatility, its basic equations possess significant generalization power on the basis of theory and data. As a consequence, the following essential *BME* features are of considerable interest as follows (Christakos 1998, 2000):

- (i) Well-known statistical interpolation techniques are derived as *special cases* of the chronotopologic BME theory. In the case, e.g., where the knowledge conditions $G = \{\bar{X}(\mathbf{p}), c_X(\mathbf{p}, \Delta\mathbf{p})\}$ and $S = \chi_h$ are met concurrently, the BME equations reduce to those of Kriging, thus rendering Kriging a special case of BME estimation (f_K is then determined by its mean and variance, which correspond to the Kriging estimation and estimation error variance, respectively).
- (ii) *BME* is a *nonlinear* chronotopologic estimator, in general.
- (iii) It incorporates *multi-sourced* knowledge bases (core and site-specific).
- (iv) It can process *higher-order* chronotopologic statistics associated with *non-Gaussian* data distributions, like skewness and kurtosis functions.
- (v) It has *extrapolation* (or *prediction*) abilities due to its incorporation of physical laws and other forms of core knowledge.
- (vi) It handles *functional*, *vector*, and *multipoint* chronotopologic estimation.
- (vii) It is *computationally* efficient, particularly in estimation cases where closed-form analytical expressions are available.

BME borrows strength from the Bayes rule and information theory, making it a versatile approach. On the one hand, the maximum entropy rule enables filtering out redundant knowledge and having useful knowledge left, which can also lead to more realistic chronotopologic estimations, and, on the other hand, the Bayes rule allows us to bring prior information (including not only statistical characteristic of the studied attribute but also physical laws, expert judgment, life experience, and skills) into considerations, which may save time/costs and lead to precise inference, even in the case that entire lack of data (Shoemaker et al. 1999; Christakos and Hristopulos 1998; Serre and Christakos 2002; Kolovos et al. 2002; McCarthy and Masters 2005; Kuhnert et al. 2010; Choy et al. 2009; Martin et al. 2012; Lang and Christakos 2019).

Another strength of the BME framework is the rigorous handling of soft data, i.e., uncertain data, such as historical records, secondary information, expert opinions, and beliefs. Therefore, BME can make maximal use of multi-sourced data and also provides a port to combine various mathematical

models (Money et al. 2011; Hu et al. 2016; Li et al. 2013a; Reyes and Serre 2014; He and Christakos 2018). This can significantly broaden the BME application range. Sometimes, soft data can be created by simple statistical methods for making up missing data in a time series case (Bayat et al. 2013; Lee et al. 2008). In other cases, the observed data may be vague, so sourcing expert advice, empirical handling of similar situations, intuition, and professional knowledge can be used to construct soft data (Serre et al. 2003; Puangthongthub et al. 2007; Yu and Wang 2013). Notice that compared to estimation without soft data, many studies have shown that including soft data can generate more accurate estimations (Christakos et al. 2001; D'Or et al. 2001; Shi et al. 2015b; Hayunga and Kolovos 2016).

The Emergence of Stochastic Logic in Geostatistics

The vast majority of the phenomena encountered in nature obey physical laws, which, in the case of classical physics, can be described in terms amenable to efficient *material causation* of logic theory. This is why BME's structure allows it to formally incorporate physical laws in a systematic and internally consistent manner. Mathematical logic is used because of the underlying deductive reasoning (if the premises are true, the conclusions are always true); on the other hand, when statistical inference is used (e.g., in statistical regression), because of the underlying inductive reasoning, the premises may be true and the conclusions wrong. Accordingly, BME has been extended in a stochastic logic context (Christakos 2002).

Indeed, beyond the conditional probability $P_G[\chi_k|\chi_d]$ used in the posterior stage of the original BME approach, other useful knowledge updating mechanisms can be expressed as the probability of *logical conditionals*, like the implication probability $P_G[\chi_d \rightarrow \chi_k]$, and the equivalence probability $P_G[\chi_d \leftrightarrow \chi_k]$, which are probabilistic expressions of the platitude that valid arguments are necessarily truth preserving. Otherwise said, these stochastic logic conditionals reveal that the relation of probability and deducibility involves the body of knowledge one actually has or the premises one actually accepts. Despite their conceptual differences, the above posterior probabilities are formally linked. For example, the standard (statistical) conditional probability is directly related to the implication probability as.

$[P_G[\chi_k|\chi_d] - 1]P_G[\chi_d] = P_G[\chi_d \rightarrow \chi_k] - 1$. Concluding, stochastic logic becomes a promising research avenue, when one realizes that the Bayes rule and the standard (statistical) conditional probability is not the only way to derive a posterior mapping probability in geostatistics.

BME in Software

There are several software libraries that can implement the BME theoretical developments, including the following: (i) BME library (BMELib) (Christakos et al. 2002), (ii) the

Spatiotemporal Epistemic Knowledge Synthesis Graphical User Interface (SEKS-GUI) (Yu et al. 2007), and (iii) the Space-Time Analysis Rendering with BME (STAR-BME) (Yu et al. 2012, 2016). The first two tools are Matlab-based compilations, while the third one was developed on Python. Each tool features a guided analysis through a sequence of screens across the different analysis tasks that start at importing data and extend all the way to visualization or results and exporting the analysis output.

Applications of BME in the Real World

Let us recall the basic function of BME approach, i.e., use known knowledge to explore the unknown natural process in space-time domain. Therefore, BME can be applied in a variety of disciplines (He and Kolovos 2018), as follows:

- The chronotopologic distribution of PM_{2.5}, ozone, NO_x concentrations, etc., was characterized and mapped for studying the human expose purpose by using BME or its combinations with land use regression models in the field of atmosphere (Christakos and Kolovos 1999; Bogaert et al. 2009; Nol et al. 2010; Yu and Wang 2010; Beckerman et al. 2013; Reyes and Serre 2014; Adam-Poupart et al. 2014; He and Christakos 2018; He et al. 2019a).
- For the lithosphere, BME was employed to study the heavy metal concentrations, electrical conductivity, moisture, organic matter in soil, and soil categories (D'Or and Bogaert 2003; Douaik et al. 2005; Brus et al. 2008; Modis et al. 2013; Gao et al. 2014; Sun et al. 2015).
- For the ecosphere, BME not only can study the natural beings (such as mercury concentration in fish tissue, young cork oak plantation, leaf area index, pedestrian mortality rates, etc., Money et al. 2011; Sedda et al. 2011; Li et al. 2013a; Fox et al. 2015) but also can depict the transmission and S/ST characteristics of a disease (such as black death, influenza, hand-foot-mouth, dengue fever, syphilis, hemorrhagic fever with renal syndrome, etc., Christakos et al. 2005; Law et al. 2006; Choi et al. 2008; Cao et al. 2010; Angulo et al. 2013; Yu et al. 2014; He et al. 2017, 2018, 2019b).
- Regarding the hydrosphere, groundwater contamination, level, and hydraulic conductivity were investigated by BME (Vyas et al. 2004; Messier et al. 2012, 2014, 2015; Yu and Chu 2010). Similarly, the inland water (such as the fecal contamination) was also characterized through BME (Coulliette et al. 2009; Money et al. 2009). Besides, the evolution of precipitation was studied in different areas with BME (Mahbub et al. 2010; Bayat et al. 2014; Zhang et al. 2016). With the largest area on the Earth, the oceans play a vital role in our planet's environmental cycle. Although the BME studies on oceans are not as much as on the other part of the Earth, but it has great potential to

show its versatile ability using the satellite remote sensing data on data fusion, spatiotemporal mapping, etc., see Li et al. (2013b), Tang et al. (2015), Shi et al. (2015b), and He et al. (2020a, 2020b).

Conclusions

BME provides a set of powerful tools supported by a rigorous theoretical framework, which are used to represent, estimate, and map natural attributes at unsampled locations under conditions of in situ uncertainty. BME displays a number of compelling characteristics on both theoretical and practical grounds, which make it stand out among mainstream geostatistics methods. On the theoretical front, BME makes no restrictive assumptions when it comes to properties like normality, linearity, and content-independency and is capable of ingesting information beyond the commonly used low-order moments and hard data. On the practical front, BME is built with the ability to integrate a variety of knowledge bases, including core knowledge bases (such as physical laws, theoretical models, and expert knowledge) and site-specific knowledge bases (such as empirical evidence, hard data, soft data, and secondary information).

Acknowledgement We would like to thank Dr. Alexander Kolovos for his valuable input to this work.

Bibliography

- Adam-Poupard A, Brand A, Fournier M, Jerrett M, Smargiassi A (2014) Spatiotemporal modeling of ozone levels in Quebec (Canada): a comparison of kriging, Land-Use Regression (LUR), and combined Bayesian maximum entropy-LUR approaches. *Environ Health Perspect* 122:970–976
- Angulo J, Yu H-L, Langousis A, Kolovos A, Wang J, Madrid AE, Christakos G (2013) Spatiotemporal infectious disease modeling: a BME-SIR approach. *PLoS One* 8:e72168
- Bayat B, Zahraie B, Taghavi F, Nasseri M (2013) Evaluation of spatial and spatiotemporal estimation methods in simulation of precipitation variability patterns. *Theor Appl Climatol* 113:429–444
- Bayat B, Nasseri M, Naser G (2014) Improving Bayesian maximum entropy and ordinary Kriging methods for estimating precipitations in a large watershed: a new cluster-based approach. *Can J Earth Sci* 51:43–55
- Beckerman BS, Jerrett M, Serre M, Martin RV, Lee SJ, van Donkelaar A, Ross Z, Su J, Burnett RT (2013) A hybrid approach to estimating national scale spatiotemporal variability of PM_{2.5} in the contiguous United States. *Environ Sci Technol* 47:7233–7241
- Bogaert P, Christakos G, Jerrett M, Yu H-L (2009) Spatiotemporal modelling of ozone distribution in the State of California. *Atmos Environ* 43:2471–2480
- Brus D, Bogaert P, Heuvelink G (2008) Bayesian maximum entropy prediction of soil categories using a traditional soil map as soft information. *Eur J Soil Sci* 59:166–177
- Cao C, Xu M, Chang C, Xue Y, Zhong S, Fang L, Cao W, Zhang H, Gao M, He Q, Zhao J, Chen W, Zheng S, Li X (2010) Risk analysis for the highly pathogenic avian influenza in Mainland China using meta-modeling. *Chin Sci Bull* 55:4168–4178
- Choi KM, Yu HL, Wilson ML (2008) Spatiotemporal statistical analysis of influenza mortality risk in the State of California during the period 1997–2001. *Stoch Env Res Risk A* 22:S15–S25
- Choy SL, O’Leary R, Mengersen K (2009) Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models. *Ecology* 90:265–277
- Christakos G (1990) A Bayesian/maximum-entropy view to the spatial estimation problem. *Math Geol* 22:763–777
- Christakos G (1991) On certain classes of spatiotemporal random-fields with applications to space-time data-processing. *IEEE Trans Syst Man Cybern* 21:861–875
- Christakos G (1998) Spatiotemporal information systems in soil and environmental sciences. *Geoderma* 85(2–3):141–179
- Christakos G (2000) Modern spatiotemporal geostatistics. Oxford University Press, New York
- Christakos G (2002) On a deductive logic-based spatiotemporal random field theory. *Theory Probab Math Stat* 66:54–65
- Christakos G (2010) Integrative problem-solving in a time of decadence. Springer, New York
- Christakos G (2017) Spatiotemporal random fields: theory and applications. Elsevier, Netherlands: Amsterdam
- Christakos G, Hristopulos DT (1998) Spatiotemporal environmental health modelling. Kluwer Academic Publishers, Boston
- Christakos G, Kolovos A (1999) A study of the spatiotemporal health impacts of ozone exposure. *J Expo Anal Environ Epidemiol* 9: 322–335
- Christakos G, Li X (1998) Bayesian maximum entropy analysis and mapping: a farewell to kriging estimators? *Math Geol* 30:435–462
- Christakos G, Serre ML (2000) BME analysis of spatiotemporal particulate matter distributions in North Carolina. *Atmos Environ* 34: 3393–3406
- Christakos G, Serre ML, Kovitz JL (2001) BME representation of particulate matter distributions in the state of California on the basis of uncertain measurements. *J Geophys Res Atmos* 106:9717–9731
- Christakos G, Bogaert P, Serre M (2002) Temporal GIS: advanced functions for field-based applications. Springer Science & Business Media, German: Berlin
- Christakos G, Kolovos A, Serre ML, Vukovich F (2004) Total ozone mapping by integrating databases from remote sensing instruments and empirical models. *IEEE Trans Geosci Remote Sens* 42: 991–1008
- Christakos G, Olea RA, Serre ML, Wang LL, Yu HL (2005) Interdisciplinary public health reasoning and epidemic modelling: the case of black death. Springer, Berlin/Heidelberg
- Coulliette AD, Money ES, Serre ML, Noble RT (2009) Space/time analysis of fecal pollution and rainfall in an Eastern North Carolina Estuary. *Environ Sci Technol* 43:3728–3735
- D’Or D, Bogaert P (2003) Continuous-valued map reconstruction with the Bayesian maximum entropy. *Geoderma* 112:169–178
- D’Or D, Bogaert P, Christakos G (2001) Application of the BME approach to soil texture mapping. *Stoch Env Res Risk A* 15:87–100
- Douaik A, Van Meirvenne M, Toth T (2005) Soil salinity mapping using spatio-temporal kriging and Bayesian maximum entropy with interval soft data. *Geoderma* 128:234–248
- Fox L, Serre ML, Lippmann SJ, Rodriguez DA, Bangdiwala SI et al (2015) Spatiotemporal approaches to analyzing pedestrian fatalities: the case of Cali, Colombia. *Traffic Inj Prev* 16:571–577
- Gao S, Zhu Z, Liu S, Jin R, Yang G, Tan L (2014) Estimating the spatial distribution of soil moisture based on Bayesian maximum entropy method with auxiliary data from remote sensing. *Int J Appl Earth Obs Geoinf* 32:54–66
- Hayunga DK, Kolovos A (2016) Geostatistical space–time mapping of house prices using Bayesian maximum entropy. *Int J Geogr Inf Sci*. <https://doi.org/10.1080/13658816.2016.1165820>

- He J, Christakos G (2018) Space-time PM2.5 mapping in the severe haze region of Jing-Jin-Ji (China) using a synthetic approach. *Environ Pollut* 240:319–329
- He J, Kolovos A (2018) Bayesian maximum entropy approach and its application: a review. *Stoch Env Res Risk A* 32(4):859–877
- He J, Christakos G, Zhang W, Wang Y (2017) A space-time study of hemorrhagic fever with Renal Syndrome (HFRS) and its climatic associations in Heilongjiang Province, China. *Front Appl Math Stat* 3:16
- He J, Christakos G, Wu J, Cazelles B, Qian Q, Mu D, Wang Y, Yin W, Zhang W (2018) Spatiotemporal variation of the association between climate dynamics and HFRS outbreaks in Eastern China during 2005–2016 and its geographic determinants. *PLoS Negl Trop Dis* 12(6):e0006554
- He J, Christakos G, Jankowski P (2019a) Comparative performance of the LUR, ANN and BME techniques in the multi-scale spatiotemporal mapping of PM2.5 concentrations in North China. *IEEE J Sel Topics Appl Earth Observ Remote Sens* 12(6):1734–1747
- He J, Christakos G, Wu J, Jankowski P, Langousis A, Wang Y, Yin W, Zhang W (2019b) Probabilistic logic analysis of the highly heterogeneous spatiotemporal HFRS incidence distribution in Heilongjiang Province (China) during 2005–2013. *PLoS Negl Trop Dis* 13(1):e0007091
- He J, Chen Y, Wu J, Stow D, Christakos G (2020a) Space-time chlorophyll-a retrieval in optically complex waters that accounts for remote sensing and modeling uncertainties and improves remote estimation accuracy. *Water Res* 171:1–17
- He M, He J, Christakos G (2020b) Space-time mapping of sea surface salinity in western Pacific ocean using contingency modelling. *Stoch Env Res Risk A* 34:355–368
- Heywood B, Brierley A, Gull S (2006) A quantified Bayesian maximum entropy estimate of Antarctic krill abundance across the Scotia Sea and in small-scale management units from the CCAMLR-2000 survey. *Ccamlr Sci* 13:97–116
- Hristopulos DT, Christakos G (2001) Practical calculation of non-Gaussian multivariate moments in spatiotemporal Bayesian maximum entropy analysis. *Math Geol* 33(5):543–568
- Hu JG, Zhou J, Zhou GM, Luo YQ, Xu XJ, Li PH, Liang JY (2016) Improving estimations of spatial distribution of soil respiration using the Bayesian maximum entropy algorithm and soil temperature as auxiliary data. *PLoS One* 11(1):e0146589
- Kolovos A, Christakos G, Serre ML, Miller CT (2002) Computational BME solution of a stochastic advection-reaction equation in the light of site-specific information. *Water Resour Res* 38(12):1318–1334
- Kolovos A, Skupin A, Jerrett M, Christakos G (2010) Multi-perspective analysis and spatiotemporal mapping of air pollution monitoring data. *Environ Sci Technol* 44:6738–6744
- Kolovos A, Angulo JM, Modis K, Papantonopoulos G, Wang J-F, Christakos G (2012) Model-driven development of covariances for spatiotemporal environmental health assessment. *Environ Monit Assess*. <https://doi.org/10.1007/s10661-012-2593-1>
- Kolovos A, Smith LM, Schwab-McCoy A, Gengler S, Yu H-L (2016) Emerging patterns in multi-sourced data modeling uncertainty. *Spat Stat* 18A:300–317. <https://doi.org/10.1016/j.spasta.2016.05.005>
- Kou X, Jiang L, Bo Y, Yan S, Linna Chai L (2016) Estimation of land surface temperature through blending MODIS and AMSR-E data with the Bayesian maximum entropy method. *Remote Sens* 2016(8):105. <https://doi.org/10.3390/rs8020105>
- Kuhnert PM, Martin TG, Griffiths SP (2010) A guide to eliciting and using expert knowledge in Bayesian ecological models. *Ecol Lett* 13:900–914
- Lang Y, Christakos G (2019) Ocean pollution assessment by integrating physical law and site-specific data. *Environmetrics* 30(3):e2547
- Law DCG, Bernstein KT, Serre ML, Schumacher CM, Leone PA, Zenilman JM, Miller WC, Rompalo AM (2006) Modeling a syphilis outbreak through space and time using the Bayesian maximum entropy approach. *Ann Epidemiol* 16:797–804
- Lee S-J, Balling R, Gober P (2008) Bayesian maximum entropy mapping and the soft data problem in urban climate research. *Ann Assoc Am Geogr* 98:309–322
- Lee S-J, Wentz EA, Gober P (2010) Space-time forecasting using soft geostatistics: a case study in forecasting municipal water demand for Phoenix, Arizona. *Stoch Env Res Risk A* 24:283–295
- Li AH, Bo YC, Chen L (2013a) Bayesian maximum entropy data fusion of field-observed leaf area index (LAI) and landsat enhanced thematic mapper plus-derived LAI. *Int J Remote Sens* 34:227–246
- Li AH, Bo YC, Zhu YX, Guo P, Bi J, He YQ (2013b) Blending multi-resolution satellite sea surface temperature (SST) products using Bayesian maximum entropy method. *Remote Sens Environ* 135: 52–63
- Li X, Li P, Zhu H (2013c) Coal seam surface modeling and updating with multi-source data integration using Bayesian geostatistics. *Eng Geol* 164:208–221
- Mahbub P, Ayoko GA, Goonetilleke A, Egodawatta P, Kokot S (2010) Impacts of traffic and rainfall characteristics on heavy metals build-up and wash-off from urban roads. *Environ Sci Technol* 44: 8904–8910
- Martin TG, Burgman MA, Fidler F, Kuhnert PM, Low-Choy S, McBride M, Mengersen K (2012) Eliciting expert knowledge in conservation science. *Conserv Biol* 26:29–38
- McCarthy MA, Masters P (2005) Profiting from prior information in Bayesian analyses of ecological data. *J Appl Ecol* 42:1012–1019
- Messier KP, Akita Y, Serre ML (2012) Integrating address geocoding, land use regression, and spatiotemporal geostatistical estimation for groundwater tetrachloroethylene. *Environ Sci Technol* 46(5): 2772–2780
- Messier KP, Kane E, Bolich R, Serre ML (2014) Nitrate variability in groundwater of North Carolina using monitoring and private well data models. *Environ Sci Technol* 48:10804–10812
- Messier KP, Campbell T, Bradley PJ, Serret ML (2015) Estimation of groundwater radon in North Carolina using land use regression and Bayesian maximum entropy. *Environ Sci Technol* 49:9817–9825
- Modis K, Vatalis KI, Sachanidis C (2013) Spatiotemporal risk assessment of soil pollution in a lignite mining region using a Bayesian maximum entropy (BME) approach. *Int J Coal Geol* 112:173–179
- Money ES, Carter GP, Serre ML (2009) Modern space/time geostatistics using river distances: data integration of turbidity and E.coli measurements to assess fecal contamination along the Raritan river in New Jersey. *Environ Sci Technol* 43:3736–3742
- Money ES, Sackett DK, Aday DD, Serre ML (2011) Using river distance and existing hydrography data can improve the geostatistical estimation of fish tissue mercury at unsampled locations. *Environ Sci Technol* 45:7746–7753
- Nol L, Heuvelink GBM, Veldkamp A, de Vries W, Kros J (2010) Uncertainty propagation analysis of an N₂O emission model at the plot and landscape scale. *Geoderma* 159:9–23
- Olea RA (1999) *Geostatistics*. Kluwer Academic Publication, Boston
- Puangthongthub S, Wangwongwatana S, Kamens RM, Serre ML (2007) Modeling the space/time distribution of particulate matter in Thailand and optimizing its monitoring network. *Atmos Environ* 41: 7788–7805
- Reyes JM, Serre ML (2014) An LUR/BME framework to estimate PM2.5 explained by on road mobile and stationary sources. *Environ Sci Technol* 48:1736–1744
- Savelieva E, Utkin S, Kazakov S, Demyanov V (2010) Modeling spatial uncertainty for locally uncertain data. In: *geoENV VII – Geostatistics for environmental applications*, pp 295–306.
- Sedda L, Atkinson PM, Filigheddu MR, Cotzia G, Dettori S (2011) Spatio-temporal analysis of tree height in a young cork oak plantation. *Int J Geogr Inf Sci* 25:1083–1096

- Serre ML, Christakos G (2002) BME-based hydrogeologic parameter estimation in groundwater flow modelling. *Acta Univ Carol Geol* 46: 566–570
- Serre ML, Kolovos A, Christakos G, Modis K (2003) An application of the holistochastic human exposure methodology to naturally occurring arsenic in Bangladesh drinking water. *Risk Anal* 23:515–528
- Shi T, Yang X, Christakos G, Wang J, Liu L (2015a) Spatiotemporal interpolation of rainfall by combining BME theory and satellite rainfall estimates. *Atmos* 6:1307–1326
- Shi Y, Zhou X, Yang X, Shi L, Ma S (2015b) Merging satellite ocean color data with Bayesian maximum entropy method. *IEEE J Sel Topics Appl Earth Observ Remote Sens* 8:3294–3304
- Shoemaker JS, Painter IS, Weir BS (1999) Bayesian statistics in genetics – a guide for the uninitiated. *Trends Genet* 15:354–358
- Sun XL, Wu YJ, Lou YL, Wang HL, Zhang C, Zhao YG, Zhang GL (2015) Updating digital soil maps with new data: a case study of soil organic matter in Jiangsu, China. *Eur J Soil Sci* 66:1012–1022
- Tang SL, Yang XF, Dong D, Li ZW (2015) Merging daily sea surface temperature data from multiple satellites using a Bayesian maximum entropy method. *Front Earth Sci* 9:722–731
- Tang Q, Bo Y, Zhu Y (2016) Spatiotemporal fusion of multiple- satellite aerosol optical depth (AOD) products using Bayesian maximum entropy method. *J Geophys Res Atmos* 121:4034–4048. <https://doi.org/10.1002/2015JD024571>
- Vyas VM, Tong SN, Uchirin C, Georgopoulos PG, Carter GR (2004) Geostatistical estimation of horizontal hydraulic conductivity for the Kirkwood-Cohansey aquifer. *J Am Water Resour Assoc* 40: 187–195
- Wei X, Chang N, Bai K (2020) A comparative assessment of multisensor data merging and fusion algorithms for high-resolution surface reflectance data. *IEEE J Sel Topics Appl Earth Observ Remote Sens* 13:4044–4059. <https://doi.org/10.1109/JSTARS.2020.3008746>
- Wu JP, He J, Christakos G (2021) Quantitative analysis and modeling of earth and environmental data: space-time and spacetime data considerations. Elsevier, New York
- Yang Y, Christakos G (2015) Spatiotemporal characterization of ambient PM_{2.5} concentrations in Shandong Province (China). *Environ Sci Technol* 49:13431–13438
- Yu HL, Chu HJ (2010) Understanding space-time patterns of groundwater system by empirical orthogonal functions: a case study in the Choshui River alluvial fan, Taiwan. *J Hydrol* 381:239–247
- Yu HL, Wang CH (2010) Retrospective prediction of intraurban spatiotemporal distribution of PM_{2.5} in Taipei. *Atmos Environ* 44: 3053–3065
- Yu HL, Wang CH (2013) Quantile-based Bayesian maximum entropy approach for spatiotemporal modeling of ambient air quality levels. *Environ Sci Technol* 47:1416–1424
- Yu HL, Kolovos A, Christakos G, Chen JC, Warnerdam S, Dev B (2007) Interactive spatiotemporal modelling of health systems: the SEKS-GUI framework. *Stoch Env Res Risk A* 21:555–572
- Yu HL, Wang CH, Liu MC, Kuo YM (2011) Estimation of fine particulate matter in Taipei using landuse regression and bayesian maximum entropy methods. *Int J Environ Res Public Health* 8(6): 2153–2169. <https://doi.org/10.3390/ijerph8062153>
- Yu H-L, Ku S-J, Kolovos A (2012) Advanced space-time predictive analysis with STAR-BME. In: Proceedings of the 20th international conference on advances in geographic information systems. ACM, pp 593–596
- Yu H-L, Angulo JM, Chen M-H, Wu J, Christakos G (2014) An online spatiotemporal prediction model for dengue fever epidemic in Kaohsiung (Taiwan). *Biom J* 56(3):428–440. <https://doi.org/10.1002/bimj.201200270>
- Yu HL, Ku SC, Kolovos A (2016) A GIS tool for spatiotemporal modeling under a knowledge synthesis framework. *Stoch Env Res Risk A* 30:665–679
- Zagouras A, Kolovos A, Coimbra CFM (2015) Objective framework for optimal distribution of solar irradiance monitoring networks. *Renew Energy* 80:153–165. <https://doi.org/10.1016/j.renene.2015.01.046>
- Zhang C, Yang Y (2019a) Can the spatial prediction of soil organic matter be improved by incorporating multiple regression confidence intervals as soft data into BME method? *Catena* 178:322–334
- Zhang C, Yang Y (2019b) Improving the spatial prediction of soil Zn by converting outliers into soft data for BME method. *Stoch Env Res Risk A* 33:855–864
- Zhang FS, Yang ZT, Zhong SB, Huang QY (2016) Exploring mean annual precipitation values (2003–2012) in a specific area (36 degrees N–43 degrees N, 113 degrees E–120 degrees E) using meteorological, elevational, and the nearest distance to coastline variables. *Adv Meteorol* 2016:2107908. <https://doi.org/10.1155/2016/2107908>, 13 p

Best Linear Unbiased Estimation

Peter K. Kitanidis

Stanford University, Civil and Environmental Engineering & Computational and Mathematical Engineering, Stanford, CA, USA

Definition

Best linear unbiased estimation (BLUE) is a widely used data analysis and estimation methodology. Earth scientists and engineers are acquainted with this methodology in solving interpolation problems, *e.g.*, using Kriging, and data assimilation, using the ensemble Kalman filter (EnKF). Although based on probabilistic concepts, the method uses only the first two moments of probability distributions, *i.e.*, mean values and variances and covariances. Estimators are linear functions of data so the expected value of the error vanishes and the expected value of squared error is as small as possible.

Introduction

In the 1960s, pioneering works (Kalman 1960; Matheron 1963; Gandin 1965) introduced BLUE methods into applications. The objective in this chapter is to summarize the key ideas and equations of this methodology.

To fix ideas, assume we have measured a quantity z , such as temperature that varies with spatial coordinate x , at a number of locations, $x_1, x_2, \dots, x_n$. The measured quantities $z_1, z_2, \dots, z_n$ constitute our *data*. Our objective is to *estimate* the value of z at intermediate locations x . There are many ways to obtain a value of z at x from the data. The vast majority of them compute a value $\hat{z}(x)$ using linear operations. For example, for $x_1 < x < x_2$, $\hat{z}(x) = \lambda_1 z_1 + \lambda_2 z_2$, where $\lambda_1 = \frac{x_2 - x}{x_2 - x_1}$, $\lambda_2 = \frac{x - x_1}{x_2 - x_1}$. The estimated function consists of straight linear segments between successive data points. By

choosing more points and different weights, one can estimate a function that is smooth, meaning it has continuous first derivatives. One may also utilize weighting schemes in which the estimate reproduces the observations only approximately in favor of more uniformity.

One issue is to select a weighting scheme that is more accurate. Another issue is to quantify the estimation uncertainty, *i.e.*, evaluate in an average sense how close the estimate may be to the true value. Both issues are challenging. Best linear unbiased estimation (BLUE) is a statistical methodology to address these issues in a systematic and logical way. The hallmark of this approach is practicality, with low informational and computational requirements. Significantly, the method uses only the first two moments of probability distributions, *i.e.*, mean values and variances and covariances. The main features of the methodology are:

- The estimate of the unknown quantity is computed as the linear function or superposition of the data (plus fixed effects).
- The weights are selected so that in a statistical (expected value) sense:
 1. the difference between the true value and the estimate is zero (unbiasedness)
 2. the squared difference is as small as possible (minimum variance or “best.”)

Basic Ideas

We conceptualize that jointly the unknowns and the observations belong in an ensemble of possible unknown-observation sets. Each member of the ensemble has a probability of being the true one. The model only specifies and the methodology only uses ensemble averages, also known as expected values, of linear and quadratic functions of the variables. Furthermore, the unknown and the observations are statistically related (or “correlated”).

The symbol $E[\cdot]$ indicates averaging, known as expected value. For illustration, for a probabilistically characterized variable z , the mean, the variance, and the average of an arbitrary quadratic function are as follows:

$$E[z] = \mu \quad (1)$$

$$E[(z - \mu)^2] = \sigma^2 \quad (2)$$

$$E[az + bz^2] = a\mu + b(\mu^2 + \sigma^2) \quad (3)$$

where a, b are deterministic variables (in other words fixed numbers).

Next, consider that we have two variables: s with mean μ and variance σ^2 and y with mean μ_y and variance σ_y^2 . Furthermore, the covariance between s and y is C_{sy} .

$$\begin{aligned} E[s] &= \mu, E[(s - \mu)^2] = \sigma^2 \\ E[y] &= \mu_y, E[(y - \mu_y)^2] = \sigma_y^2, E[(s - \mu)(y - \mu_y)] = C_{sy} \end{aligned} \quad (4)$$

These relations express what we know about these two variables. *E.g.*, our best estimate of s is μ with mean squared error σ^2 . They also express the interdependence between two variables. When y is observed, the estimate of s can be improved. We look for the best estimate $\hat{s}$ that is a linear combination of the old estimate and the observed value of y :

$$\hat{s} = \mu + \lambda(y - \mu_y) \quad (5)$$

where λ is a weight. The estimation error is

$$\hat{s} - s = \mu + \lambda(y - \mu_y) - s \quad (6)$$

with expected value 0, meaning that the estimator is unbiased. We computed expected values by treating s and y as random variables with the probabilistic properties defined above.

We select λ by requiring minimum variance, *i.e.*, minimum mean squared error. The mean squared error is

$$\begin{aligned} E[(\hat{s} - s)^2] &= E[(s - \mu)^2 + \lambda^2(y - \mu_y)^2 - 2\lambda(s - \mu)(y - \mu_y)] \\ &= \sigma^2 + \lambda^2\sigma_y^2 - 2\lambda C_{sy} \end{aligned} \quad (7)$$

Setting the derivative with respect to λ equal to 0 we get $2\lambda\sigma_y^2 - 2C_{sy} = 0$, and thus $\lambda = \frac{C_{sy}}{\sigma_y^2}$. Consequently, the best linear unbiased estimator (BLUE) is

$$\hat{s} = \mu + \frac{C_{sy}}{\sigma_y^2}(y - \mu_y) \quad (8)$$

with mean squared error

$$E[(\hat{s} - s)^2] = \sigma^2 - \frac{C_{sy}^2}{\sigma_y^2} \quad (9)$$

We can express the results in terms of the familiar correlation coefficient ρ . Using that $C_{sy} = \rho\sigma\sigma_y$,

$$\hat{s} = \mu + \rho \frac{\sigma}{\sigma_y}(y - \mu_y) \quad (10)$$

with mean squared error

$$E[(\hat{s} - s)^2] = (1 - \rho^2)\sigma^2 \quad (11)$$

Thus, the updated estimate is the old estimate plus a correction that is linear in $y - \mu_y$ and a coefficient that is the correlation coefficient between the observation and the quantity we want to estimate times a ratio of normalizing factors, $\frac{\sigma_y}{\sigma_y}$. The mean squared error depends on the correlation coefficient and vanishes when $|\rho| = 1$.

Application: Kriging

We turn our attention to an interpolation method. Consider a function $z(\mathbf{x})$, where $\mathbf{x}$ denotes case-specific coordinates, like in 1D, 2D, or 3D. An example may be the precipitation over a watershed, or the concentration of a solute in an aquifer, or the pressure inside a reactor.

Given n observations of z at locations with coordinates $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, our objective is to estimate the value of z at location $\mathbf{x}_0$. A common model is the following:

$$\begin{aligned} E[z(\mathbf{x}_i)] &= \mu, \text{ for } i = 0, 1, 2, \dots, n \\ E[(z(\mathbf{x}_i) - \mu)(z(\mathbf{x}_j) - \mu)] &= C_{ij}, \text{ for } i, j = 0, 1, 2, \dots, n \end{aligned} \quad (12)$$

where μ is the mean and C_{ij} is the covariance between values at points $\mathbf{x}_i$ and $\mathbf{x}_j$.

We simplify the notation by adopting $z_i = z(\mathbf{x}_i)$. The unbiased linear estimator is

$$\hat{z}_0 = \mu + \sum_{i=1}^n \lambda_i (z_i - \mu) \quad (13)$$

The mean squared error is:

$$E[(\hat{z}_0 - z_0)^2] = \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j C_{ij} - 2 \sum_{i=1}^n \lambda_i C_{i0} + C_{00} \quad (14)$$

We select the λ weights to minimize this expression. We set the derivative with respect to each weight equal to 0, thus obtaining a system of n linear equations with n unknowns:

$$\sum_{j=1}^n \lambda_j C_{ij} = C_{i0}, \text{ for } i = 1, 2, \dots, n \quad (15)$$

Solving this system, we obtain the λ values that minimize the mean squared error that becomes

$$E[(\hat{z}_0 - z_0)^2] = C_{00} - \sum_{i=1}^n \lambda_i C_{i0} \quad (16)$$

This approach is known as simple Kriging. In practice, the main challenge is to select appropriate μ and C_{ij} values.

An important variant is when the mean values of y are modeled as the same everywhere but with an unknown value μ . The unbiased linear estimator then is

$$\hat{y}_0 = \sum_{i=1}^n \lambda_i y_i, \text{ where } \sum_{i=1}^n \lambda_i = 1 \quad (17)$$

The mean squared error is given by Eq. (14). The challenge is to find the weights that minimize this expression subject to the constraint that they must sum up to 1. This constrained optimization problem is solved using the method of Lagrange multipliers. We define the Lagrangian

$$\begin{aligned} \mathcal{L} &= \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j C_{ij} - 2 \sum_{i=1}^n \lambda_i C_{i0} + C_{00} \\ &\quad + 2v \left(\sum_{i=1}^n \lambda_i - 1 \right) \end{aligned} \quad (18)$$

then set the derivative of $\mathcal{L}$ with respect to each weight λ and the Lagrange multiplier v equal to 0, thus obtaining a system of $n + 1$ linear equations with $n + 1$ unknowns.

$$\begin{aligned} \sum_{j=1}^n \lambda_j C_{ij} + v &= C_{i0}, \text{ for } i = 1, 2, \dots, n \\ \sum_{j=1}^n \lambda_j &= 1 \end{aligned} \quad (19)$$

This method is known as ordinary Kriging.

The approach can be extended in a number of ways. For example, universal Kriging, where the mean is variable and in the form $\sum_{k=1}^p f_k(\mathbf{x}) \beta_k$, where $f_1, f_2, \dots, f_p$ are known functions of the coordinates $\mathbf{x}$ and $\beta_1, \beta_2, \dots, \beta_p$ are deterministic but unknown coefficients. The main challenge in the application of these methods is in the selection of an appropriate model. Kriging and related issues are covered in books on linear Geostatistics (for example, Kitanidis 1997). See article *Kriging*.

Inverse Problems

Basic Form

We now turn our attention to a formulation that is suitable for solving inverse problem. Inverse problems are estimation

problems for which data alone are not sufficient to provide a unique answer. These problems are underdetermined or ill posed: There are many solutions that reproduce the data. BLUE can handle such problems because it includes probabilistic modeling of the unknown.

We adopt a matrix-vector notation, or else the notation gets out of hand. The unknowns are arranged in an m -dimensional column vector $\mathbf{s}$ and the observations in an n -dimensional column vector $\mathbf{y}$. The relation between the two vectors is linear

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{v} \quad (20)$$

where $\mathbf{H}$ is a given n by m matrix, and $\mathbf{v}$ is random with mean $\mathbf{0}$ and covariance matrix $\mathbf{R}$. In many applications $\mathbf{R} \propto \mathbf{I}$, where $\mathbf{I}$ is the identity matrix of size n .

The $\mathbf{s}$ is also random with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{Q}$, and $\mathbf{s}$ is uncorrelated from $\mathbf{v}$. Then, $\mathbf{y}$ has mean $\mathbf{H}\boldsymbol{\mu}$ and covariance $\mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R}$, while the cross-covariance between $\mathbf{s}$ and $\mathbf{y}$ is $\mathbf{Q}\mathbf{H}^T$. Then, following the familiar BLUE methodology, we obtain the following algorithm. First compute $\mathbf{A}^T$, using linear system solvers,

$$(\mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R})\mathbf{A}^T = \mathbf{H}\mathbf{Q} \quad (21)$$

or, explicitly,

$$\mathbf{A} = \mathbf{Q}\mathbf{H}^T (\mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R})^{-1} \quad (22)$$

Then, the best estimate $\hat{\mathbf{s}}$ and matrix of mean squared estimation errors $\boldsymbol{\Sigma}$ are

$$\hat{\mathbf{s}} = \boldsymbol{\mu} + \mathbf{A}(\mathbf{y} - \mathbf{H}\boldsymbol{\mu}) \quad (23)$$

$$\boldsymbol{\Sigma} = \mathbf{Q} - \mathbf{A}\mathbf{H}\mathbf{Q} \quad (24)$$

The only requirement is that $\mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R}$ is invertible. These equations are very general; for example, they can give the simple Kriging equations and are used in the Kalman filter.

Unknown Coefficients in the Mean

Consider now that the mean vector $\boldsymbol{\mu}$ is not fully predetermined but instead is

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \quad (25)$$

where $\mathbf{X}$ is a known m by p matrix and $\boldsymbol{\beta}$ is an p -dimensional column vector of unknown coefficients. The BLUE method gives the following algorithm (Kitanidis 1995).

First, compute two matrices: the m by n matrix $\mathbf{A}$ and the p by m matrix $\mathbf{M}$ from the linear system

$$\left[\begin{array}{c|c} \mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R} & \mathbf{H}\mathbf{X} \\ \hline (\mathbf{H}\mathbf{X})^T & \mathbf{0} \end{array} \right] \left[\begin{array}{c} \mathbf{A}^T \\ \mathbf{M} \end{array} \right] = \left[\begin{array}{c} \mathbf{H}\mathbf{Q} \\ \mathbf{X}^T \end{array} \right] \quad (26)$$

The matrix of coefficients is a 4×4 block matrix, with overall size $(n + p) \times (n + p)$. Then, the mean is computed through

$$\hat{\mathbf{s}} = \mathbf{A}\mathbf{y} \quad (27)$$

and the covariance is

$$\boldsymbol{\Sigma} = \mathbf{Q} - \mathbf{X}\mathbf{M} - \mathbf{Q}\mathbf{H}^T\mathbf{A}^T \quad (28)$$

These equations have many applications, like they can yield the ordinary Kriging equations.

The Kalman Filter

The Model

Let us first discuss two important concepts: (1) dynamic system and (2) state-space representation underlying the Kalman filter.

The evolution of many systems encountered in applications can be described by a linear discrete-time state-space structure, as follows:

$$\mathbf{s}(k+1) = \boldsymbol{\Phi}_k \mathbf{s}(k) + \mathbf{u}_k + \mathbf{w}_k \quad (29)$$

where $\mathbf{s}(k)$ is the state at stage k , $\boldsymbol{\Phi}_k$ is a known transition matrix, $\mathbf{u}_k$ is a known (deterministic) input, and $\mathbf{w}_k$ is a random (probabilistically defined) input.

The relation between state and observation is also linear:

$$\mathbf{y}(k+1) = \mathbf{H}_{k+1} \mathbf{s}(k+1) + \mathbf{v}_{k+1} \quad (30)$$

where $\mathbf{y}(k+1)$ is data obtained at $k+1$, $\mathbf{H}_{k+1}$ is a known observation matrix, and $\mathbf{v}_{k+1}$ is a random (probabilistically defined) noise term.

An important part of the model is that the stochastic terms $\mathbf{w}$ and $\mathbf{v}$ that drive the process are modeled as discrete white noise processes, which means the following:

For all k and l ,

$$E[\mathbf{w}_k] = \mathbf{0} \quad (31)$$

$$E[\mathbf{w}_k \mathbf{w}_l^T] = \begin{cases} \mathbf{Q}_k, & k = l \\ 0, & k \neq l \end{cases} \quad (32)$$

$$E[\mathbf{v}_k] = \mathbf{0} \quad (33)$$

$$E[\mathbf{v}_k \mathbf{v}_l^T] = \begin{cases} \mathbf{R}_k, & k = l \\ 0, & k \neq l \end{cases} \quad (34)$$

$$E[\mathbf{w}_k \mathbf{v}_l^T] = \mathbf{0} \quad (35)$$

$$E[\mathbf{s}(0) \mathbf{v}_k^T] = \mathbf{0} \quad (36)$$

$$E[\mathbf{s}(0) \mathbf{w}_k^T] = \mathbf{0} \quad (37)$$

The model is versatile and can be applied to many actual problems. The state-space structure is a natural way to represent the evolution of most systems. The crucial idea is that $\mathbf{s}(k)$ is all we need to know about the system and its past history in order to describe its evolution at $k+1, k+2, \dots$. Many models are linear or can be approximated as linear. The representation of uncertainty as additive white noise is a modeling assumption of proven usefulness. When the assumptions of serial independence in $\mathbf{w}$ or $\mathbf{v}$ and independence between $\mathbf{w}$ and $\mathbf{v}$ are violated, there are ways to modify the model to handle these issues (for example, Grewal and Andrews 2015, p. 200).

Kalman Filter Algorithm

The Kalman filter algorithm is repeated at every step, which is defined as the transition from stage k to $k+1$. We may omit the index k or $k+1$ for Φ , $\mathbf{u}$, $\mathbf{w}$, $\mathbf{Q}$, $\mathbf{H}$, $\mathbf{v}$, and $\mathbf{R}$. (In other words, the same recursion formulas will be used, even though these quantities may be different each time.)

The state of the system at time k given observations up to time k has mean $\hat{\mathbf{s}}(k|k)$ and covariance $\Sigma(k|k)$. These two quantities summarize all information from the initial conditions and measurements up to time t .

The objective is to obtain the state of the system at time $k+1$ given observations up to time $k+1$, described through $\hat{\mathbf{s}}(k+1|k+1)$ and $\Sigma(k+1|k+1)$. This is accomplished in two steps: prediction (also known as forecasting or extrapolation) and updating (also known as correction or conditioning).

Prediction

The one-step prediction problem is to find the mean and covariance of $\mathbf{s}(k+1)$ given information up to time k and utilizing the state transition equation Eq. (29).

$$\hat{\mathbf{s}}(k+1|k) = \Phi \hat{\mathbf{s}}(k|k) + \mathbf{u} \quad (38)$$

$$\Sigma(k+1|k) = \Phi \Sigma(k|k) \Phi^T + \mathbf{Q} \quad (39)$$

Updating

Now, we must combine two sources of information regarding $\mathbf{s}(k+1)$: Information up to k , which is what we got at the prediction step, and new information, from data $\mathbf{y}(k+1)$. This is the updating step. We already know the solution to this problem (see preceding section Inverse Problems). The most intuitive form of the equations for the correction step is the *linear feedback* or *linear correction* form, given next:

$$\begin{aligned} \hat{\mathbf{s}}(k+1|k+1) &= \hat{\mathbf{s}}(k+1|k) \\ &\quad + \mathbf{K}(\mathbf{y}(k+1) - \mathbf{H}\hat{\mathbf{s}}(k+1|k)) \end{aligned} \quad (40)$$

where

$$\mathbf{K} = \Sigma(k+1|k) \mathbf{H}^T [\mathbf{H} \Sigma(k+1|k) \mathbf{H}^T + \mathbf{R}]^{-1} \quad (41)$$

The covariance is

$$\Sigma(k+1|k+1) = \Sigma(k+1|k) - \Sigma(k+1|k) \mathbf{H}^T \mathbf{K}^T \quad (42)$$

The updated covariance is the prediction covariance minus the reduction through the measurements.

The linear feedback coefficient $\mathbf{K}$ is known as the Kalman gain. The updated estimate is the predicted mean plus a correction that is linear in the feedback. When the feedback is zero, there is no need for correction. Otherwise, the correction is weighted by the Kalman gain.

Kalman Filter with Unknown Inputs

In many applications, the unknown input that drives the dynamic system cannot be well captured through a white noise process, as when the input is sometimes near zero and sometimes quite pronounced. We add the term $\mathbf{G}_k \boldsymbol{\theta}_k$, where $\mathbf{G}_k$ is an $m \times p$ matrix of known coefficients and $\boldsymbol{\theta}_k$ is a $p \times 1$ vector of unknown coefficients (Kalman filter with unknown inputs).

$$\mathbf{s}(k+1) = \Phi_k \mathbf{s}(k) + \mathbf{u}_k + \mathbf{G}_k \boldsymbol{\theta}_k + \mathbf{w}_k \quad (43)$$

The solution is obtained using the familiar BLUE methodology (Kitanidis 1987).

Successive Linearization

BLUE methods can be adjusted to yield approximate solutions to nonlinear problems. For example, consider the inverse problem with a nonlinear forward map

$$\mathbf{y} = \mathbf{h}(\mathbf{s}) + \mathbf{v} \quad (44)$$

when $\mathbf{h}$ is the n -dimensional vector function of an m -dimensional vector. A common approach is the method of successive linearization. At each step of the iterative procedure we start with an initial estimate of $\tilde{\mathbf{s}}$ and we seek a better estimate $\hat{\mathbf{s}}$. Let the derivative of $\mathbf{h}$ with respect to $\mathbf{s}$ at $\tilde{\mathbf{s}}$ be $\tilde{\mathbf{H}}$:

$$\tilde{\mathbf{H}} = \left. \frac{\partial \mathbf{h}}{\partial \mathbf{s}} \right|_{\mathbf{s}=\tilde{\mathbf{s}}} \quad (45)$$

The n by m matrix $\tilde{\mathbf{H}}$ is known as the Jacobian matrix. Assuming that $\mathbf{s}$ is close to $\tilde{\mathbf{s}}$, approximate,

$$\mathbf{h}(\mathbf{s}) = \mathbf{h}(\tilde{\mathbf{s}}) + \tilde{\mathbf{H}}(\mathbf{s} - \tilde{\mathbf{s}}) \quad (46)$$

Then, the observation equation can be rewritten as

$$\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{h}(\tilde{\mathbf{s}}) + \tilde{\mathbf{H}}\tilde{\mathbf{s}} = \tilde{\mathbf{H}}\mathbf{s} + \mathbf{v} \quad (47)$$

This is the familiar linear observation equation except that the matrix of observation matrix and also the data vector depend on the most recent estimate $\tilde{\mathbf{s}}$. So, we can use the familiar BLUE equations by using:

$$\begin{array}{ll} \tilde{\mathbf{H}} & \text{instead of } \mathbf{H} \\ \tilde{\mathbf{y}} & \text{instead of } \mathbf{y} \end{array} \quad (48)$$

This approach, which we call the *basic iteration*, is simple enough:

1. Start with a reasonable initial estimate $\tilde{\mathbf{s}}$.
2. Compute $\tilde{\mathbf{H}}$ and $\tilde{\mathbf{y}}$.
3. Use the equations for the solution of the linear problem to obtain $\tilde{\mathbf{s}}$.
4. Update $\tilde{\mathbf{s}} = \hat{\mathbf{s}}$ and return to step 1 until convergence is achieved.

Variable Transformation

A related issue is that of variable transformation. Suppose the unknown is conductivity, such as the hydraulic conductivity of a porous medium, or the electric conductivity of a fluid or a solid. This quantity cannot be negative and in some cases it makes no sense to assume it could be zero. However, the BLUE approach does not make a provision for this constraint. One way to circumvent this problem by applying the BLUE methods to a nonlinear transformation of the variable, like the logarithm of conductivity that can be positive and negative (e.g., Tarantola 2005).

Algorithms for Large Problems

Requirements for higher-resolution solutions and trends in data collection technologies lead to problems with increasingly larger numbers of unknowns and observations. The storage and computational cost of implementing the standard or “textbook” methods increases at least quadratically with m . Thus, despite the rapid improvement in computational technology, they become practically impossible to apply for problems of, say $m \sim 10^6$.

However, special technologies have been developed for large problems (reviewed in Ghorbanidehno et al. 2020). Some of these methods use low-rank approximations of covariance matrices, which are easy to store and use in computations, and matrix-free methods, which means that the Jacobian matrix $\mathbf{H}$ is not explicitly computed. One such method for inverse problems is the principal component geostatistical approach (PCGA) (Kitanidis and Lee 2014). The most popular such method is the ensemble Kalman (EnKF), which is described in the next section.

The Ensemble Kalman Filter

The ensemble Kalman filter (EnKF) (Evensen 1994, 2009) is an ingenious in its simplicity implementation of the Kalman filter with computational costs that scale roughly linearly with m , and thus can be used to solve large problems, such as in atmospheric modeling or meteorological forecasting.

The foundation of this approach is that a random vector $\mathbf{x}$ can be represented approximately through an ensemble of N *realizations* or samples from its pdf. For example, the state vector of a dynamic system is given through a matrix with columns the N equiprobable independent realizations from its distribution.

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N] \quad (49)$$

where $\mathbf{x}_i$ is m -dimensional column vector. See article *Ensemble Kalman Filter*.

Relation to Bayesian Methods

Most recent developments in data analysis and estimation are within the realm of Bayesian analysis. The BLUE methodology gives the same results in most cases as the Bayesian approach with Gaussian distributions. Rather than thinking of the two approaches as competitive, one should consider that they provide complementary insights. These results, whether interpreted as BLUE or Bayesian analysis, remain among the most commonly used in practical applications.

Machine learning (ML) methods, (see, for example, Geron 2019) have made great strides in dealing with the same problems where BLUE methods are popularly used. However, such methods typically require much more data to train and implement.

Summary and Conclusions

Best linear unbiased estimation (BLUE) methods are probability-based inference methods that utilize only the first two statistical moments, mean and variance-covariance values, instead of complete probability distributions. In addition to requiring little information, BLUE methods are computationally efficient because they rely on linear (like matrix-vector multiplications) operations. The methodology seeks an estimate that is a linear function of the data that, in an average sense over the ensemble of possible values, has an error that (a) has zero mean and (b) has a squared value that is as small as possible. The methodology has been extended to deal with some nonlinear problems through linearization. New algorithms, like EnKF, allow us to solve large-dimensional problems in a computationally efficient way. Among the many applications of BLUE, it is worth mentioning the following: (1) interpolation methods, like Kriging; (2) inverse problem applications; and (3) dynamic data assimilation methods, using various flavors of Kalman filter. Although more sophisticated methods have been developed, mainly in the Bayesian and machine learning frameworks, they usually require more information and are computationally more expensive to run than the BLUE approach.

References

- Evensen G (1994) Sequential data assimilation with a non-linear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J Geophys Res* 99('C5'):10.143–10.162. <https://doi.org/10.1029/94JC00572>
- Evensen G (2009) *Data assimilation: the ensemble Kalman filter*, 2nd edn. Springer, Berlin
- Gandin LS (1965) *Objective analysis of meteorological fields*. Isr. Program for Sci. Trans., Jerusalem
- Geron A (2019) *Hands-on machine learning with Scikit-learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems*, 2nd edn. O'Reilly, Sebastopol
- Ghorbanidehno H, Kokkinaki A, Lee J, Darve E (2020) Recent developments in fast and scalable inverse modeling and data assimilation methods in hydrology. *J Hydrol* 591(125):266. <https://doi.org/10.1016/j.jhydrol.2020.125266>
- Grewal MS, Andrews AP (2015) *Kalman theory and practice*, 4th edn. Prentice-Hall, Englewood Cliffs
- Kalman RE (1960) A new approach to linear filtering and prediction problems. *J Basic Eng* 82(1):35–45. <https://doi.org/10.1115/1.3662552>
- Kitanidis PK (1987) Unbiased minimum-variance linear state estimation. *Automatica* 23(6):775–778. [https://doi.org/10.1016/0005-1098\(87\)90037-9](https://doi.org/10.1016/0005-1098(87)90037-9)
- Kitanidis PK (1995) Quasilinear geostatistical theory for inversing. *Water Resour Res* 31(10):2411–2419. <https://doi.org/10.1029/95WR01945>
- Kitanidis PK (1997) *Introduction to geostatistics*. Cambridge University Press, Cambridge
- Kitanidis PK, Lee J (2014) Principal component geostatistical approach for large-dimensional inverse problems. *Water Resour Res* 50: 5428–5443. <https://doi.org/10.1002/2013WR014630>
- Matheron G (1963) *Principles of Geostatistics*. Econ Geol 58: 1246–1266
- Tarantola A (2005) *Inverse problem theory and methods for model parameter estimation*. SIAM, Philadelphia

Big Data

Xiaogang Ma

Department of Computer Science, University of Idaho,
Moscow, ID, USA

Definition

It is hard to use a single definition to cover all the characteristics of big data. A general understanding is that big data are a field dealing with data sets that are too large or complex for traditional platforms and tools to handle. The term “big data” first appeared in the mid-1990s. Later, Laney (2001) used the three Vs to describe the dimensions of challenges in data management: Volume, Variety, and Velocity. The three Vs have been well received and have become a common framework to describe the characteristics big data. Other V terms, such as Value, Veracity, and Variability have also been used in various publications to extend the discussion on big data. In the original three Vs, Volume refers to the size of data. So far, big data are reported in terabytes or petabytes, and exabytes and zettabytes are already on the horizon. Variety refers to the heterogeneous structures and types in data. Big data often include structured, semi-structured, and unstructured formats, especially the latter two. Velocity refers to the rate at which data are generated and processed. The advancement in hardware and platform has led to real-time or near real-time generation of massive data sets, such as those on social media. In the other three Vs, Value means that a high value can be obtained from big data though the value density is low in the raw data. Veracity refers to the (in)reliability in the source and quality of data. Variability refers to changing rate or other characteristics in the data flow. In recent years, a topic under discussion among data scientists is the transformation from “big data” to “smart data.” In comparison, in the community of geomathematics and geoinformatics researchers

have been working on both data integration and intelligent analysis. Such discussion and progress reflect people's desire for better platforms and tools to support their work with big data. Most recently, the FAIR principles (Findable, Accessible, Interoperable, Reusable) have set up a roadmap for refining the capabilities of data sharing and reuse in the cyberinfrastructure. Data are not isolated from other objects in research activities. To efficiently implement FAIR principles, we also need to consider many other entities, agents, and activities in a scientific process, such as research activities, samples, software packages, data analysis methods, publications, organizations, researchers, research questions, and more. With clear identification of those objects and meaningful connection among them, we will be able to develop functions to make smart recommendations on datasets to retrieve, hypotheses to explore, methods to use for data analysis, and researchers to collaborate with. Eventually we envision a so-called smart data ecosystem to leverage big data and facilitate new discoveries in geosciences.

Introduction

The rapid development in facilities for data generation and collection, infrastructure for data storage and transmission, and social and cultural change in data sharing and reuse enable the big data of nowadays. In geoscience, scientists are now able to access massive datasets on the Internet, and many large-scale scientific projects have been leveraging big data, machine learning and high performance computation to make ground-breaking discoveries in geoscience (Gil et al. 2019). We envision that global open data and open science endeavors (Nosek et al. 2015) will result in an emergent smart data ecosystem: a complex network that includes data, researchers, research questions, publications, software, instruments, institutions, data repositories, web-based search engine results, and more. The interconnection and interaction of people, data, scientific hypotheses and other facilities will finally lead to intercreation in geoscience in the coming decades. Yet, the machine-readability of key components in the data ecosystem remains a very challenging task. There are various objects and relationships in the ecosystem, and many of them lack a common mechanism for definition, metadata encoding, identifier, and interconnection – for example, via an Application-Programming-Interface (API). Taking data as an example, it is only recently that the global science community developed the FAIR principles: findable, accessible, interoperable, and reusable (Wilkinson et al. 2016) though means for each component have been in existence for decades on the Internet. The geoscience community has taken pioneering steps in developing implementation plans and best practices of FAIR (Stall et al. 2019).

Besides data, there are also standard digital representation of other objects, such as identifiers and metadata schemas for software packages, researchers (e.g., ORCID), organizations, samples, and research activities. Nevertheless, the interconnections among those objects are still limited, and the data ecosystem is still far from mature enough to provide efficient support to research activities. Many researchers are still spending about 80% of their time on data preparation before the actual analysis and knowledge discovery can be conducted (i.e., the 80/20 rule). To achieve a smart data ecosystem for geoscience, there are still many developments to be explored.

Efficient and precise data recommendation and discovery will bring significant improvements to geoscience research with big data. In the past decades, several global geoscience data infrastructures have been established through various communities of practice. For example, data repositories of many geoscience disciplines have been built in the USA through the support from NSF and other sources. Yet, facing such a flourishing data environment, researchers still often struggle to find datasets appropriate for their research questions or interests. A data infrastructure does not constitute a list of isolated data repositories. Instead, tools are required to make interconnections among data sources to focus on research questions and, more importantly, to make context-aware recommendations of data to researchers, research questions, or research programs to help data discovery. The core of such data recommendations is the similarity between a dataset and other objects such as research questions, software packages, and researchers. Algorithms for similarity computation and recommendation are usually developed by using records from the semantic representation of data and various other objects, i.e., a machine-readable encoding of what these objects “mean” or are defined as. Any enriched metadata and relationship descriptions in a semantic representation will help reduce the uncertainty of input data for those algorithms. With minor adaptation, similarity algorithms will also apply to other objects, such as recommending software packages to a dataset or a research question, recommending a collaborator to a researcher, or recommending a research proposal to a grant program.

Moreover, data-intensive geoscience of nowadays is often cross-disciplinary, and needs teamwork to assemble experts with different knowledge backgrounds. This calls for theories and approaches to efficiently unite expertise in computer science, mathematics, statistics, information science, and geoscience, to discover and interpret interesting patterns in the data. We want to call for communities of practice to shift the time that geoscience researchers spend on data preparation into innovative work in scientific discovery. The goal is to transform the 80/20 rule in data-intensive geoscience research into 50/50 or even 20/80. This entry does not provide detailed solutions to certain small questions in big data. Instead, it

intends to draw directions from the broad geomathematics and geoinformatics community about the characteristics of big data and smart data, leading to a call to action, with the hope of seeing increasing improvements, extensions, and discussions of progress in the near future.

Infrastructure and Characteristics of Big Data

Big data are the outcome of many infrastructure and technological developments in the past decades (Kitchin 2014). The rise of computers in the 1950s brought digitization to data. The appearance of the Internet in the 1970s enabled computers to be linked together to transmit data. The widespread use of personal computers in the 1980s and 1990s transformed the mode of how geoscientists do their job. The fast development of the Internet and the Web in the 1990s and 2000s enabled a lot of data portals, especially those built by governmental agencies. The Internet of Things, linked open data, cloud computing, and machine learning in the 2010s quickly increased the rate and scale of data generation and processing. In comparison, those developments in infrastructure and technologies have also facilitated the evolution of mathematical geosciences. Merriam (2004) summarized the six stages of geomathematics: Origins (1650–1833), Formative (1833–1895), Exploration (1895–1941), Development (1941–1958), Automated (1958–1982), and Integration (1982–). The three stages since the 1940s were absolutely characterized by the application of computers. In particular, we are able to see the quick development of spatial statistics, simulation and modeling, and large dataset processing to tackle all the steps in a geoscience workflow. Moreover, geoscience data nowadays are not only being integrated but also being processed with many intelligent technologies.

The rapid development in infrastructure and technology enables big data and bring new opportunities to various organizations and individuals. For example, NASA is one of the biggest generators of geoscience data across the world. An article published in 2019 reported that NASA was generating 12.1 TB of data every day in 2016 (Shannon 2019). Those data were from nearly 100 missions and thousands of sensors, instruments, and platforms around the Earth and space. The article also anticipated that some of NASA missions will soon be able to generate 24 TB of data every day. Besides its long-term data storage programs such as the Earth Observing System Data and Information System (EOSDIS) and the Distributed Active Archive Centers (DAACs), NASA is also collaborating with commercial organizations such as Amazon and Google to use external cloud platforms for data storage. To efficiently handle and process the big data, NASA has put a lot of efforts on machine learning algorithms and artificial intelligence solutions, with the aim to realize seamless connections between data generation, transmission,

cleansing, processing, and decision-making. For individuals, with the thriving approaches of open data, crowdsourcing, and citizen science, anyone can be a data contributor and also a data user. Governmental agencies such as NASA and USGS are making massive geoscience data open and free to use. Scientists are able to analyze real-time data stream transmitted back from a vessel carrying out investigation in the ocean. A geologist in the field can take a photo of an outcrop with a smartphone and then use an app to help identify the rock and mineral types. Publishers are establishing new mechanisms to allow or even require authors to publish their data with papers. A student can paste a question on social media, then many people will suggest publications, datasets, or software packages for him.

The usage of the infrastructure, including Internet, data, algorithms, and computing facilities is being leveraged to a new level. They are not only capable of representing what is known, but can also uncover the origin and reason and help generate ideas for exploring new findings. The skill-set of a geoscientist in the era of big data should include not only the expertise from traditional geoscience disciplines but also many new theories and methods from other disciplines. Essentially, geoscientists who plan to conduct data-intensive research should be aware of the characteristics of big data. Besides the six Vs (Volume, Variety, Velocity, Value, Veracity, and Variability) mentioned above, a few other characteristics were discussed by Kitchin (2014), including exhaustivity, resolution, indexicality, relationality, and flexibility. The “exhaustivity” means that big data projects tend to capture datasets for the whole population or much larger sampling sizes than traditional small data studies. For example, a very recent report on NY Time (Koettl et al. 2020) used satellite images, videos taken by first responders, social media posts, and public radio reports to analyze the patterns of the Almeda fire at the Rogue Valley in southern Oregon and help discover the reasons for the huge destruction caused by the fire. The slogans in big data community, such as “more is better” or “when you can, keep everything” perhaps are driven by the belief that larger data population will increase the chance to find new insights. Nevertheless, the extremely large data also require special methods and tools for storage and processing. The “resolution” means datasets of fine graininess. A typical example is remote sensing images. The “indexicality” means that datasets can have unique and persistent labels and identifiers. For example, Digital Object Identifiers (DOIs) are increasingly used in data sharing and reuse. The “relationality” means that different datasets can be conjoined to answer new questions. The intention is that the sum of data is better than the parts. An interesting object that might appear in the future is “data study” or “data review,” which is similar to the review article in journal publications. The “flexibility” means that big data systems are flexible in

structure and volume. New attributes can be added easily, and the scale of data collection can expand quickly.

Open Data to Change the Mode of Geoscience

Openness, provenance, and reproducibility are readily recognized as features of science nowadays (Nosek et al. 2015; Gil et al. 2019). Increasingly, scientists share data as a part of their outputs, and they also use data shared by others to advance their own research (Kitchin 2014; Boulton et al. 2015). Open science and open data endeavors not only promote communication and reproducibility of scientific outputs, but also strengthen public trust and inform effective decision- and policy-making (Tilmes et al. 2013). In a provenance-aware cyberinfrastructure, if we treat a dataset as an object, then it is related to other objects such as: data creator, scientific subjects, data type, software packages, source and derived publications, data repositories, research questions, samples, instruments, research programs, institutions, and more. The technological framework of the Semantic Web (Berners-Lee et al. 2001) provides the necessary capabilities for categories, identifiers, descriptions, and interconnections among the above-mentioned objects. Thus, Semantic Web is an appropriate environment to capture provenance of open data (Groth and Moreau 2013). To facilitate a virtuous circle of data sharing, discovery, reuse, publication and sharing in the data ecosystem, good quality data, metadata, and provenance must all be published. As Gil et al. (2019) pointed out, provenance and reproducibility will be an essential feature of geoscience contributions in the future. A better semantic representation for objects and their interrelationships in open data and open science will help lead to many innovative studies: enriched metadata of datasets for web crawlers, evaluation and ranking of data FAIRness, and provenance tracing of shared data, just to name a few. This will in turn lay the ground for other innovative work, such as evaluating the long tail of data reuse, reproducibility of scientific workflows by tracing connected objects, and scalability of machine-readable networks.

The importance of open data for scientific innovation has not only been addressed by funding agencies but also by the broader international scientific community. Researchers themselves have been working on principles, guidelines, and implementation plans to facilitate better data management and stewardship using cyberinfrastructure. A very recent work is the FAIR data principles (Wilkinson et al. 2016). In FAIR, Findability means data can be found by common search tools. Accessibility means both metadata and data are able to be retrieved and examined. Interoperability means comparable data will be used together through common vocabulary and formats. Reusability means that data are able to be used by others as a result of good metadata,

provenance, and licenses. The FAIR principles are becoming a growing consensus among different disciplines to improve the quality of open data, including geoscience (Stall et al. 2019). Given the massive amounts of data, it is hard to implement the FAIR principles manually. Tools and platforms are needed to help researchers to check different parts of the FAIRness of their data, and complement the missing or incomplete information to meet the FAIR principles. The enriched data will also be easier for users to discover, access, and reuse. To translate FAIR principles into actionable steps, a system of metrics is needed to represent the detailed parts in those principles, and then to be implemented in the automated tools and platforms mentioned above. Thus, the FAIRness of data will be measurable through the metrics embedded in the tools and platforms.

In this entry, the many objects accompanying open data and big data are called an ecosystem because they form a socio-technical system (Berman et al. 2018). Besides the many technological elements and barriers discussed above, there are also considerable social, ethical, and legal topics related to data. A data creator and/or owner needs to consider their rights and licenses when publishing data. A data user needs to know the license before using other people's data. Moreover, appropriate technologies are needed to accredit data sharing. The academic community needs to change its culture and standards to take data management and sharing as part of research performance measures. The government needs to develop suitable guidelines to encourage best practices in data stewardship and prevent harmful consequences. In today's cyberinfrastructure, researchers are able to share not only data but also many other resources, such as software code, methods and techniques, and workflows. Those objects, their attributes, and their relationships form the technological foundation of the data ecosystem (Berman et al. 2018; Noy et al. 2019). One part of this ecosystem is a way to track the provenance of objects and another is a design of appropriate incentives. These in turn imply the design of appropriate metrics to be used as resources in articulating how incentives lead to value. Therefore, what needs to be considered is the design in the larger context of the science enterprise, that is, the interconnection and interaction of factors and objects in a data ecosystem.

A Smart Data Ecosystem for Geoscience: A Vision for the Future

In recent years, the geoscience community has made noteworthy achievements in regard to data facility building blocks, infrastructure for interoperability, and informatics applications (Nativi and Fox 2010; Yang et al. 2019). Geoscientists are currently taking a leadership role to turn the FAIR principles of open data into practice (Stall et al. 2019).

Funding agencies are increasing their support of data-intensive research; an example of which is one of NSF's ten big ideas on Harnessing the Data Revolution (HDR). Various efforts on best practices are evident in communities such as Research Data Alliance (RDA), NSF-EarthCube, AGU-Earth and Space Science Informatics section (ESSI), GSA-Geoinformatics division, and Earth Science Information Partners (ESIP). For instance, the EarthCube and ESIP communities have conducted fruitful applications of schema.org for geoscience data discovery (Lingerfelt et al. 2017), which was even earlier than the release of the Google Dataset Search (Noy et al. 2019). Such works and discussions demonstrate geoscientists' desire to transform big data into smart data and thus gain better support of data for their research.

To create a smart data ecosystem and transform the 80/20 rule, better machine-readable metadata and more automation are needed in the current cyberinfrastructure. Structured and enriched semantic representation is needed for both data and other objects in the data ecosystem. Work on FAIR metrics however is still in an early stage. Many FAIRness pilot systems applied a questionnaire approach. A user needs to answer questions step by step in order to receive an over-all evaluation result. In Wilkinson et al. (2017), part of the metrics design is machine-readability and protocol for metadata access, but there is no pilot portal for the design yet. Machine-readable FAIR metrics may be the key to increased automation in the data ecosystem. Those metrics are derived from the detailed items listed in each of the four FAIR principles. For example, the metrics need to at least include the scheme of unique and persistent identifiers for data, the metadata associated with each data identifier, the protocol for data and metadata access, ontologies, vocabularies and other community standards used in the data, provenance about the origin of the data, and licenses for data usage, just to name a few. Based on those metrics, tools and platforms will be built to set up a pipeline that helps users examine and enrich the standardization, structure, and description of their data and metadata. The process of data and metadata sharing is bottom-up and needs contribution from various individuals and organizations, and a framework for such activities is the essential foundation. Semantic representation, machine-readable metrics, and interconnections among objects are able to represent a framework of structured information. Even light and simple semantics in the framework will increase the machine-readability a lot, such as the structured data markup enabled by schema.org for web pages (Noy et al. 2019). With clear identification of objects in the data ecosystem and meaningful connections among them, functions can be developed to make smart recommendations on datasets to retrieve, hypotheses to explore, functions to use for data analysis, and researchers to collaborate with.

Eventually, a smart data ecosystem will implement the FAIR principles, leverage various open data resources, and

facilitate new discoveries in geoscience. Data-intensive geoscience research of nowadays has several characteristics: research question-focused, cross-disciplinary, and teamwork with complementary expertise (Hazen et al. 2019; Shipley and Tikoff 2019). Geoscience research also benefits from the recent discussion and development on theories and methods of data science. The above narrative identifies the following topics of interest for future methodological and technological developments towards a smart data ecosystem for geoscience: (1) machine-readable FAIR metrics for open data; (2) representation of objects and relationships in the data ecosystem for provenance tracing; (3) intelligent recommendation and suggestion of data and other resources in open science; and (4) data science methods in cross-disciplinary geoscience research.

Technological Approaches Toward a Smart Data Ecosystem

The following four subsections will introduce more details of a few technological topics that can be explored in the future. It is noteworthy that they are just part of the many research topics in big data. They have a focus on data sharing, management, and reuse. Many other topics in machine learning and data mining are not covered here.

Machine-Readable FAIR Metrics of Geoscience Data

As mentioned above, to put FAIR data principles into practice in geoscience, machine-readable FAIR metrics data must be developed. There are community initiatives focusing on different aspects of those principles, such as data identifiers, data types, domain-specific vocabularies, data citation, and a data seal of approval for repositories. Research is needed to leverage the existing building blocks to automate a set of metrics to represent the detailed elements in the FAIR principles, and incorporate those technological components and interfaces into new tools. Those tools will enable individual researchers to evaluate and improve the FAIRness of their data before sharing. The tools will also be used by data facility managers to check the FAIRness of published data and make updates.

Semantic representations for FAIR principles, such as ontologies, must be further developed. They will define/inform key elements in the machine-readable FAIR metrics. The knowledge engineering work for those semantic representations will need to incorporate community standards such as the Provenance Ontology (PROV-O), schema.org/Dataset, and the Dataset Catalog recommendation (DCAT) as well as the existing experimental FAIR ontology studies. To employ those semantic representations (e.g., ontologies) in tools for FAIR metrics evaluation, the ideal situation is an automated or semiautomated process. In such a process, the tools will access existing services to generate results for the metrics,

without or with minor user control. To achieve such automation, a number of existing services and building blocks for open data should be reflected in the FAIR metrics, and interfaces and interactions to them should be established in the tools. Those services and building blocks include DataCite, Data Type Registry, Linked Open Vocabulary Service FAIR sharing standard and vocabulary registry, Data Citation Index, and dataset attributes specified in schema.org/Dataset, to name a few. Future developments in cyberinfrastructure will bring more building blocks for all aspects of the FAIR principles and automate the FAIR metrics evaluation process.

Semantic Representation of Scientific Workflows for Provenance Documentation

The current technological environment for a data ecosystem is the Web, wherein data are but only one of the many objects that are involved in a scientific research workflow. On the Web, researchers will follow linked identifiers to such objects and their corresponding metadata, which in turn often document the provenance of scientific workflows. Increasingly, metadata can be retrieved through object identifiers (often Uniform Resource Locators (URLs)), such as metadata of publications and data. A workflow includes various types of objects, and the metadata of those objects provide sources for provenance documentation. In the past decade, researchers have made progress on identifiers of many types of objects in the data ecosystem. Besides DOIs for publications, recent initiatives include ORCID for researchers, DataCite and DOIs for data, IGSN for geo-samples, OrgID for organizations, RAiD for research activities, and more. Most of those identifiers provide an API service for accessing associated metadata. Those metadata contain contextual information that can be used in subsequent data analysis. For example, Bohannon (2017) showed how to use ORCID metadata to find migratory patterns of scientists over the course of their careers. There have also been experimental studies trying to make connections among those identifiers, such as the ODIN work between ORCID and DataCite, which was able to identify and connect scientists and datasets across multiple services.

Previous successful work on provenance documentation of workflows stands ready to be extended. In the Global Change Information System (GCIS), various objects in climate assessment and reporting were recognized as classes and properties in a new mixture of science and provenance terminology: the GCIS ontology (Tilmes et al. 2013). The collected provenance in GCIS provided structured inputs for the traceability and reproducibility of scientific workflows (Zeng et al. 2019). URLs were used as object identifiers in GCIS. For example, a URL (<http://data.globalchange.gov/article/10.1080/01490419.2010.491031>) was used to represent an article. It points to a web page with detailed metadata about the article, and also the connections between this article and other objects, such as where it was referenced, the source data used

in the article, the figures derived from this article, and more. In the Deep Carbon Observatory (DCO) – Data Science program, the Global Handle System was used to assign a unique and resolvable identifier (i.e., a handle) to each object in the DCO data portal (Ma et al. 2017). The object types were further defined in the DCO ontology to give specific machine-readable meanings of those objects. Our experience of knowledge engineering and semantic representation shows that it is also beneficial to refer to the existing standard ontologies and community best practices to avoid redundant work.

Context-Aware Data and Resource Recommendation for Geoscience

To reduce the time that researchers spend on data search and manipulation, and change the 80/20 rule, one approach is to make data recommendations that match researcher needs by using the idea of scientific similarity. Existing algorithms on similarity computations must be improved to make finer recommendations. Similarity computations have been a continuous research topic in the Semantic Web community, and have also received attention in geoscience (Zheng et al. 2015). Similarity will be computed between data and a researcher, data and a research question, data and methods, and more. Results of such similarity computation will provide suggestions for data recommendations. For example, if the similarity between a dataset and researcher is very high, then it is verily likely that the researcher can make use of the data in his work, so the dataset can be recommended to the researcher. Besides data, similarity computations have been conducted for various other objects, such as software programs, research topics, collaborators, and grant programs. In a data ecosystem, recommendations of those resources and objects can also be made based on the result of similarity computation.

The results of similarity computations are able to be used in many ways in a data ecosystem. For instance, if the similarity between a dataset and a researcher is computed, the resulting similarity value represents to which extent the dataset matches the researcher. The similarity value is a key input for making a data recommendation or not. However, similarity alone is just one of many components to be considered in making recommendations. For example, to make finer recommendations of data to researchers, collaborative filtering (Herlocker et al. 2004) is a way to be used. Collaborative filtering considers users' contributions such as ratings or reviews about items in recommendation making. It is based on the theory that similar users have similar patterns of rating behavior, and similar items obtain similar ratings. Collaborative filtering takes a matrix of existing user-item ratings as the main source of input, such as the data citation records. In a data ecosystem, we anticipate many kinds of user inputs and feedback will be obtained and used in collaborative filtering to improve the precision of resource recommendations.

Data Science Approaches for Geoscience in the Era of Big Data

The theoretical foundation of data science is under fast development, and will benefit from many existing disciplines such as computer science, statistics, mathematics, information science, and domain-specific applications (Fox and Hendler 2014). Geoscience has experienced many challenges and opportunities over the last 20 years and has made changes in their infrastructure in response to the emergent data ecosystem. However, nascent advances in international communities, spurred by RDA, CODATA, EUDAT, NSF-EarthCube, and other communities offer an inflection point for greater capabilities (such as those mentioned in previous sections). Data-driven geoscience often needs multidisciplinary collaboration and teamwork of researchers from different knowledge backgrounds. This feature will potentially make science of team science necessary in geoscience. The practices of team science in geoscience will in turn contribute to the theoretical foundation of data science. The data ecosystem is underpinned by abundant data sources, and by computational facilities, such as Giovanni, IRIS, HydroShare, and others. Progress of geoscience in the coming decade needs renewed attention to both data integration and intelligent analysis.

Exploratory data analysis will be an effective way to tackle big data in geoscience. Similar to many other disciplines, geoscience conventionally applies the hypothesis-driven research method in the research process. In a data ecosystem, geoscience researchers may often face such a situation: while plenty of data are already collected, there is no research hypothesis. The method of exploratory data analysis, which has been proven functional in the domain of statistics (Tukey 1977), may offer clues for hypothesis generation in geoscience. The term “exploratory” explains the purpose of the method: it is flexible and will help look for things that we believe are not there or to be there (Tukey 1977). Exploratory data analysis is comparable to abduction in an iterative process of data-driven discovery (Hazen et al. 2019). Here abduction means forming a plausible explanation for an observation before deduction and induction are taken.

Science of team science can be applied to leverage the multidisciplinary expertise in big geoscience projects. Science of team science studies focus on collaborative and team-based research and elaborate strategies and methods to enhance the effectiveness of science process and outcomes (Grudin 1994). In a recent report released by the National Research Council, recommendations have been made on how to enhance the effectiveness of team science (NRC 2015). One recommendation is about how to use research networking tools in team assembly and task analysis. A key topic in science of team science is to obtain a better understanding of the multi-network structure of scientific collaboration. A network is a structure made up of a set of objects (i.e., nodes), the attributes of the objects, and the

interrelationships (i.e., edges) between those objects. In the data ecosystem, the various types of objects and relationships form complex networks, and methods are needed to gain insights into them.

Summary and Conclusions

Geoscience, like many other disciplines, faces the challenges and opportunities of big data. A smart data ecosystem will significantly shift the time that researchers spend away from data cleansing and preparation into data analysis and scientific discovery. Although cyberinfrastructure development in the past decades has provided a new and powerful data ecosystem, the full potential of the ecosystem is yet to be explored, and many extensions still need to be developed to make such an ecosystem smarter. This entry provides a brief review of the concepts in big data, open data, and open science and offers a perspective on future work that should be explored, such as machine-readable and automated FAIR metrics of open data, semantic representation of key components in the data ecosystem, algorithms for similarity computation and resource recommendation, and network analysis to facilitate the science of team science. Many existing platforms, services, and building blocks are able to be leveraged, such as those in DataCite, Data Type Registry, Linked Open Vocabulary Service, FAIR sharing standard and vocabulary registry, and Data Citation Index. Such extensions and developments inherently call for participation from communities of practice including many stakeholders. A new capacity for data-intensive geoscience, once established, is able to be transferred to other disciplines with minor adaptation, and will serve as a functional foundation for many new research paths to explore big data.

Cross-References

- [Data Life Cycle](#)
- [Data Mining](#)
- [Database Management System](#)
- [Exploratory Data Analysis](#)
- [Spatial Data](#)

Bibliography

- Berman F, Rutenbar R, Hailpern B, Christensen H, Davidson S, Estrin D, Franklin M, Martonosi M, Raghavan P, Stodden V, Szalay AS (2018) Realizing the potential of data science. *Commun ACM* 61(4):67–72
- Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. *Sci Am* 284(5):34–43

- Bohannon J (2017) Vast set of public CVs reveals the world's most migratory scientists. *Sci News*. <https://doi.org/10.1126/science.aal1189>
- Boulton G, Babini D, Hodson S, Li J, Marwala T, Musoke Mg, Uhler PF, Wyatt S (2015) Open data in a big data world: an international accord. International Council for Science (ICSU), International Social Science Council (ISSC), The World Academy of Sciences (TWAS), InterAcademy Partnership (IAP), Paris, p 15. https://icsu.org/cms/2017/04/open-data-in-big-data-world_long.pdf. Accessed 30 May 2018
- Fox P, Hendler J (2014) The science of data science. *Big Data* 2(2):68–70
- Gil Y, Pierce SA, Babaie H, Banerjee A, Borne K, Bust G, Cheatham M, Ebert-Uphoff I, Gomes C, Hill M, Horel J (2019) Intelligent systems for geosciences: an essential research agenda. *Commun ACM* 62(1):76–84
- Groth P, Moreau L (eds) (2013) PROV-overview: an overview of the PROV family of documents. <http://www.w3.org/TR/prov-overview/>. Accessed 2 June 2019
- Grudin J (1994) Computer-supported cooperative work: history and focus. *Computer* 27(5):19–26
- Hazen RM, Downs RT, Eleish A, Fox P, Gagne O, Golden JJ, Grew ES, Hummer DR, Hystad G, Krivovichev SV, Li C, Liu C, Ma X, Morrison SM, Pan F, Pires AJ, Prabhu A, Ralph J, Rumyon SE, Zhong H (2019) Data-driven discovery in mineralogy: recent advances in data resources, analysis, and visualization. *Engineering* 5(3):397–405
- Herlocker JL, Konstan JA, Terveen LG, Riedl JT (2004) Evaluating collaborative filtering recommender systems. *ACM Trans Inf Syst* 22(1):5–53
- Kitchin R (2014) The data revolution: big data, open data, data infrastructures & their consequences. Sage, London, p 222
- Koetli C, Kim C, Jordan D, Ray A (2020) How an Oregon wildfire became one of the most destructive. <https://www.nytimes.com/video/us/100000007341513/oregon-fires-path.html>. Accessed 20 Sept 2020
- Laney D (2001) 3D data management: controlling data volume, velocity and variety. In: Meta group. Available at: <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>. Accessed 17 Sept 2020
- Lingerfelt E, Fils D, Shepherd A (2017) Project 418: using web architecture patterns to address discovery and facilitate approaches to Open and FAIR. AGU 2017 Fall Meeting, New Orleans. <https://www.earthcube.org/sites/default/files/doc-repository/p418agu.pdf>. Accessed 15 Apr 2018
- Ma X, West P, Zednik S, Erickson J, Eleish A, Chen Y, Wang H, Zhong H, Fox P (2017) Weaving a knowledge network for Deep Carbon Science. *Front Earth Sci* 5:36. <https://doi.org/10.3389/feart.2017.00036>
- Merriam D (2004) The quantification of geology: from abacus to Pentium: a chronicle of people, places, and phenomena. *Earth Sci Rev* 67(1–2):55–89
- Nativi S, Fox P (2010) Advocating for the use of informatics in the earth and space sciences. *EOS Trans Am Geophys Union* 91(8):75–76
- Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD, Chin G, Christensen G, Contestabile M, Dafeo A, Eich E, Freese J, Glennerster R, Goroff D, Green DP, Hesse B, Humphreys M, Ishiyama J, Karlan D, Kraut A, Lupia A, Mabry P, Madon T, Malhotra N, Mayo-Wilson E, McNutt M, Miguel E, Paluck EL, Simonsohn U, Soderberg C, Spellman BA, Turitto J, VandenBos G, Vazire S, Wagenmakers EJ, Wilson R, Yarkoni T (2015) Promoting an open research culture. *Science* 348(6242):1422–1425
- Noy N, Burgess M, Brickley D (2019) Google dataset search: building a search engine for datasets in an open web ecosystem. In: Proceedings of the World Wide Web conference, San Francisco, pp 1365–1375. <https://doi.org/10.1145/3308558.3313685>
- NRC (National Research Council) (2015) Enhancing the effectiveness of team science. The National Academies Press, Washington, DC, p 280. <https://doi.org/10.17226/19007>
- Shannon M (2019) How does NASA use big data?. <https://bigdata-madesimple.com/how-does-nasa-use-big-data>. Accessed 20 Sept 2020
- Shipley TF, Tikoff B (2019) Collaboration, cyberinfrastructure, and cognitive science: the role of databases and dataguides in 21st century structural geology. *J Struct Geol* 125:48–54
- Stall S, Yarmey L, Cutcher-Gershenfeld J, Hanson B, Lehnert K, Nosek B, Parsons M, Robinson E, Wyborn L (2019) Make scientific data FAIR. *Nature* 570(7759):27–29
- Tilmes C, Fox P, Ma X, McGuinness D, Privette AP, Smith A, Waple A, Zednik S, Zheng J (2013) Provenance representation for the National Climate Assessment in the Global Change Information System. *IEEE Trans Geosci Remote Sens* 51(11):5160–5168
- Tukey JW (1977) Exploratory data analysis. Addison-Wesley, Reading, p 688
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J (2016) The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 3:160018. <https://doi.org/10.1038/sdata.2016.18>
- Wilkinson MD, Sansone SA, Schultes E, Doorn P, da Silva Santos LOB, Dumontier M (2017) A design framework and exemplar metrics for FAIRness. *Sci Data* 5(1):1–4. <https://doi.org/10.1101/225490>
- Yang C, Yu M, Li Y, Hu F, Jiang Y, Liu Q, Sha D, Xu M, Gu J (2019) Big Earth data analytics: a survey. *Big Earth Data* 3(2):83–107
- Zeng Y, Su Z, Barmpadimos I, Perrels A, Poli P, Boersma Kf, Frey A, Ma X, de Bruin K, Gossen H, Timmermans W (2019) Towards a traceable climate service: assessment of quality and usability of essential climate variables. *Remote Sens* 11(10):1186. <https://doi.org/10.3390/rs11101186>
- Zheng JG, Fu L, Ma X, Fox P (2015) SEM+: tool for discovering concept mapping in Earth science related domain. *Earth Sci Inf* 8(1):95–102

Binary Mathematical Morphology

Aditya Challa¹, Sravan Danda¹ and B. S. Daya Sagar²

¹Computer Science and Information Systems, APPCAIR, Birla Institute of Technology and Science, Pilani, Goa, India

²Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, Karnataka, India

Definition

Binary morphology encompasses all operators from mathematical morphology (MM) on binary images. MM operators are defined on a complete lattice. Binary images constitute a specific lattice. This allows the definition of the MM operators to be simplified. In this entry we shall give an overview of different binary morphological operators. Detailed explanation of these operators can be found in their respective

chapters and the books Najman and Talbot (2013), Soille (2004), and Dougherty and Lotufo (2003).

A binary image can be represented as a set in $\mathbb{R}^2$ or $\mathbb{Z}^2$. Equivalently, one can also represent it as a function, $f: E \rightarrow \{0, 1\}$ where E is either $\mathbb{R}^2$ or $\mathbb{Z}^2$. We follow the set representation in this entry. If the domain is $\mathbb{Z}^2$ it is referred to as discrete binary image.

Basic MM Operators on Binary Images

The following are the basic operators from mathematical morphology adapted to binary images.

1. *Dilation*: Given a binary image X and a structuring element B , the dilation is defined as

$$\delta_B(X) = \bigcup_{b \in B} X_b = \bigcup_{x \in X} B_x \quad (1)$$

where X_b denotes the set X translated by b .

2. *Erosion*: Given a binary image X and a structuring element B the erosion is defined as

$$\varepsilon_B(X) = \bigcap_{b \in \hat{B}} X_b \quad (2)$$

where $\hat{B}$ denotes the reflection of B and X_b denotes the translation of X by b .

3. *Opening*: Given a binary image X and a structuring element B the opening is defined as

$$\gamma_B(X) = \delta_B \circ \varepsilon_B(X) \quad (3)$$

4. *Closing*: Given a binary image X and a structuring element B the closing is defined as

$$\phi_B(X) = \varepsilon_B \circ \delta_B(X) \quad (4)$$

Filtering on Binary Images

A morphological filter on a binary image is any operator that is increasing and idempotent and can be obtained by basic dilation/erosion combinations. The opening/closing operators above form the basic filters. Several families of filters can be constructed from these operators such as:

1. *Granulometries*: The idea of a granulometry is similar to filtering out particle based on their sizes. Using the opening operator we can define it as

$$\text{Granulometry} = \gamma_{nB} \circ \gamma_{(n-1)B} \circ \dots \circ \gamma_{2B} \circ \gamma_B \quad (5)$$

Not all structuring elements B are allowed. Usually only disk and line-segments are considered.

2. *Alternating Sequential Filters*: For binary images where noise is spread across different scales and contains both bright and dark distortions, using alternate sequences of opening and closing would help.

$$\text{ASF} = \phi_{nB} \circ \gamma_{nB} \circ \phi_{(n-1)B} \circ \gamma_{(n-1)B} \circ \dots \circ \phi_B \circ \gamma_B \quad (6)$$

One can also start with closing instead of opening as in the above definition.

On the other hand one can define an *algebraic filter*, which is any operator that is increasing and idempotent. Examples of algebraic filters include Area Opening, Attribute Openings, Annular Opening, Convex Hull Closing, etc.

Other Operators on Binary Images

A host of operators can be defined on binary images. A few are stated as follows:

1. *Geodesic Dilation/Erosion*: The geodesic operations force the output to belong to a mask Y . A single step of geodesic dilation is defined as

$$\text{Geo} - \text{Dilate}^{(1)}(X) = \delta_B(X) \cap Y \quad (7)$$

One can suitably define geodesic erosion as well. These operators can be repeated to obtain higher order dilations or erosions.

2. *Hit or Miss Transform (HMT)*: Defined on discrete binary images, these operators extract certain neighborhood configurations in the binary image.
3. *Thinning/Thickening*: Using the HMT transform above, one can obtain the thinning by removing the identified neighborhoods from the existing image and thickening by adding the identified neighborhoods to the existing image.
4. *Skeletonization* involves reducing the binary image to a one-dimensional caricature to extract the most discerning features in the image.

Summary

In this entry, elementary MM operators on binary images, i.e., binary dilation, binary erosion, binary opening, and binary closing are defined. Using these elementary operators, more

complex operators, namely, granulometries, alternating sequential filters, geodesic dilation/erosion, hit or miss transform, thinning/thickening, and skeletonization are then described and intuitively explained.

Cross-References

- [Mathematical Morphology](#)
- [Morphological Closing](#)
- [Morphological Dilation](#)
- [Morphological Erosion](#)
- [Morphological Filtering](#)
- [Morphological Opening](#)

Bibliography

- Dougherty ER, Lotufo RA (2003) Hands-on morphological image processing, vol 59. SPIE Press, Bellingham
- Najman L, Talbot H (eds) (2013) Mathematical morphology: from theory to applications. John Wiley & Sons, Inc. <https://doi.org/10.1002/9781118600788>
- Soille P (2004) Morphological image analysis. Springer, Berlin/Heidelberg. <https://doi.org/10.1007/978-3-662-05088-0>

Binary Partition Tree

Sravan Danda¹, Aditya Challa² and B. S. Daya Sagar³

¹Computer Science and Information Systems, APPCAIR, Birla Institute of Technology and Science, Pilani, Zuari Nagar, India

²Computer Science and Information Systems, Birla Institute of Technology and Science, Pilani, Zuari Nagar, India

³Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, Karnataka, India

Definition

A binary partition tree (BPT) is a hierarchy, used to represent a set of regions in a grey-scale image. This hierarchy is constructed using the homogeneity between the regions. The leaf nodes represent the basic regions and the other nodes represent the regions obtained by merging pairs of neighboring regions. The root represents the entire image. See Salembier and Garrido (2000) for more details.

Illustration

To illustrate the definition, consider the image on the left of Fig. 1. This image consists of five regions with the grey-scale value as denoted in the brackets. The corresponding binary partition tree is shown on the right.

To construct the binary partition tree, one starts with merging the neighboring regions with highest affinity. In this case, we have regions A and B which have the highest affinity with respect to their grey-scale values. Hence these regions are merged. The value for combined region is taken to be the average of regional values. So, node AB is assumed to have the value $0.5(75 + 60) = 67.5$. The next closest pair is that of (C, D) . Accordingly, node CD is given the value 20. Then we have that node CD is closer to node AB than node E . Hence, nodes (AB, CD) are combined to get node $ABCD$. Finally region E is merged to obtain the region $ABCDE$.

Construction and Processing the BPT

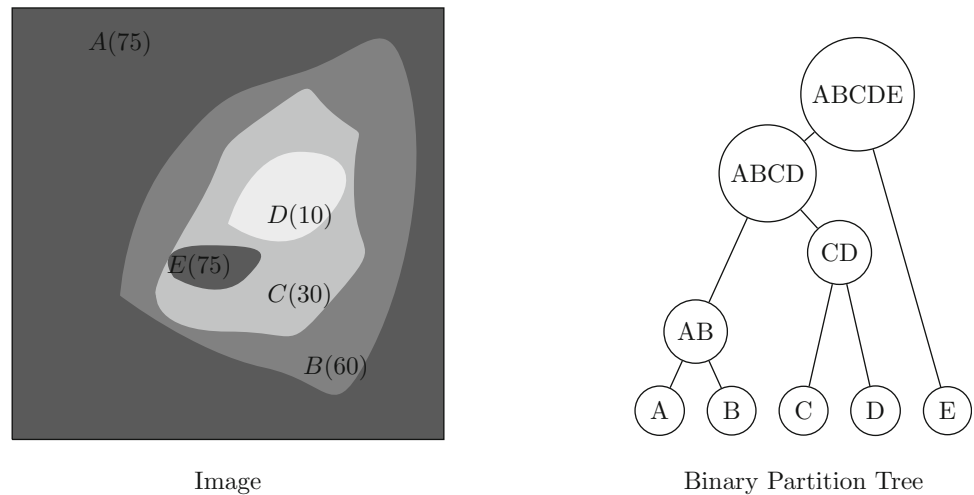
In the simple example illustrated in Fig. 1 we use the average of grey-scale values. In general the construction of binary partition tree requires the input of two functions -

1. Merging Order: A function which indicates how close the two regions are. In the simple case of Fig. 1 we used the difference between grey-scale values. This can be a more complicated function. See Garrido et al. (1998) for some examples and comparisons.
2. Region Model: One also needs the function which dictates the value of regions which are merged. In Fig. 1 we used the average. However, one might consider taking the weighted average or any other intricate function.

After the BPT is constructed, one can process this to obtain efficient partitions. For instance one can use metrics such as PSNR (peak signal to noise ratio) to cut the hierarchy and obtain suitable partitions. Further, BPT also encodes the spatial relationships between existing objects.

BPT in Edge-Weighted Graphs

BPT can also be defined on edge-weighted graphs. Recall that an edge-weighted graph G is a tuple (V, E, W) where V denotes the vertices, E denotes the edges which is a subset of $V \times V$ and W denotes the edge-weight assigned to each edge in E . Then using the minimum spanning tree construction, one can construct the BPT with altitude as described in Najman et al. (2013). A fast library for implementing BPT on edge-weighted graphs can be found in Perret et al. (2019).

Binary Partition Tree,**Fig. 1** Figure illustrating the binary partition tree**Summary**

In this chapter, a binary partition tree is described as a tool used for hierarchical segmentation of grey-scale images. It is illustrated with the help of a toy example as to how it works. Two types of constructing a BPT are outlined, namely, merging-order based model and region model. Then, a generalization of BPT suited to edge-weighted graphs is briefly mentioned.

Cross-References

► [Mathematical Morphology](#)

Bibliography

- Garrido L, Salembier P, Garcíá D (1998) Extensive operators in partition lattices for image sequence analysis. *Signal Process* 66(2):157–180. [https://doi.org/10.1016/S0165-1684\(98\)00004-8](https://doi.org/10.1016/S0165-1684(98)00004-8)
- Najman L, Cousty J, Perret B (2013) Playing with Kruskal: algorithms for morphological trees in edge-weighted graphs. In: Hendriks CLL, Borgfors G, Strand R (eds) *Mathematical morphology and its applications to signal and image processing, 11th international symposium, ISMM 2013, Uppsala, Sweden, May 27–29, 2013. Proceedings. Lecture notes in computer science, vol 7883*. Springer, Heidelberg, pp 135–146. https://doi.org/10.1007/978-3-642-38294-9_12
- Perret B, Chierchia G, Cousty J, Guimarães SJF, Kenmochi Y, Najman L (2019) Higua: hierarchical graph analysis. *SoftwareX* 10:100335. <https://doi.org/10.1016/j.softx.2019.100335>. <http://www.sciencedirect.com/science/article/pii/S235271101930247X>
- Salembier P, Garrido L (2000) Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval. *IEEE Trans Image Process* 9(4):561–576. <https://doi.org/10.1109/83.841934>

Bonham-Carter, Graeme

Eric Grunsky
Department of Earth and Environmental Sciences, University of Waterloo, Merrickville, ON, Canada

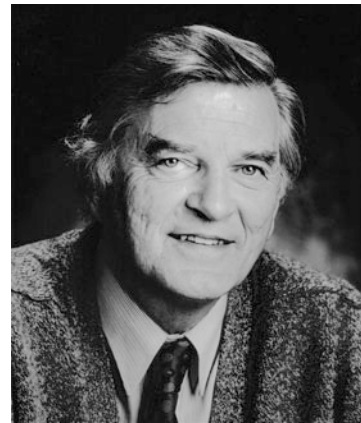


Fig. 1 Graeme Bonham-Carter, courtesy of Dr. Bonham-Carter

Biography

Graeme Bonham-Carter is a mathematical geologist who is known for his research in simulation modeling and geospatial analysis. Graeme was born in England in 1939. He obtained a BA in Natural Sciences and Geology, Cambridge University in 1962, an MA in Geology (1963), and a Ph.D. in Geology (1966) at the University of Toronto followed by a postdoctoral fellowship at Stanford University

from 1966 to 1969. He taught at University of Rochester from 1969 to 1975. He was a research scientist at the Geological Survey of Canada (GSC) from 1980 to 2006. Since retiring he consulted for Canadian Mining Industry Research Organization (CAMIRO), for GSC, and for mineral exploration companies.

While at Stanford, Graeme developed a simulation model of deltas, and published a textbook on computer simulation (Harbaugh and Bonham-Carter 1970). At University of Rochester, he modeled wind-driven circulation of lakes (Bonham-Carter and Thomas 1973). At GSC in Ottawa, he worked on early applications of GIS in geology. With Frits Agterberg he developed the Weights of Evidence (WoE) methodology that was widely adopted for use in mineral prospectivity modeling. In 1994 he published *Geographic Information Systems for Geoscientists: Modelling with GIS* (Bonham-Carter 1994). He led a series of *Mineral Potential Mapping with GIS* short courses that were given first in Canada, then with Gary Raines from USGS in the USA, Australia, and Brazil (Kemp et al. 1999). His interests have included modeling of dispersal of metals around smelters (Bonham-Carter et al. 2006), distribution of metals using selective leaches (Bonham-Carter and Hall 2010), and analysis of portable XRF data (Hall et al. 2014).

Graeme was awarded the 1998 William Christian Krumbein Medal by the IAMG. In 1996 he succeeded Daniel Merriam as Editor-in-Chief of the IAMG journal *Computers & Geosciences (C&G)*, an IAMG journal. In 2008, Graeme co-edited the book *Progress in Geomathematics*, a Festschrift in honor of Frits Agterberg (Bonham-Carter and Cheng 2008). He served as President of the IAMG 2000–2004.

Bibliography

- Bonham-Carter GF (1994) *Geographic information Systems for Geoscientists: modelling with GIS*. Pergamon Press, Oxford. 398 p
- Bonham-Carter GF, Thomas JH (1973) Numerical calculation of steady wind-driven currents in Lake Ontario and the Rochester embayment. In: *Proceedings 16th conference on great lakes research*, Ohio State University, Huron, Ohio, 16–18
- Bonham-Carter GF, Cheng Q (eds) (2008) *Progress in geomathematics*. Springer, Heidelberg. 554 p
- Bonham-Carter GF, Hall GEM (2010) Final report on results of soil geochemistry using selective leaches, multi-media techniques for direct detection of covered unconformity uranium deposits in the Athabasca Basin, project 08E01. CAMIRO Exploration Division. 280 p. <https://www.appliedgeochemists.org/other-publications/camiro-project-08e01>
- Bonham-Carter GF, Henderson PJ, Kliza DA, Kettles IM (2006) Comparison of metal distributions in snow, peat, lakes and humus around a Cu smelter in western Quebec, Canada. *Geochem Explor Environ Anal* 6(2):215–228
- Hall GEM, Bonham-Carter GF, Thomas JH, Buchar AK (2014) Evaluation of portable X-ray fluorescence (pXRF) in exploration and mining: phase 1, control reference materials. *Geochem Explor Environ Anal* 14(2):99–123
- Harbaugh JW, Bonham-Carter GF (1970) *Computer simulation in geology*. Wiley-Interscience, New York. 575 p
- Kemp LD, Bonham-Carter GF, Raines GL (1999) Arc-WofE: Arcview extension for weights of evidence mapping. <http://www.ige.unicamp.br/wofe>

Bootstrap

Oktay Erten and Clayton V. Deutsch
Centre for Computational Geostatistics, University of
Alberta, Edmonton, AB, Canada

Definition

The *bootstrap* is a resampling technique that is used to make inferences about a sampling distribution of a statistic. Its application becomes relevant when there are a few *independent and identically distributed* (iid) observations ($n \leq 25$) which are representative of the *population* for inference. Considering a small *sample* of n observations, the bootstrap randomly samples n observations with replacement, resulting in *resamples* (or bootstrap samples). A statistic of interest is calculated from each resample, and the pooled set of calculated statistics form a *bootstrap distribution*. The *confidence* in the true value of the statistic can be quantified from the bootstrap distribution.

Introduction

In many geoscience applications (i.e., mining, hydrogeology, petroleum, soil science, meteorology, and climatology), there may be only a small sample (tens of observations) available due to the cost of sampling. Spatial modeling of an attribute, conditioned on a few observations, leads to substantial uncertainty associated with the value of the attribute at an unsampled location. Geostatistics provides established techniques to quantify this uncertainty (Pyrz and Deutsch 2014); however, one should also consider that the estimated model parameters (i.e., distribution functions and variograms/covariances) themselves are also uncertain due to the small sample size. Therefore, the uncertainty in model parameters should be quantified and incorporated into attribute modeling.

The traditional bootstrap (Efron and Tibshirani 1993) is a Monte Carlo resampling algorithm used to assess the uncertainty in estimated statistics (i.e., mean, median, and variance) by sampling the empirical data with replacement. In the traditional bootstrap, the sample, deemed representative of the underlying population, is assumed to be composed of iid observations. In other words, the traditional bootstrap does

not consider the spatial correlation between the observations. However, in many geoscience applications, the sampling locations are determined based on a specific sampling strategy, and the attribute has spatial correlation at distances larger than the sample spacing; therefore, the observations cannot be considered independent. Solow (1985) proposed a simulation methodology with LU decomposition of a covariance matrix for obtaining resamples from the spatially correlated observations. A number of studies have also been presented in the literature, which extends the traditional bootstrap procedure to the spatial context (Deutsch 2004; Feyen and Caers 2006; Journel and Bitanov 2004; Olea and Pardo-Igúzquiza 2011). The difference between the traditional and spatial bootstrap is that the latter takes into account the covariance function describing the spatial correlation between the observations.

One of the useful applications of the traditional bootstrap is the quantification of variability in the estimated slope $\hat{\beta}_0$ and intercept $\hat{\beta}_1$ coefficients in a regression model $y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ for $i = 1, \dots, n$. The residuals $e_i = y_i - \hat{y}_i$ are randomly sampled with replacement, and the resample is generated by using the same x_i value, but with y_i^* obtained from $y_i^* = \hat{y}_i + e_i^*$. As for the spatial bootstrap, the popular application is the assessment of uncertainty in the distribution and the experimental variogram of a spatial attribute. Pardo-Igúzquiza and Olea (2012) proposed a computer program through which the spatially correlated observations $\{z(\mathbf{u}_i), i = 1, \dots, n\}$ are sampled with replacement, and the experimental variogram $\{\hat{\gamma}_j^*(\mathbf{h}), j = 1, \dots, m\}$ is computed for each resample $\{z_j^*(\mathbf{u}_i), i = 1, \dots, n; j = 1, \dots, m\}$. The uncertainty in the experimental variogram $\hat{\gamma}(\mathbf{h})$ of the attribute z is quantified from the bootstrap distribution of each lag estimate. Erten et al. (2020) proposed a methodology for assessing the uncertainty in the experimental variogram considering the nonstationary case. In their methodology, they first estimate the drift model with its uncertainty; they then generate a number of likely drift models using maximum likelihood estimation. The residuals are calculated for each drift model. Each residual set is randomly sampled with replacement, and the resamples from each residual set are combined according to the probability assigned to the corresponding drift models. For each resample, the experimental variogram is computed, and the percentile confidence intervals are calculated for each lag. Khan and Deutsch (2016) extended the spatial bootstrap procedure to the multivariate context by considering the spatial correlation of K attributes sampled at n locations $\{z_k(\mathbf{u}_i), i = 1, \dots, n; k = 1, \dots, K\}$. In their methodology, they first generate m bootstrap realizations of the parameter of interest (mean and variance) for each attribute. They then use each bootstrap realization as a reference distribution to normal score transform the original multivariate sample. The cosimulation of the normal scored attributes is performed, and the final back-transformation of the realizations is carried out with m transform tables. A review of the traditional and

spatial bootstrap methods is given with examples in the next two sections, followed by a discussion section explaining how the domain limits can be considered in the spatial bootstrap procedure. Finally, the summary and conclusion section is presented.

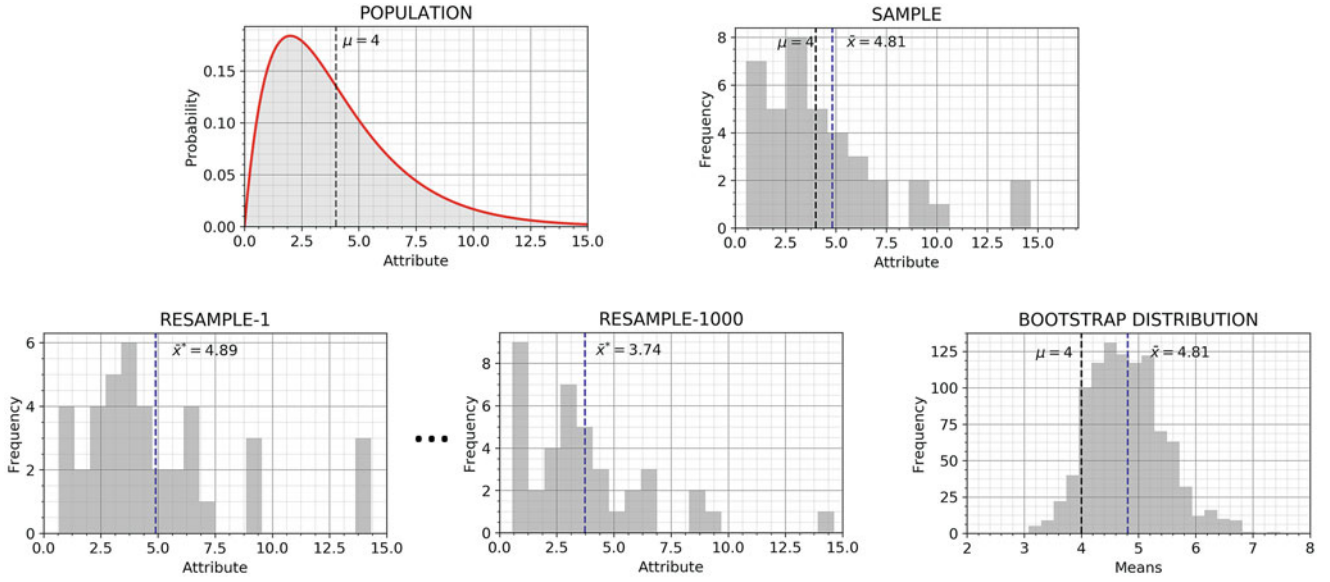
Traditional Bootstrap Resampling

To make any statistical inference without uncertainty, one needs access to the entire population. However, there is only a relatively small sample available, and the statistic must be estimated from the available sample. Consider a population distribution $F(x)$ and a small sample of n iid observations $\{X_i, i = 1, \dots, n\}$ sampled from $F(x)$. Because $F(x)$ is unknown, any statistic of interest, for example, the mean $\bar{X}$ must be estimated from the sampling distribution $F_X(x)$. The uncertainty in the estimated mean can be calculated by:

$$\text{Var}\{\bar{X}\} = \frac{\sigma_X^2}{n} \quad (1)$$

where σ_X^2 is the variance of the sample and n is the number of observations. It is clear from Eq. 1 that as the number of observations increases, the uncertainty decreases. The traditional bootstrap procedure quantifies the uncertainty in the estimated mean by randomly sampling the available data (n) with replacement; that is, we first generate n uniformly distributed random numbers $\{p_i, i = 1, \dots, n\}$ where p lies in the range $0 \leq p \leq 1$, and the corresponding quantiles $\{X_{i,j}^* = F_X^{-1}(p_{i,j}), i = 1, \dots, n; j = 1, \dots, m\}$ are read from the sampling distribution. This procedure is repeated many times, say $m = 1000$ for a stable distribution. Considering each resample $\{X_{i,j}^*, i = 1, \dots, n; j = 1, \dots, m\}$, some observations may appear multiple times and others may not appear at all due to the randomness of the selection of the observations. The mean $\bar{X}_j^*$ is estimated from each resample, and the distribution of the pooled set of means is referred to as the Monte Carlo approximation to the bootstrap distribution, as presented in Fig. 1. As can be seen from Fig. 1, the expected value $E(X) = \mu$ of the population is 4. The small sample of 30 observations was randomly drawn from this population, and the estimated mean of the sampling distribution is $\bar{X} = 4.81$. The observations were randomly drawn with replacement from the sampling distribution, resulting in the resamples $\{X_{i,j}^*, i = 1, \dots, n = 30; j = 1, \dots, m = 1000\}$, each with varying estimated means ($\bar{X}_1^* = 4.89, \bar{X}_{1000}^* = 3.74$). One can see from Fig. 1 that the average (or theoretical mean) of the bootstrap distribution is equal to the sample mean.

The bootstrap confidence intervals can be constructed from the bootstrap distribution. For example, considering 95% confidence interval, the extreme values corresponding to 5% of the bootstrap distribution ($\alpha = 0.05$) are discarded;



Bootstrap, Fig. 1 Traditional bootstrap procedure for $n = 30$ iid observations

that is, the 2.5% smallest and 2.5% largest values are not considered. Given that the bootstrap distribution presented in Fig. 1 is composed of 1000 means, the 25 smallest mean values and 25 largest mean values should be discarded. Once m estimated means are sorted in ascending order of magnitude, the 95% confidence interval is calculated by:

$$\bar{X}_{(\alpha/2) \cdot m}^* < \mu < \bar{X}_{(1-\alpha/2) \cdot m}^* \quad (2)$$

where μ is the expected value of the underlying population; $\bar{X}_{(\alpha/2) \cdot m}^*$ and $\bar{X}_{(1-\alpha/2) \cdot m}^*$ are the estimated means corresponding to the 25th and 975th ranks, respectively. Provided that the original sample is representative of the population and consists of random observations, the bootstrap samples should yield reliable confidence estimates. Considering the example given in Fig. 1, the 95% confidence (or probability) interval for the expected value of the population is $Prob\{\mu \in [3.64, 6.28]\}$.

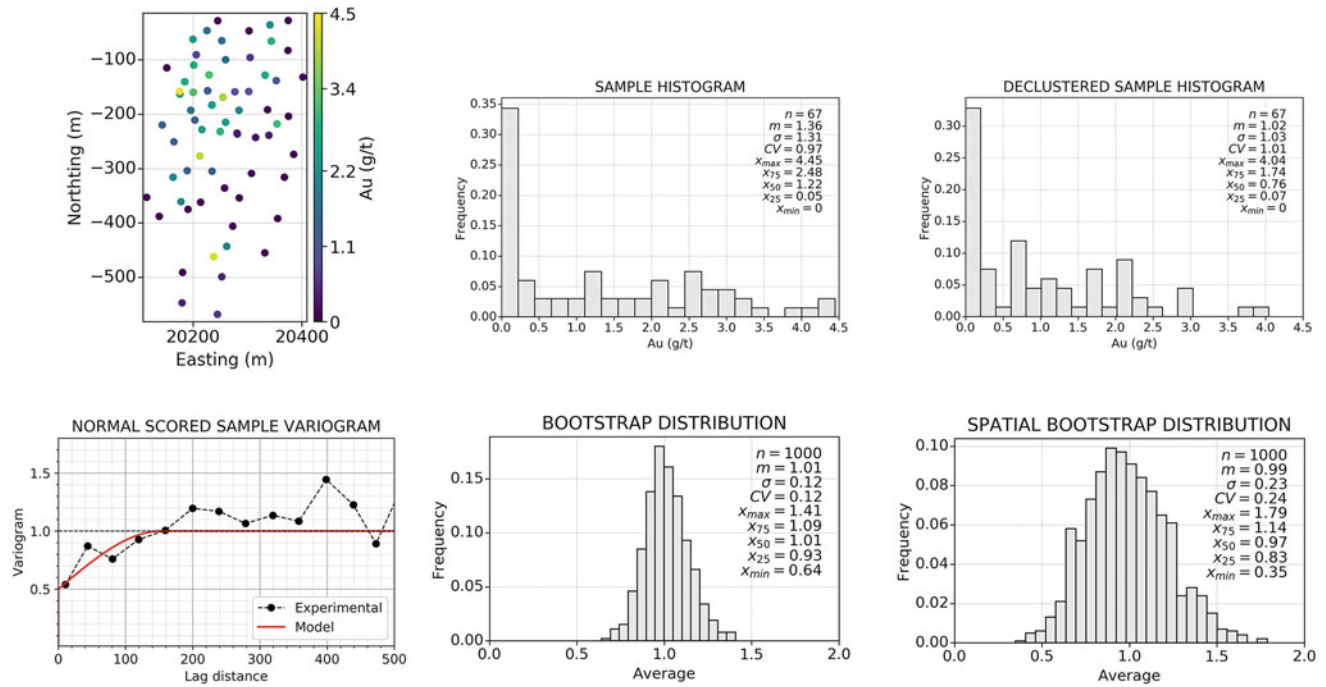
Spatial Bootstrap Resampling

Consider that Z is an attribute sampled at n locations $\{z(\mathbf{u}_i), i = 1, \dots, n\}$ and modeled by a second-order stationary random function $Z(\mathbf{u})$. Because the observations are spatially correlated, the uncertainty in the estimated mean $\bar{z}$ cannot be calculated using Eq. 1. Considering the spatial correlation between the observations, the variance in the sample mean is in fact equal to the average covariance between the observations; that is,

$$\bar{C}(n, n) = \frac{\sigma^2}{n_{\text{eff}}} \quad (3)$$

where $\bar{C}(n, n)$ is the average covariance, σ^2 is the population variance which is equal to one in the case of normal scored observations, and n_{eff} is the effective number of observations indicating the spatial degrees of freedom in the distribution of the mean. The spatial bootstrap procedure for a single variable is given in the following steps:

1. Define a representative distribution $F_Z(z)$ of the sample of n correlated observations.
2. Transform the spatial data $\mathbf{z}^T = (z_1, z_2, \dots, z_n)$ into the corresponding normal scores $\mathbf{y}^T = (y_1, y_2, \dots, y_n)$ using the normal score transform $\mathbf{y} = G^{-1}(F(\mathbf{z}))$.
3. Fit an analytical model to the experimental variogram of the normal scores of the data, and infer the covariance model $C(\mathbf{h})$.
4. Calculate the $n \times n$ data-to-data covariance matrix $\mathbf{C}$ from the covariance model $C(\mathbf{h})$.
5. Decompose the covariance matrix $\mathbf{C}$ as the product of lower triangular matrix and upper triangular matrix using the Cholesky factorization; that is, $\mathbf{C} = \mathbf{L} \cdot \mathbf{L}^T$, where $\mathbf{L}$ is an $n \times n$ lower triangular matrix.
6. Simulate the unconditional realization of $\mathbf{y}$; that is, $\mathbf{y} = \mathbf{L} \cdot \mathbf{w}$, where $\mathbf{y}$ is the $n \times 1$ unconditional realization, $\mathbf{L}$ is the $n \times n$ lower triangular matrix, and $\mathbf{w}$ is the $n \times 1$ independent (or uncorrelated) standard Gaussian deviates.
7. Convert the simulated Gaussian values to the corresponding probability values; that is, $\mathbf{p} = \mathbf{G}(\mathbf{y})$, where $\mathbf{p}$ is the $n \times 1$ vector of probability values $[0, 1]$.



Bootstrap, Fig. 2 Spatial bootstrap procedure for $n = 67$ spatially correlated observations

8. Draw $\mathbf{z}^*$ values from the empirical conditional cumulative distribution of $\mathbf{z}$ using the probability values; that is, $\mathbf{z}^* = F_{\mathbf{Z}}^{-1}(\mathbf{p})$.
9. Calculate the statistic of interest from the generated realization $\mathbf{z}^*$.
10. Repeat steps 6–9 for a large number m (at least 1000) of resamples.

The LU simulation method, due to its efficiency, is generally used to generate unconditional realizations of n observations with the correct covariance. It is also noted that the LU decomposition of the covariance matrix is required only once. The example given in Fig. 2 indicates that the Au (g/t) grade is sampled at $n = 67$ locations. One can notice that the high grade zone is located in the upper half of the domain. The cell declustering was applied using 85×85 m grid, and the estimated mean of the declustered sample histogram is 1.02. When compared to the mean of the naïve (equal weighted) histogram (1.36), there is a significant decrease in the mean. The experimental variogram of the normal scores of the empirical Au (g/t) data was computed and fitted to an analytical model. Because the second-order stationary assumption holds, the covariance function was obtained through $C(\mathbf{h}) = C(0) - \gamma(\mathbf{h})$. It can be seen from Fig. 2 that the uncertainty in the mean is much less in the case where the spatial correlation between the observations was not considered (Eq. 1). When taking into account the spatial correlation, n_{eff} (Eq. 3) was calculated to be 28.78 which is much less than the number of observations $n = 67$ in the sample. In both

cases, the means of the bootstrap distributions are rather close to the original sample mean.

Discussion

It can be counterintuitive that the spatial bootstrap uncertainty increases when there is greater correlation between the data, yet those data provide more information on the attribute at unsampled locations. It can also be counterintuitive that the spatial bootstrap uncertainty does not consider domain limits. The uncertainty for a small domain with limits close to the data should be less than for a large domain with vast unsampled regions. The spatial bootstrap provides prior uncertainty in a parameter, that is, before conditioning to the data and considering domain limits. The spatial bootstrap parameters used in subsequent spatial modeling are updated by the conditioning data and the domain limits to provide a posterior uncertainty.

Summary and Conclusions

The use of bootstrap resampling technique provides a practical way of accounting for the uncertainty in the global model parameters (i.e., covariance functions, distribution functions) in many geoscience applications. There are two assumptions to consider in the bootstrap procedure: (1) The sample is representative of the underlying population, and (2) the

sample consists of iid observations. The second assumption is violated in the case of spatial samples; therefore, the covariance function describing the spatial correlation between the observations should be taken into account when bootstrapping the spatially correlated observations. Considering the case where there is only a small sample available for spatial modeling of the attribute of interest, neglecting the uncertainty in the model parameters may lead to a bias in the final attribute model.

Two examples show how the bootstrap distributions of the estimated statistic of interest can be constructed through both traditional and spatial bootstrap procedures. In both cases, the bootstrap distributions are normally distributed with a mean equal to approximately the sample mean and a variance calculated through Eqs. 1 and 3, respectively. The percentile confidence intervals can be calculated from the bootstrap distributions, which assess the uncertainty in the estimated statistic. It has also been shown through the second example that in the case of spatial correlation, the uncertainty in the model parameters is expected to be larger than that of the independence case. This is due to the fact that the effective number of observations (n_{eff}) (or number of independent observations) is reduced due to the spatial correlation. The reduced n_{eff} indicates the redundancy between the observations considering the spatial correlation.

Cross-References

- [Kriging](#)
- [Maximum Likelihood](#)
- [Sequential Gaussian Simulation](#)

Bibliography

- Deutsch CV (2004) A statistical resampling program for correlated data: spatial bootstrap. Center for Computational Geostatistics Annual Report Papers
- Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman & Hall, Boca Raton
- Erten O, Pardo-Igúzquiza E, Olea RA (2020) Assessment of experimental semivariogram uncertainty in the presence of a polynomial drift. *Nat Resour Res* 29(2):1087–1099
- Feyen L, Caers J (2006) Quantifying geological uncertainty for flow and transport modeling in multi-modal heterogeneous formations. *Adv Water Resour* 29(6):912–929
- Journel AG, Bitanov A (2004) Uncertainty in N/G ratio in early reservoir development. *J Pet Sci Eng* 44:115–130
- Khan KD, Deutsch CV (2016) Practical incorporation of multivariate parameter uncertainty in geostatistical resource modeling. *Nat Resour Res* 25(1):51–70
- Olea RA, Pardo-Igúzquiza E (2011) Generalized bootstrap method for assessment of uncertainty in semivariogram inference. *Math Geosci* 43(2):203–228
- Pardo-Igúzquiza E, Olea RA (2012) VARBOOT: a spatial bootstrap program for semivariogram uncertainty assessment. *Comput Geosci* 41:188–198
- Pyrz MJ, Deutsch CV (2014) Geostatistical reservoir modelling, 2nd edn. Oxford University Press
- Solow AR (1985) Bootstrapping correlated data 1. *Math Geol* 17(7):769–775

Burrough, Peter Alan

Andrew U. Frank

Geoinformation, Technical University, Vienna, Austria

Biography

Peter Burrough was born in Wimborne (England) in 1944 and died in Leiden (Netherlands) 2009. He studied chemistry at the University of Sussex (Brighton) and obtained a Bachelor of Science there in 1965. When he started his doctoral studies at the University of Oxford, he switched to soil sciences; he obtained his doctorate in 1969 with a thesis on “Studies of Soil Survey Methodology.” In the first years he worked on soil surveys in Malaysia but returned to academe for teaching and research and was eventually appointed Professor of Physical Geography and Geographical Information Systems at the Geographical Institute of Utrecht University in 1984.

Peter Burrough was a gifted and productive researcher, picking up new ideas, often from other disciplines and adapted them to geoscience. In 1981 he published an article in *Nature* “Fractal dimensions of landscapes and other environmental data” a subject introduced by Mandelbrot only a few years earlier. Goodchild (2009) summarized the contribution as: “it is the deviations from self-similarity that provide insights into the nature of geographic phenomena.”

Burrough saw early on the contribution computers can make to the practice of geoscience; he looked beyond automated cartography and saw the potential for analyzing geoscientific data. He had the broad academic background to blend different scientific disciplines and bring them to bear on geoscientific topics.

Peter Burrough is probably best known for his book *Principles of Geographical Information Systems for Land Resources Assessment* published in 1986 (Burrough 1986). It was the first comprehensive book on the then novel subject; it went through various editions and is still a widely used text book.

The contribution of this book, however, was beyond just to be the first text book; it showed the importance of geographic information science as a tool and as a science; it demonstrated how scientific principles from mathematics, geometry, and

statistics can be applied to geoscientific topics to yield new insights, helpful for the then emerging environmental sciences.

When affordable computers became powerful enough, Burrough expanded with his graduate students the, then common, static “map algebra” (Tomlin 1990) to a **dynamic** “PC Raster” system (Wesseling et al. 1996) which visualized dynamic environmental processes in space and time described by partial differential equations.

Peter Burrough was always aware of the errors in geoscientific data and the influence of such errors on results and produced with his graduate students a software to explore the effects (Heuvelink 2007). He co-organized a workshop on geographic objects with indeterminate boundaries (Burrough and Frank 1996) which tried to bridge the gap in understanding the influence of uncertainty and error in spatial data in physical and social science data.

Peter Burrough was an inspiring teacher and a mentor for many graduate students, many of whom became successful

researchers and university professors themselves, continuing Burrough’s scientific contributions.

Bibliography

- Burrough PA (1981) Fractal dimensions of landscapes and other environmental data. *Nature* 294:240–242
- Burrough PA (1986) *Principles of Geographical Information Systems for Land Resources Assessment*. Oxford University Press, Oxford
- Burrough PA, Frank AU (eds) (1996) *Geographic objects with indeterminate boundaries*. Taylor and Francis, Bristol
- Goodchild MF (2009) A research appreciation. *Trans GIS* 13(3): 247–252
- Heuvelink GBM (2007) The Data Uncertainty Engine (DUE): a software tool for assessing and simulating uncertain environmental variables. *Comput Geosci* 33(2):172–190
- Tomlin CD (1990) *Geographic information systems and cartographic modeling*. Prentice Hall, Englewood Cliffs
- Wesseling CG, Karssenber DJ, Burrough PA, van Deursen WP (1996) Integrating dynamic environmental models in GIS: the development of a Dynamic Modelling language. *Trans GIS* 1(1):40–48

Cartogram

Michael T. Gastner
Division of Science, Yale-NUS College, Singapore,
Singapore

Synonyms

Anamorphic maps

Definition

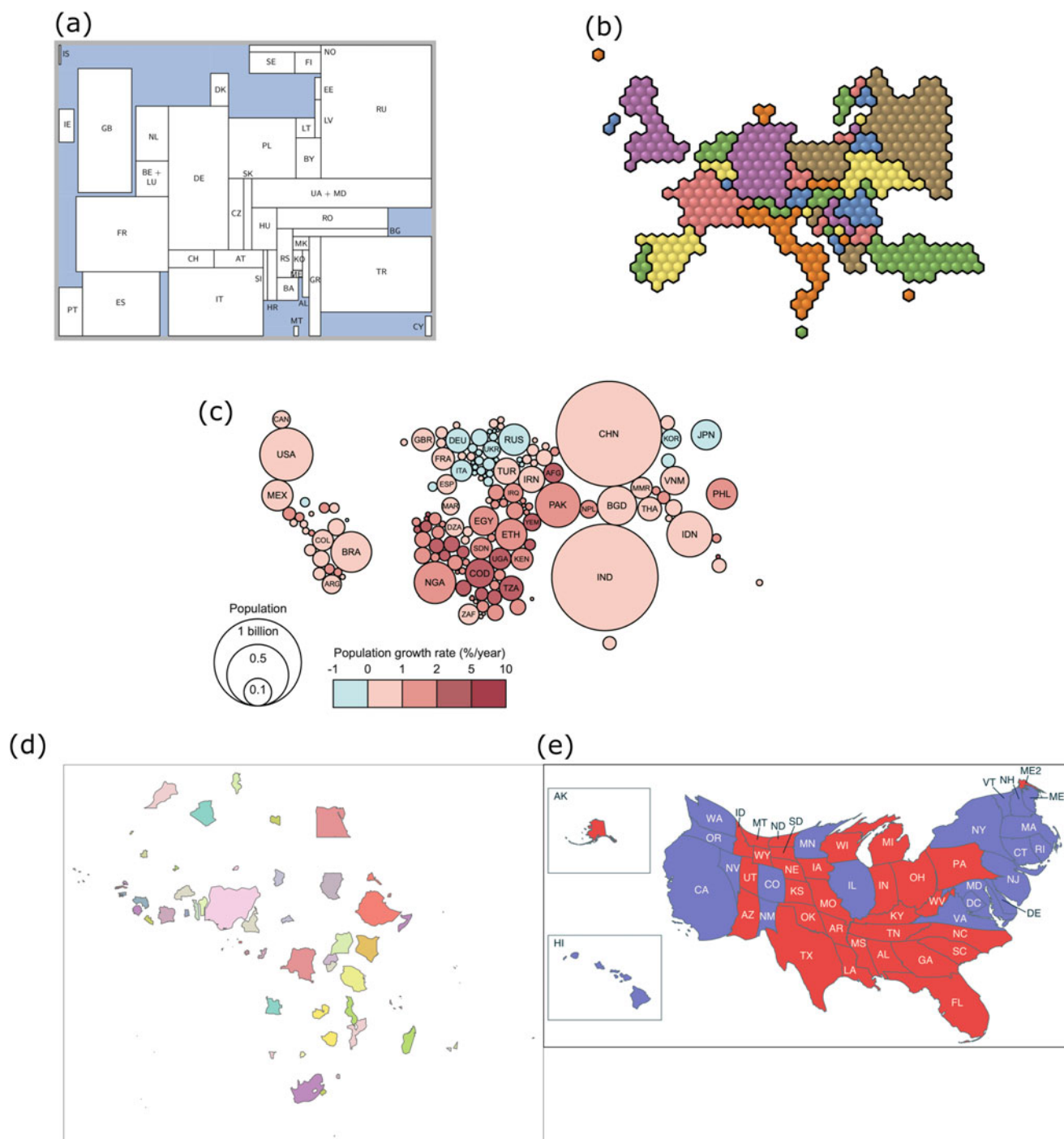
Cartograms are thematic maps in which the sizes of geographic objects appear in proportion to a quantitative mapping variable. There are two types of cartograms. If the mapping variable is represented by the areas of the depicted regions, the map is called an *area cartogram*. If distances between points are a substitute for the mapping variable, the map is called a *distance cartogram*. Both cartogram types visualize quantitative data by distorting familiar geographic features. However, area cartograms and distance cartograms have distinct applications, and their construction requires different mathematical and algorithmic techniques. Suitable mapping variables for area cartograms are quantities that can be associated with nonoverlapping regions on a map (e.g., countries or provinces). Area cartograms are frequently used in human geography to visualize population distributions (“isodemographic maps”) or economic indicators (e.g., gross domestic product). Distance cartograms usually represent travel times (“time-space maps”), but other measures of separation (e.g., travel costs or fuel consumption) can also be used.

Area Cartogram

The mathematical objective of an area cartogram (also known as “value-by-area map”) is to transform geographic regions, given in the form of multipolygons, into new multipolygons with areas that match quantitative data associated with each region. The output should maintain certain properties of conventional geographic maps (e.g., the locations and shapes of regions) and maintain the regions’ topology (i.e., regions should share a common border on the cartogram if and only if they share a common border in geographic space). It is generally impossible to achieve all these objectives simultaneously. A variety of area cartogram types have been developed that relax different constraints. The main categories of area cartograms are rectangular cartograms, mosaic cartograms, circular cartograms, noncontiguous cartograms, and contiguous cartograms (Fig. 1). For a review of area cartograms, including further cartogram types, see Tobler (2004), and Nusrat and Kobourov (2016).

Rectangular Cartograms

In a rectangular cartogram, each region is represented by a rectangle with an area equal to a numeric input (Fig. 1a). Rectangular cartograms are the oldest cartogram type with hand-drawn examples dating back to the late nineteenth century (see Tobler 2004 for a historical account). Maintaining the correct adjacency of regions is generally impossible in rectangular cartograms. Algorithms have been created to automate the construction of rectangular cartograms with correct topology, but only after landlocked regions with fewer than four neighbors are manually merged with one of the surrounding regions (Buchin et al. 2012). A variation of the rectangular cartogram with fewer topological constraints is the rectilinear cartogram, in which each region can be composed of multiple rectangles.



Cartogram, Fig. 1 Area cartograms. (a) Rectangular cartogram depicting the population of Europe (Buchin et al. 2012). (b) Mosaic cartogram for the same data (Cano et al. 2015). (c) Circular cartogram illustrating the predicted world population in 2030 (Inoue 2011). (d) Noncontiguous cartogram representing the 2005 population of African

countries (Jeworutzki 2020). (e) Contiguous cartogram for the 2016 US presidential election. The area of each state in the cartogram is proportional to the number of electors. Red states were won by Republicans, blue states by Democrats (Gastner et al. 2018)

Mosaic Cartograms

Mosaic cartograms divide the map into a set of small square-shaped or hexagonal tiles, each of equal size and orientation (Fig. 1b). The most suitable thematic mapping variables for

mosaic cartograms are those that take small integer values so that one tile corresponds to one discrete unit (e.g., a seat in parliament). Tiles are assigned to regions such that contiguous geographic regions are represented by contiguous sets of tiles.

For some data, topological errors are inevitable because the details of real geographic boundaries cannot be represented by the discrete shapes of the tiles. To minimize the topological errors, Dorling (1996) developed a cellular automaton algorithm for the construction of mosaic cartograms. A recently proposed algorithm aims to improve the preservation of the regions' shapes (Cano et al. 2015).

Circular Cartograms

Circular cartograms represent each region by a circle with an area that is proportional to the thematic mapping variable (Fig. 1c). Circles are placed such that they may touch, but do not overlap. Dorling's (1996) circular cartogram algorithm and a recent refinement by Inoue (2011) aim to place the centers of the circles approximately at the same position where the region centroids would be on a conventional map. Unlike rectilinear and mosaic cartograms, circular cartograms always leave unfilled space in the interior of the map. The restriction to circular shapes may also force neighboring geographic regions to appear separated on the cartogram.

Noncontiguous Cartograms

Noncontiguous cartograms deliberately introduce empty space between the depicted regions (Fig. 1d) and, thus, make no attempt to represent the topological relations between the regions. The empty space between polygons allows noncontiguous cartograms to depict the shapes of the regions accurately. Olson (1976) described a construction method for noncontiguous cartograms, in which each polygon is shrunk isotropically towards its centroid. The method tends to create excessive empty space between the polygons. It can also lead to region overlaps if polygons are concave or contain holes. A fully automated fail-safe algorithm has yet to be described in the literature. Hence, the construction of noncontiguous cartograms currently relies on manual adjustments with image editing software.

Contiguous Cartograms

Contiguous cartograms correctly represent common borders and tripoints between the depicted regions (Fig. 1e). Construction methods for contiguous cartograms can be divided into two distinct classes. The first class ("boundaries-only" methods) restricts itself to moving a finite number of boundary points from their geographic position (e.g., given as latitude and longitude) to their positions on a cartogram. The second class creates a cartogram by establishing a "density-equalizing map projection" that shifts every point in continuous geographic space to a new position, regardless of whether the point is on a boundary.

Boundaries-only algorithms can influence the polygon shapes more directly than algorithms that involve continuous map projections. Consequently, several boundaries-only algorithms explicitly seek a compromise between

maintaining polygon shapes and achieving correct polygon areas (House and Kocmoud 1998; Keim et al. 2004). A disadvantage of the boundaries-only approach is that it does not attribute any semantics to points that are not on a boundary. By contrast, a density-equalizing map projection makes it possible to show additional data that are resolved at a finer level than the polygons to be displayed (e.g., addresses, roads, or distributions on a fine-grained grid; see Hennig 2013).

A density-equalizing map projection is a two-dimensional function $\mathbf{T} = (T_x; T_y)$ that satisfies

$$\det J_{\mathbf{T}} \equiv \frac{\partial T_x}{\partial x} \frac{\partial T_y}{\partial y} - \frac{\partial T_y}{\partial x} \frac{\partial T_x}{\partial y} = \frac{\rho(x, y)}{\bar{\rho}} \quad (1)$$

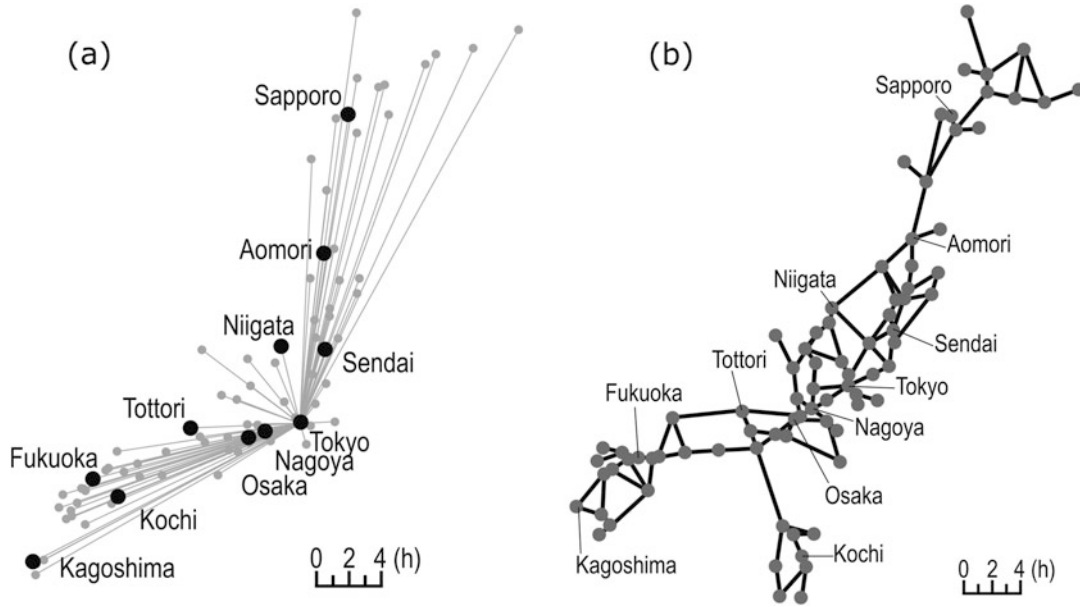
where $\det J_{\mathbf{T}}$ is the Jacobian determinant of $\mathbf{T}$, (x, y) are coordinates on an equal-area projection, $\rho(x, y)$ is the local density of the mapping variable (e.g., population per square kilometer), and $\bar{\rho}$ is the spatial average. Equation (1) does not uniquely specify a projection. Various additional constraints have been proposed to determine a unique solution. Tobler (1973) suggested selecting the most angle-preserving projection among all solutions to Eq. (1). In practice, Tobler's optimization algorithm failed to converge. However, several alternatives, based on analogies from physics, have led to satisfactory results (Gusein-Zade and Tikunov 1993; Sun 2020). The technique by Gastner et al. (2018) treats $\rho(x, y)$ as the density of a fluid that equilibrates over time. This method solves Eq. (1) by calculating a vortex-free and mass-conserving velocity field. The calculation, performed with fast Fourier transforms, has been implemented as a web application (<https://go-cart.io/>; see Tingsheng et al. 2019).

Distance Cartogram

The objective of a distance cartogram (also called "linear cartogram") is to place a set of points in a flat two-dimensional geometric space such that the distances between the points are proportional to a quantitative measure of separation. Hereinafter, we assume that the measure of separation is travel time, which is by far the most frequently used mapping variable for distance cartograms.

There are two broad categories of distance cartograms. In the first category ("central-point cartogram"), only travel times from a central location are represented on the cartogram (Fig. 2a). The second category ("network cartograms") aims to represent travel times in a network whose links are not necessarily connected to a unique central node (Fig. 2b).

The construction of central-point cartograms is straightforward. Points can be placed at a distance that is strictly proportional to the travel time, and the angle of the connecting line to the central point can be chosen arbitrarily (e.g., to



Cartogram, Fig. 2 Distance cartograms. (a) Central-point cartogram showing the travel time from Tokyo by train in 1965. (b) Network cartogram based on pairwise travel times between 81 Japanese cities. (Figure from Shimizu and Inoue (2009))

match the angle on a conventional map). Network cartograms pose greater challenges because it is generally impossible to simultaneously satisfy the distance constraints for all point pairs. Instead, the aim is to minimize the deviation from proportionality. A common objective function is

$$\min f(x_1, \dots, x_n, y_1, \dots, y_n) = \sum_{(i,j) \in L} \left(t_{ij} - \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \right)^2 \quad (2)$$

where x_i and y_i are the coordinates of point i on the cartogram, and t_{ij} is the travel time from i to j . L is the set of all links whose travel times are to be visualized. Minimizing f is equivalent to multidimensional scaling, a common mathematical technique for dimension reduction in statistics. The solution to Eq. (2) is not unique because any rotation and translation leaves f unchanged. To select a unique solution, it is sensible to conduct the following post-hoc steps (Ewing and Wolfe 1977):

- Scale the cartogram coordinates such that the mean of all distances between pairs of points matches the mean interpoint distance in physical space.
- Shift the cartogram coordinates such that the centroid of the points on the cartogram is the same as the centroid of their physical locations.
- Rotate the cartogram coordinates around the centroid such that the rotated coordinates minimize the sum of the squared distances from the physical coordinates.

If L contains all $n(n-1)/2$ possible links between n points, good numerical results were obtained using the technique described above (Marchand 1973). However, if some point pairs are missing from L , there can be multiple solutions to Eq. (2) which differ by more than an overall rotation and translation. In this case, the algorithm should aim to minimize angular distortion to produce readable cartograms (Shimizu and Inoue 2009).

It is often informative to show more than the transformed network locations $(x_1, y_1), \dots, (x_n, y_n)$ on the cartogram. For example, coastlines and administrative boundaries can add valuable context. In this case, the transformation from the physical locations $(x_i^p, y_i^p), i \in \{1, \dots, n\}$ to the cartogram locations (x_i, y_i) must be spread into a continuous two-dimensional space. Ideally, this task is achieved by a continuous map projection $\mathbf{T}$ that interpolates smoothly between the n points in the network such that $\mathbf{T}(x_i^p, y_i^p) = (x_i, y_i)$. However, standard interpolation methods can result in graticule cells whose winding orientation is locally inverted. That is, there can be coordinates (x, y) where the Jacobian determinant $\det J_{\mathbf{T}}(x, y)$ is negative, especially where fast long-distance connections are geographically close to slow short-distance connections. To alleviate this problem, Spiekermann and Wegener (1994) proposed a modification of multidimensional scaling, called “stepwise multidimensional scaling.” However, a rigorous proof that this method avoids local inversions in $\mathbf{T}$ remains an open research challenge.

Summary

Conventionally, geographic maps are designed to faithfully represent metric properties such as areas, distances, or angles. Cartograms offer an alternative map design with the primary aim of visualizing quantitative statistical data instead of geometric properties. Data associated with nonoverlapping two-dimensional regions can be shown on area cartograms, for which many different designs have been developed over the past decades. Data that quantify the degree of separation between points (e.g., travel time) can be visualized with distance cartograms. Although both types of cartograms appear inevitably distorted, they can be effective alternatives to traditional thematic maps (e.g., choropleth maps or proportional-symbol maps).

Cross-References

- [Data Visualization](#)
- [Fast Fourier Transform](#)
- [Optimization in Geosciences](#)
- [Tobler, Waldo](#)

Bibliography

- Buchin K, Speckmann B, Verdonchot S (2012) Evolution strategies for optimizing rectangular cartograms. In: Xiao N, Kwan M-P, Goodchild MF, Shekhar S (eds) *Geographic information science*. Springer, Berlin, pp 29–42
- Cano RG, Buchin K, Castermans T, Pieterse A, Sonke W, Speckmann B (2015) Mosaic drawings and cartograms. *Comput Graph Forum* 34:361–370
- Dorling D (1996) *Area cartograms: their use and creation*. School of Environmental Sciences, University of East Anglia, Norwich
- Ewing GO, Wolfe R (1977) Surface feature interpolation on two-dimensional time-space maps. *Environ Plan A* 9:429–437
- Gastner MT, Seguy V, More P (2018) Fast flow-based algorithm for creating density-equalizing map projections. *Proc Natl Acad Sci U S A* 115:E2156–E2164
- Gusein-Zade SM, Tikunov VS (1993) A new technique for constructing continuous cartograms. *Cartogr Geogr Inf Syst* 20:167–173
- Hennig B (2013) *Rediscovering the world: map transformations of human and physical space*. Springer, Berlin
- House D, Kocmoud C (1998) Continuous cartogram construction. *Proc Visualization 98* (Cat. No. 98CB36276):197–204
- Inoue R (2011) A new construction method for circle cartograms. *Cartogr Geogr Inf Sci* 38:146–152
- Jeworutzki S (2020) cartogram: Create Cartograms with R. <https://CRAN.R-project.org/package=cartogram>. Accessed 19 Sep 2021
- Keim D, North S, Panse C (2004) CartoDraw: a fast algorithm for generating contiguous cartograms. *IEEE Trans Vis Comput Graph* 10:95–110
- Marchand B (1973) Deformation of a transportation surface. *Ann Am Assoc Geogr* 63:507–521
- Nusrat S, Kobourov S (2016) The state of the art in cartograms. *Comput Graph Forum* 35:619–642
- Olson JM (1976) Noncontiguous area cartograms. *Prof Geogr* 28:371–380
- Shimizu E, Inoue R (2009) A new algorithm for distance cartogram construction. *Int J Geogr Inf Sci* 23:1453–1470
- Spiekermann K, Wegener M (1994) The shrinking continent: new time-space maps of Europe. *Environ Plan B* 21:653–673
- Sun S (2020) Applying forces to generate cartograms: a fast and flexible transformation framework. *Cartogr Geogr Inf Sci* 47:381–399
- Tingsheng S, Duncan IK, Gastner MT (2019) go-cart.io: a web application for generating contiguous cartograms. *Abstr Int Cartogr Assoc* 1:333
- Tobler WR (1973) A continuous transformation useful for districting. *Ann N Y Acad Sci* 219:215–220
- Tobler W (2004) Thirty five years of computer cartograms. *Ann Am Assoc Geogr* 94:58–73

Chaos in Geosciences

Frits Agterberg¹ and Qiuming Cheng²

¹Central Canada Division, Geomathematics Section, Geological Survey of Canada, Ottawa, ON, Canada

²State Key Lab of Geological Processes and Mineral Resources, China University of Geosciences, Beijing, China

Definition

Randomness is widespread in the sciences including geoscience. The concepts of probability and random deviations from an average value have been known to exist since the sixteenth century (Hacking 2006). During the second part of the twentieth century, it was found that various random variables could be modeled as fractals or multifractals. It was also discovered that nonlinear differential equations can result in varied forms of chaos closely linked to fractals or multifractals. According to Turcotte (1997), the introduction of the concept of chaos stems from the original work by Lorenz (1963) who employed nonlinear equations to describe thermal convection in a fluid. Although his equations were deterministic, they resulted in a chaotic solution. Infinitesimal variations in initial conditions led to unpredictable differences in the outcome. Later, Motter and Campbell (2013) showed that the Lorenz attractor has a fractal dimension.

May (1976) introduced the logistic map to represent population dynamics of animal species with an annual breeding cycle. Schuster (1995) and Turcotte (1997) explain May's logistic map model in detail and present other examples of deterministic chaos. Bak (1996) introduced a new theory of complex systems using geoscience examples including his sandpile model for avalanche sizes that satisfies Zipf's (1949) law for fractals describing a straight line relationship on log-log plots. Sagar et al. (2003) expanded Bak's approach by simulating avalanches in sand dune formation experiments.

Introduction

A question to which new answers are being sought is: Where does the randomness in Nature come from? Nonlinear process modeling is providing new clues to answer this question. Deterministic models can produce chaotic results with fractal properties. A famous early example was described in Poincaré's (1899) study of the three-body problem in mechanics. This author observed that in some systems it may happen that small differences in the initial conditions produce very great ones in the final phenomena, and "... prediction becomes impossible." Nevertheless, the randomness created by nonlinear processes obeys its own specific laws. Many scientists including Turcotte (1997) have pointed out that, in chaos theory, otherwise deterministic Earth process models can contain terms that generate purely random responses. Examples include the logistic equation and the van der Pol equation with solutions that contain unstable fixed points or bifurcations.

Multifractals, which are spatially intertwined fractals and were anticipated by Hans de Wijs (1951), provide a novel approach to problem solving in situations where the attributes display strongly positively skewed frequency distributions. The original de Wijs model is discussed in more detail in the entry by S. Xie elsewhere in this Encyclopedia (also see Xie and Bao 2004). Another example of this type will be given later in the current entry. The multifractal method of local singularity mapping uses short-distance variability to delineate places with relatively strong enrichment or depletion of element concentration on geochemical maps and in other applications (Cheng and Agterberg 2009) offering a new approach for mineral exploration and regional environmental assessment. Such nonlinear developments are closely related to statistical theory of extreme events (Beirlant et al. 2005).

Model of de Wijs

A relatively simple multiplicative cascade model in 1-D, 2-D, or 3-D is the model of de Wijs (1951). This model is graphically illustrated in Figs. 1 and 2. In the original model of de Wijs, a block of rock is divided into two equal parts. The concentration value (ξ) of a chemical element in the block can be written as $(1 + d) \times \xi$ for one half and $(1 - d) \times \xi$ for the other half so that total mass is preserved. The coefficient of dispersion d is assumed to be independent of block size. For the first cell at the beginning of the process, ξ can be set equal to unity. This implies that all concentration values are divided by their overall regional average concentration value (μ). The index of dispersion (d) is independent of cell size. In 2-D space, two successive subdivisions into quarters result in 4 and 16 cells with concentration values as shown in Fig. 1.

	$(1+d)$	$(1-d)$		
	$(1+d)^2$	$(1+d)(1-d)$	$(1-d)^2$	
	$(1+d)(1-d)$	$(1-d)^2$		
$(1+d)^4$	$(1+d)^3(1-d)$	$(1+d)^2(1-d)^2$	$(1+d)(1-d)^3$	$(1-d)^4$
$(1+d)^3(1-d)$	$(1+d)^2(1-d)^2$	$(1+d)(1-d)^3$	$(1-d)^4$	
$(1+d)^2(1-d)^2$	$(1+d)(1-d)^3$	$(1-d)^4$		
$(1+d)(1-d)^3$	$(1-d)^4$			
$(1-d)^4$				

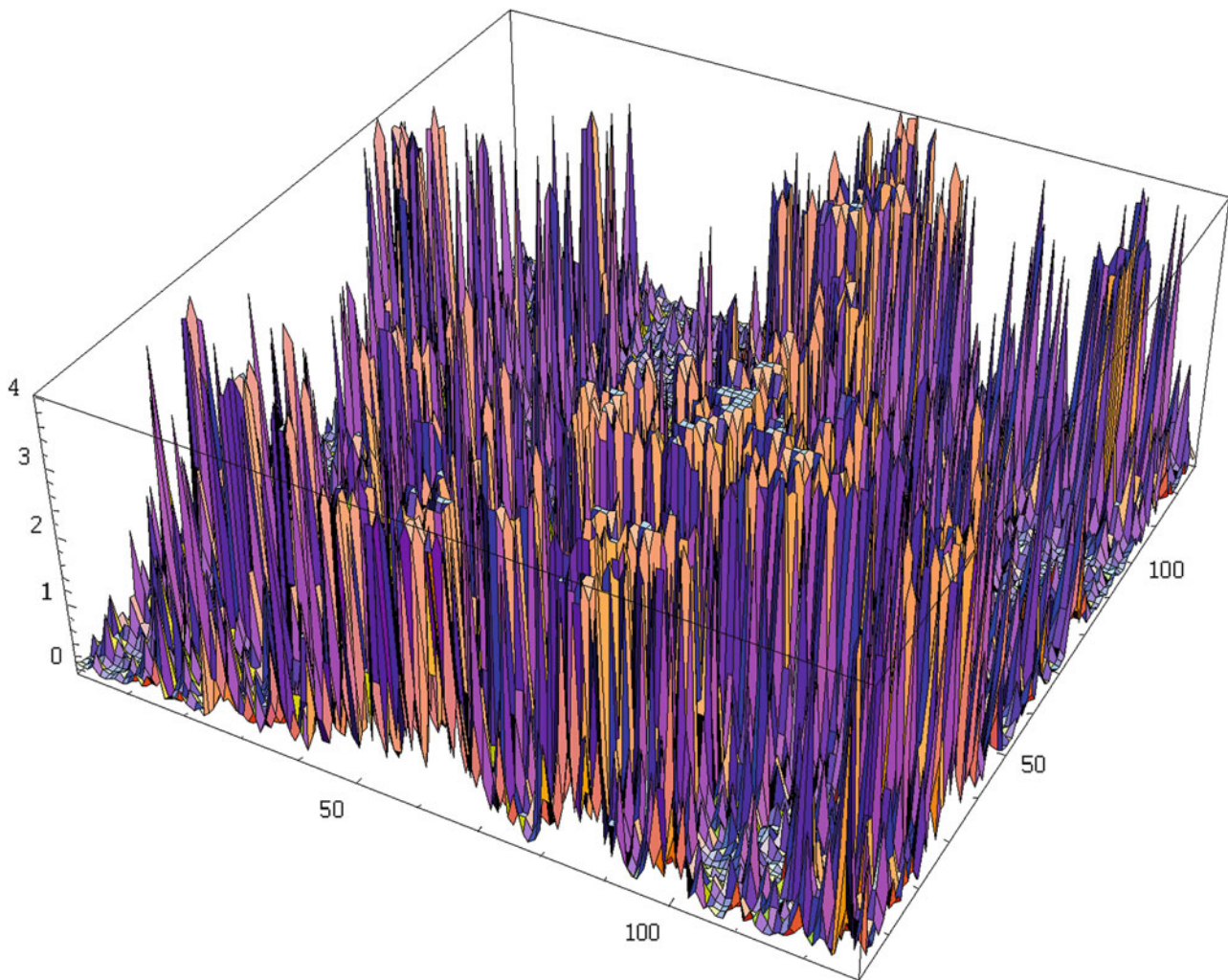
Chaos in Geosciences, Fig. 1 First stages of two-dimensional cascade model of de Wijs. Overall mean concentration value was set equal to one; d = dispersion index. Non-Random index matrix corresponds to (4×4) squares distribution of concentration values. (Source: Agterberg 2007, Fig. 2a)

The maximum element concentration value after k subdivisions is $(1 + d)^k$ and the minimum value is $(1 - d)^k$.

The model of de Wijs is a random cascade in which larger and smaller values are assigned to cells using a discrete random variable. Multifractal patterns generated by a random cascade of this type have more than a single maximum. The frequency distribution of the element concentrations at any stage of this process is called "log-binomial" because logarithmically transformed concentration values satisfy the binomial distribution model. The log-binomial converges to a lognormal distribution although its upper and lower value tails do remain weaker than those of the lognormal (Agterberg 2007). Because of its property of self-similarity, the model of de Wijs was recognized to be a multifractal by Mandelbrot (1989), who adopted this approach for various applications to the Earth's crust.

Figure 2 shows a 2-D log binomial pattern for $d = 0.4$ and $k = 14$. Increasing the number of subdivisions for the model of de Wijs (as in Fig. 1) to 14 results in the (128×128) pattern shown where values greater than 4 were truncated to more clearly display most of the spatial variability. The frequency distribution of all values is log-binomial and approximately lognormal except in the highest-value and lowest-value tails that are thinner. When the number of subdivisions becomes large, the end product cannot be distinguished from that of multiplicative cascade models in which the dispersion index D is modeled as a continuous random variable with mathematical expectation equal to 1 instead of the Bernoulli variable allowing the values $+d$ and $-d$ only. The lognormal model often provides good first approximations for regional background distributions of trace elements. The multifractal method used for estimating d is similar to the method for separation of geochemical anomalies from background introduced by Cheng (1994) and Cheng et al. (1994).

The model of de Wijs is a multifractal that satisfies the property of scale-independence. Its multifractal property



Chaos in Geosciences, Fig. 2 Realization of model of de Wijs (see Fig. 1 in 2-D for $d = 0.4$ and $n = 14$). Overall average value is equal to 1. Values greater than 4 were truncated. (Source: Agterberg 2007, Fig. 3)

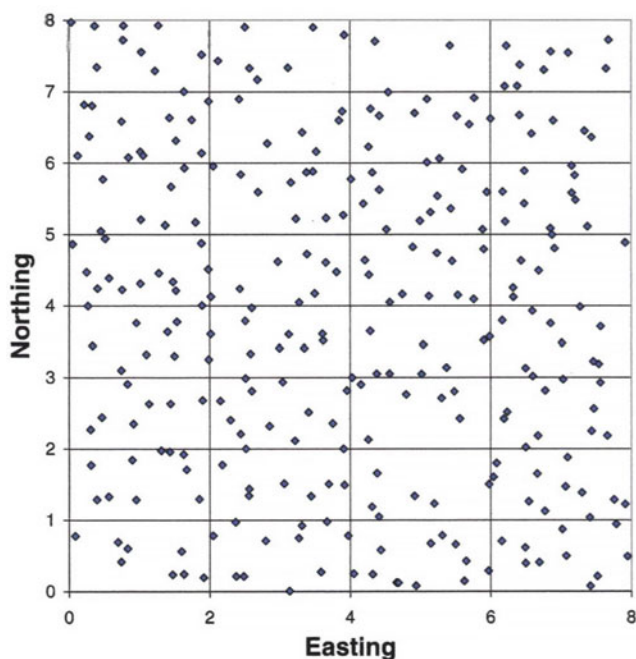
provides an example of renormalization group models producing fractal statistics explicitly utilizing the principle of scale invariance. Other methods of this type are discussed by Turcotte (1997), who in the last chapter of his book also provides examples of self-organized criticality not only resulting in patterns with fractal properties but basically originating from deterministic differential equations.

Gold in South Saskatchewan till Example

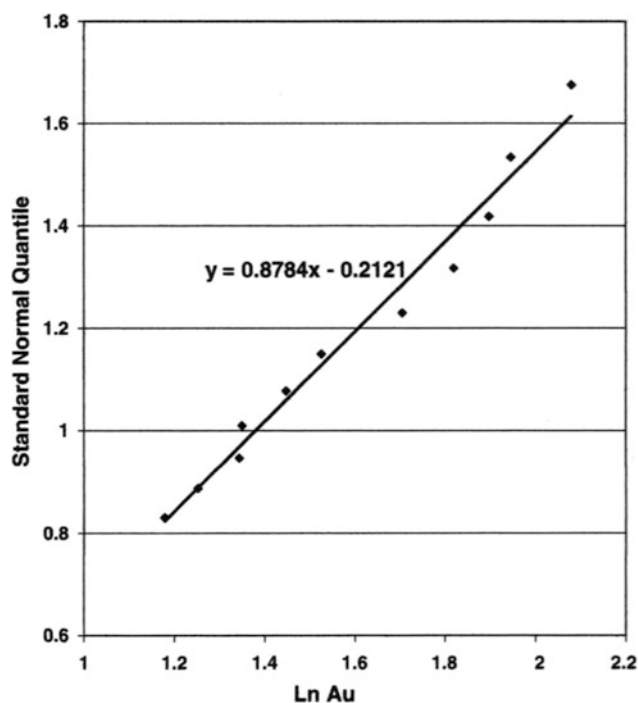
The approach described in the previous section is illustrated by application to gold content of till samples obtained during a systematic ultra-low density (1 site per 800 km²) geochemical reconnaissance survey in southern Saskatchewan (see Fig. 3). There are 389 observed gold concentration values ranging from below detection limit to 77 ppb. All Au values

below detection limit (= 2 ppb) were set equal to 1 ppb. The arithmetic mean gold concentration value is 2.5 ppb. In total, only about one-third of 389 observed values exceed this average. Overall, average gold value of 2.5 ppb slightly overestimates true mean gold concentration value because most gold values below 2 ppb are probably less than 1 ppb.

A lognormal Q - Q plot for Au is shown in Fig. 4 with its line of best fit. Because natural logarithms are used, the standard deviation of logarithmically transformed concentration values can be set equal to the inverse of the slope of the best-fitting straight line. It amounts to 2.37 ppb for Au. Original sampling design for the geochemical reconnaissance study in Southern Saskatchewan is described by Garrett and Thorleifson (1995) who selected an 80 × 80 km grid as the basic sampling cell, which was successively subdivided into 40 × 40 km, 20 × 20 km, and finally 10 × 10 km grid cells for sampling with randomly selected 1 × 1 km target



Chaos in Geosciences, Fig. 3 Study area of 64 cells with locations of till samples, Southern Saskatchewan. Unit of distance for east-west and north-south directions is 78.125 km. (Source: Agterberg 2007, Fig. 9)



Chaos in Geosciences, Fig. 4 Straight line fitted to part of lognormal $Q-Q$ plot for 64 average gold concentration values for cells. (Source: Agterberg 2007, Fig. 10)

cells. Compositional data for many other elements and minerals were determined as well. Gold was selected for this example because its frequency distribution seems to be unimodal, and Au had been used extensively in other multifractal studies (cf. Cheng 1994).

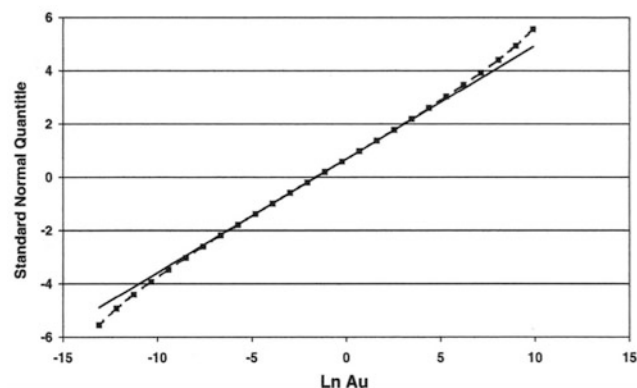
For the case study described here, a modified grid was used for Southern Saskatchewan resulting in a square-shaped study area consisting of 64 square cells measuring slightly less than 80 km on a side (see Fig. 3). The purpose of this redefinition of boundaries of study area was to minimize edge effects in the multifractal analysis. Data from outside the square study area were not used. In total, 290 of the original 389 (on average, about 4.5/cell) were retained. The Au logarithmic variance estimated from the 379 original till samples is 2.369.

de Wijs (1951) derived the following equation for the variance of logarithmically transformed element concentration values:

$$\sigma^2 = \frac{n}{4} (\ln \eta)^2$$

where σ^2 is the logarithmic variance (base e) and $\eta = (1 + d)/(1 - d)$. Application of this equation yields $n = 26.1$. This apparent maximum number of subdivisions of the environment would correspond to square cells measuring approximately 75 m, and 40 m on a side, respectively. Clearly, these cells are much larger than the very small areas where the till was actually sampled for the purpose of chemical analysis. It shows that the regional model of de Wijs for Au does not apply at the local sampling scale.

The log-binomial distribution for gold is shown in the lognormal $Q-Q$ plot of Fig. 5. Because the negative binomial is discrete, the preceding estimate of n was rounded off to the nearest integer. The parameters of the distribution shown are



Chaos in Geosciences, Fig. 5 Lognormal $Q-Q$ plots for discrete binomial frequency distributions of gold concentration values resulting from the model of the Wijs. Overall average Au concentration value $\xi = 1.82$ ppb, dispersion index $d = 0.433$, and apparent number of subdivisions of the environment $n = 26$. (Source: Agterberg 2007, Fig. 13)

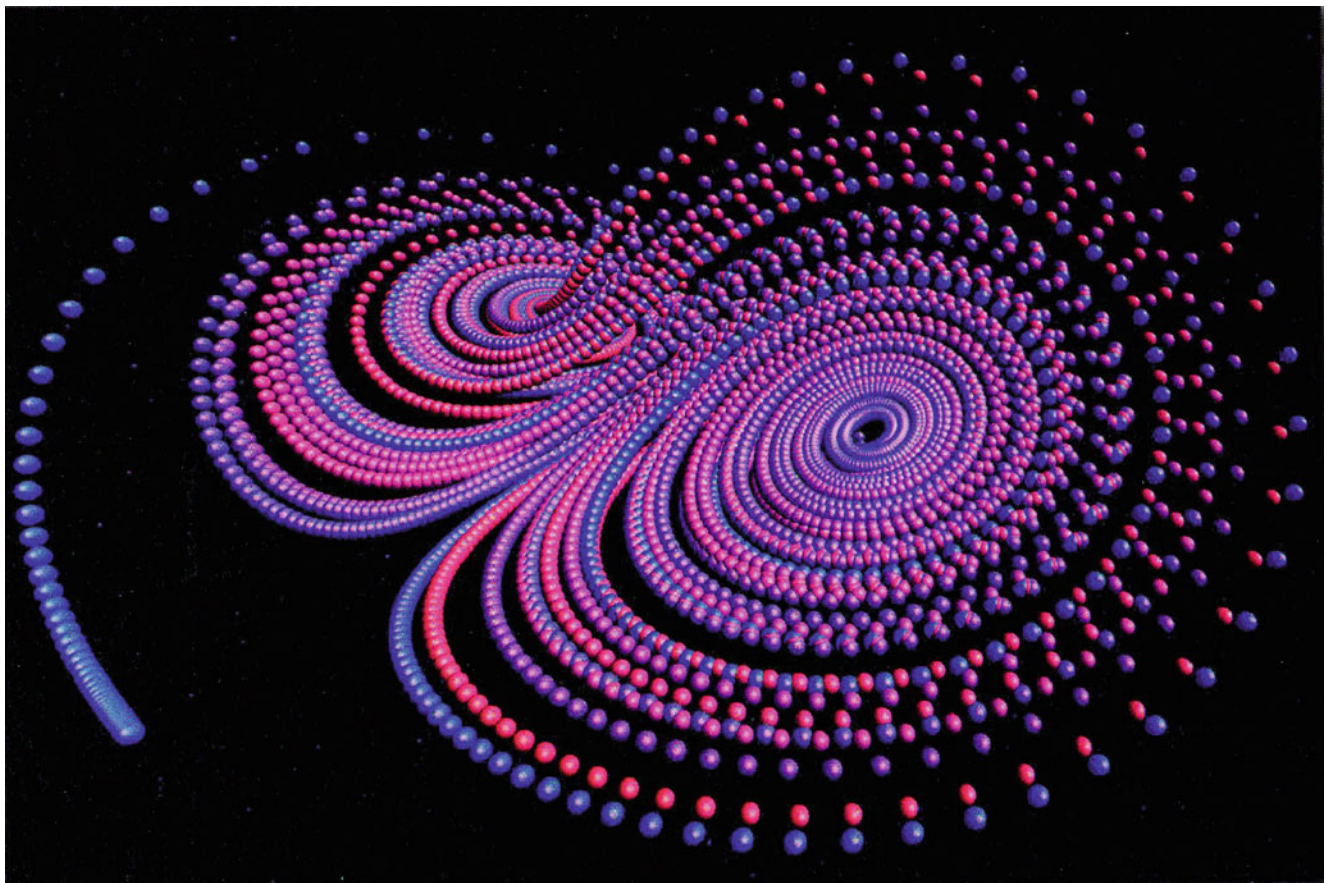
$\xi = 1.82$ ppb, $d = 0.433$, and $n = 26$ for gold (Fig. 5). The mean value ξ was determined by forcing the log-binomial and its lognormal approximation through the centers of clusters of points on lognormal Q - Q plots for original data. This result remains approximate but is interesting in that it permits shape comparison of the log-binomial and lognormal tails.

The Lorenz Attractor

It can be argued that nonlinear process modeling originated in 1963 when the meteorologist Edward Lorenz (1963) published the paper entitled “Deterministic nonperiodic flow” (cf. Motter and Campbell 2013). Earlier, he had accidentally discovered how miniscule changes in the initial state of a time series computer simulation experiment grew exponentially with time to yield results that did not resemble one another although the equations controlling the process were the same in all experiments (Lorenz 1963). It turned out that the notion of absolute predictability, although it had been assumed intuitively to hold true in classical physics, is in practice false for most systems. In the title of a talk Lorenz

gave in 1972, he aptly referred to this chaotic behavior as the “butterfly effect” asking if the flap of a butterfly in Brazil could set off a tornado in Texas (Fig. 6).

Lorenz (1963) presented chaos theory by using a three-variable system of nonlinear ordinary differential equations now known as the Lorenz equations. The model derives from a truncated Fourier expansion of the partial differential equations describing a thin, horizontal layer of fluid that is heated from below and cooled from above. The equations can be written as: $dx/dt = a(-x + y)$, $dy/dt = cx - y - xz$, and $dz/dt = -bz + xy$, where x represents intensity of convective motion, y is proportional to the temperature difference between the ascending and descending convective currents, and z indicates the deviation of the vertical temperature profile from linearity. The parameters b and c represent particulars of the flow geometry and rheology. Lorenz set them equal to $b = 8/3$ and $c = 10$, respectively, leaving only the Rayleigh number a to vary. For small c , the system has a stable fixed point at $x = y = z = 0$, corresponding to no convection. If $24.74 > c \geq 1$, the system has two symmetrical fixed points, representing two steady convective states. At $c = 24.74$, these two convective states lose stability; at $c = 28$, the system



Chaos in Geosciences, Fig. 6 The Lorenz attractor as revealed by the never-repeating trajectory of a single chaotic orbit (Motter and Campbell 2013). The spheres represent iterations of the Lorenz equations, calculated using the parameters originally used by Lorenz (1963). Spheres are

colored according to iteration count. The two lobes of the attractor resemble a butterfly, a coincidence that helped earn sensitive dependence on initial conditions. Hence, its nickname: “the butterfly effect.” (Source: Agterberg 2014, Fig. 12.9)

shows nonperiodic trajectories. Such trajectories orbit along a bounded region of 3-D space known as a chaotic attractor, never intersecting themselves (Fig. 6). For larger values of c , the Lorenz equations exhibit different behaviors that have been catalogued by Sparrow (1982).

Conceptionally, chaos is closely connected to fractals and multifractals. The attractor's geometry can be related to its fractal properties. For example, although Lorenz could not resolve this particular problem, his attractor has fractal dimension equal to approximately 2.06 (Motter and Campbell 2013, p. 30). An excellent introduction to chaos theory and its relation to fractals can be found in Turcotte (1997). The simplest nonlinear single differential equation illustrating some aspects of chaotic behavior is the logistic equation that can be written as $dx/dt = x(1 - x)$ where x and t are non-dimensional variables for population size and time, respectively. There are no parameters in this equation because characteristic time and representative population size were defined and used. The solution of this logistic equation has fixed points at $x = 0$ and 1, respectively. The fixed point at $x = 0$ is unstable in that solutions in its immediate vicinity diverge away from it. On the other hand, the fixed point at $x = 1$ has solutions in its immediate vicinity that are stable. Introducing the new variable, $x_1 = x - 1$, and neglecting the quadratic term, the logistic equation has the solution $x_1 = x_{10} \cdot e^{-t}$ where x_{10} is assumed to be small but constant. All such solutions "flow" toward $x = 1$ in time. They are not chaotic.

Chaotic solutions evolve in time with exponential sensitivity to the initial conditions. The so-called logistic map arises from the recursive relation $x_{k+1} = a \cdot x_k (1 - x_k)$ with iterations for $k = 0, 1, \dots$. May (1976) found that the resulting iterations have a remarkable range of behavior depending on the value of the constant a that is chosen. Turcotte (1997, Sect. 10.1) discusses in detail that there now are two fixed points at $x = 0$ and $1 - a^{-1}$, respectively. The fixed point at $x = 0$ is stable for $0 < a < 1$ and unstable for $a > 1$. The other fixed point is unstable for $0 < a < 1$, stable for $1 < a < 3$, and unstable for $a > 3$. At $a = 3$ a so-called flip bifurcation occurs. Both singular points are unstable and the iteration converges on an oscillating limit cycle. At $a = 3.449479$, another flip bifurcation occurs and there suddenly are four limit cycles. The approach can be generalized by defining successive constants a_i ($i = 1, 2, \dots, \infty$) leading to limit cycles that satisfy the iterative relation $a_{i+1} - a_i = F^{-1} (a_i - a_{i-1})$ where $F = 4.669202$ is the so-called Feigenbaum constant. Turcotte (1997, Figs. 10.1–10.6) shows examples of this approach. Sornette et al. (1991) and Dubois and Cheminée (1991) have treated the eruptions of the Piton de la Fournaise on Réunion Island and Mauna Lao and Kilauea in Hawaii using this generalized logistic model.

The preceding examples of deterministic processes result in chaotic results. However, one can ask the question of whether there exist deterministic processes that can fully explain fractals and multifractals? The power-law models related to fractals and multifractals can be partially explained on the basis of the concept of self-similarity. Various chaotic patterns observed for element concentration values in rocks and orebodies could be partially explained as the results of multiplicative cascade models such as the model of de Wijs. These models invoke random elements such as increases of element concentration that are not deterministic.

Successful applications of nonlinear modeling in geoscience also include the following: Rundle et al. (2003) showed that the Gutenberg-Richter frequency-magnitude relation is a combined effect of the geometrical (fractal) structure of fault networks and the nonlinear dynamics of seismicity. Most weather-related processes taking place in the atmosphere, including cloud formation and rainfall, are multifractal (Lovejoy and Schertzer 2007). Other space-related nonlinear processes include "current disruption" and "magnetic reconnection" scenarios (Uritsky et al. 2008). Within the solid Earth's crust, processes involving the release of large amounts of energy over very short intervals of time including earthquakes (Turcotte 1997), landslides (Park and Chi 2008), flooding (Gupta et al. 2007), and forest fires (Malamud et al. 1998) are nonlinear and result in fractals or multifractals.

Summary and Conclusions

Nonlinear process modeling is providing new clues to answers of where the randomness in nature comes from. From chaos theory it is known that otherwise deterministic Earth process models can contain terms that generate purely random responses. The solutions of such equations may contain unstable fixed points or bifurcations. It was shown that multifractals provide a novel way of approach to problem solving in situations where the attributes display strongly positively skewed frequency distributions. Randomness in the geosciences has been known to exist since the sixteenth century but it is only fairly recently that chaos theoretical explanations using deterministic differential equations are being developed. The results of chaos theory often are shown to possess fractal or multifractal properties. The model of de Wijs applied to gold concentration values in till was used as an example in this entry. It can be assumed that future developments of chaos theory will help to explain various forms of randomness and multifractals in the geosciences.

Cross-References

- Additive Logistic Normal Distribution
- De Wijs Model
- Frequency Distribution
- Lognormal Distribution
- Multifractals
- Self-Organizing Maps

Bibliography

- Agterberg FP (2007) New applications of the model of de Wijs in regional geochemistry. *Math Geosci* 39(1):1–26
- Agterberg FP (2014) *Geomathematics: theoretical foundations, applications and future developments*. Springer, Dordrecht
- Bak P (1996) *How nature works*. Copernicus, Springer, New York
- Beirlant J, Goegebeur Y, Segers J, Teugels J (2005) *Statistics of extremes*. Wiley, New York
- Cheng Q (1994) Multifractal modeling and spatial analysis with GIS: gold mineral potential estimation in the Mitchell-Sulphurets area, northwestern British Columbia. Unpublished doctoral dissertation, University of Ottawa
- Cheng Q, Agterberg FP (2009) Singularity analysis of ore-mineral and toxic trace elements in stream sediments. *Comput Geosci* 35: 234–244
- Cheng Q, Agterberg FP, Ballantyne SB (1994) The separation of geochemical anomalies from background by fractal methods. *J Geochem Explor* 51(2):109–130
- de Wijs HJ (1951) Statistics of ore distribution I. *Geol Mijnb* 13:365–375
- Dubois J, Cheminée H (1991) Fractal analysis of eruptive activity of some basaltic volcanoes. *J Volcanol Geotherm Res* 45:197–208
- Garrett RG, Thorleifson LH (1995) Kimberlite indicator mineral and till geochemical reconnaissance, southern Saskatchewan. Geological Survey of Canada Open File 3119, pp 227–253
- Gupta VK, Troutman B, Dawdy D (2007) Towards a nonlinear geophysical theory of floods in river networks: an overview of 20 years of progress. In: *Nonlinear dynamics in geosciences*. Springer, New York, pp 121–150
- Hacking I (2006) *The emergence of probability*, 2nd edn. Cambridge University Press, New York
- Lorenz EN (1963) Deterministic nonperiodic flow. *J Atmos Sci* 20: 130–141
- Lovejoy S, Schertzer D (2007) Scaling and multifractal fields in the solid earth and topography. *Nonlinear Process Geophys* 14:465–502
- Malamud BD, Morein G, Turcotte DL (1998) Forest fires: an example of self-organized critical behavior. *Science* 281:1840–1842
- Mandelbrot B (1989) Multifractal measures, especially for the geophysicist. *Pure Appl Geophys* 131:5–42
- May RM (1976) Simple mathematical models with very complicated dynamics. *Nature* 261:459–467
- Motter E, Campbell DK (2013) Chaos at fifty. *Physics Today*, May issue, pp 27–33
- Park NW, Chi KH (2008) Quantitative assessment of landslide susceptibility using high-resolution remote sensing data and a generalized additive model. *Int J Remote Sens* 29(1):247–264
- Poincaré H (1899) *Les méthodes nouvelles de la mécanique céleste*. Gauthier-Villars, Paris
- Rundle JB, Turcotte DL, Sheherbakov R, Klein W, Sammis C (2003) Statistical physics approach to understanding the multiscale dynamics of earthquake fault systems. *Rev Geophys* 41:1019. <https://doi.org/10.1029/2003RG000135>
- Sagar BSD, Murthy MBR, Radhakrishnan P (2003) Avalanches in numerically simulated sand dune dynamics. *Fractals* 11(2):183–193
- Schuster HG (1995) *Deterministic chaos*, 3rd edn. Wiley, Weinheim
- Sornette A, Dubois J, Cheminée JL, Sornette D (1991) Are sequences of volcanic eruptions deterministically chaotic. *J Geophys Res* 96(11): 931–945
- Sparrow C (1982) *The Lorenz equation: bifurcations, chaos, and strange attractors*. Springer, New York
- Turcotte DL (1997) *Fractals and chaos in geology and geophysics*, 2nd edn. Cambridge University Press, Cambridge, UK
- Uritsky VM, Donovan E, Klimas AJ (2008) Scale-free and scale-dependent modes of energy release dynamics in the night time magnetosphere. *Geophys Res Lett* 35(21):L21101, 1–5 pages
- Xie S, Bao Z (2004) Fractal and multifractal properties of geochemical fields. *Math Geol* 36(7):847–864
- Zipf GK (1949) *Human behavior and the principle of least effect*. Addison-Wesley, Cambridge, MA

Chayes, Felix

Richard J. Howarth
Department of Earth Sciences, University College London
(UCL), London, UK



Fig. 1 Felix Chayes in 1970. (Reproduced with permission of the Geophysical Laboratory, Carnegie Institution of Washington)

Biography

The quantitative petrographer (Charles) Felix Chayes was born in New York City, on 10 May 1916. He majored in geology at New York University (BA 1936), then moved to Columbia University, NY, to study the alkaline intrusives of Bancroft, Ontario, Canada (MA 1939; PhD 1942). Joining the US Bureau of Mines in 1942, he introduced grain-fragment counting to monitor the composition of quarried granite products.

In 1947 he joined the Geophysical Laboratory, Carnegie Institution, Washington, DC, and studied the compositional variation of fine-grained New England granites, using “modal analysis” (i.e., “point-counting” the proportions of mineral

grains in a thin-section; Chayes 1956) and undertook similar investigations elsewhere. Advised by statisticians Joseph Cameron (1922–2000) and William Youden (1900–1971) from the National Bureau of Standards (NBS), Washington, he tried to alert geologists to problems posed by data in which “the sum of the variates is itself a variable . . . [which] is actually, or practically, constant. The effect . . . is to generate negative correlation” (Chayes 1948, p. 415). He introduced the term “closed table” for one having a constant-sum property. Realizing that major-element geochemical data caused similar ratio-correlation problems which affected both petrographic attributes and derived CIPW- and Niggli-normative compositions, he repeatedly drew attention to these difficulties thereafter. Advised by statistician William Kruskal (1919–2005) of the University of Chicago, he developed a test for the significance of such spurious correlations, but it subsequently became apparent that there were problems with it.

In 1961, Spanish petrologist José Luis Brändle began work with Chayes to assemble a database of major-element analyses of Cenozoic volcanic rocks and in 1971, funded by the National Science Foundation, Chayes began to undertake data-retrieval requests. He became Chair of the International Union of Geological Sciences (IUGS) Subcommittee on Electronic Databases for Petrology in 1984.

With the advice of NBS statisticians Cameron, Calyampudi Rao (1920–) and Geoffrey Watson (1921–1998), Chayes and Danielle Velde (1936–; née Métais) were early users of discriminant analysis to distinguish between igneous rock types, based on their major-element chemical compositions (Chayes and Métais 1964), but later work revealed the inherent variability of such procedures when using small numerical sample sizes.

Chayes (1968) also used least-squares mixing-models to estimate the composition of the unexposed “hidden zone” of the Skaergaard layered intrusion, East Greenland.

Chayes and Brändle sought to enlarge their database to include both trace-element and petrographic data. This resulted in the IUGS “IGBADAT” database for igneous petrology and its support software; keeping it updated dominated Chayes’ later years – its fifth release in 1994 contained 19,519 rock analyses. After 1986, he worked part-time at the Geophysical Laboratory and at the National Museum of Natural History, Washington. He died on February 28, 1993, as the result of a car accident.

A Charter Member (1968–1993) of the International Association for Mathematical Geology (IAMG), he was awarded its William Christian Krumbein Medal in 1984. The IAMG’s Felix Chayes Prize for Excellence in Research in Mathematical Petrology was established in his memory in 1997. For more information, see Howarth (2004).

Bibliography

- Chayes F (1948) A petrographic criterion for the possible replacement origin of rocks. *Am J Sci* 246:413–425
- Chayes F (1956) *Petrographic modal analysis. An elementary statistical appraisal.* Wiley, New York, p 113
- Chayes F (1968) A least squares approximation for estimating the amounts of petrographic partition products. *Mineral Petrogr Acta* 14:111–114
- Chayes F, Métais D (1964) On distinguishing basaltic lavas of intra- and circum-oceanic types by means of discriminant functions. *Trans Am Geophys Union* 45:124
- Howarth RJ (2004) Not “Just a Petrographer”: The life and work of Felix Chayes (1916–1993). *Earth Sci Hist* 23:343–364

Cheng, Qiuming

Frits Agterberg
Central Canada Division, Geomathematics Section,
Geological Survey of Canada, Ottawa, ON, Canada



Fig. 1 Qiuming Cheng, courtesy of Prof. Cheng

Biography

Professor Qiuming Cheng was born in 1960 in Taigu County, China. He received a BSc in mathematics (1982) and MSc in mathematical geology (1985) at Changchun University of Earth Sciences where he was appointed Associate Professor. In 1991 he became a PhD student at the University of Ottawa. Within 3 years, he completed his doctoral thesis (Cheng 1994)

recognized by the university as its best doctoral dissertation for that year. The university's citation included the sentence "Dr. Cheng has demonstrated his potential as a world-class scientist, producing results that are important not only in the geosciences but also in physics and mathematics." This prediction was amply fulfilled. After a one-year post-doctorate fellowship at the Geological Survey of Canada in Ottawa, Dr. Cheng was awarded a professorship at York University with cross-appointment in the Department of Earth and Space Science and Department of Geography. In 2004 he also became a Changjiang Scholar in China and Director of the State Key Laboratory of Geological Processes and Mineral Resources (GPMR) which is part of the China University of Geosciences with campuses in Wuhan and Beijing. In 2017 he retired from York in order to perform research and teach fulltime at CUG Beijing. In 2019 he was elected to membership in the Chinese Academy of Sciences.

Professor Cheng's research focuses on the development of mathematical geocomplexity theory for modeling nonlinear geo-processes and quantitative prediction of mineral resources (see e.g., Cheng 2005). His pioneering work on new fractal theory and local singularity analysis (Cheng 2007) has made major impacts on several geoscientific sub-disciplines including those concerned with extreme geological events originated from nonlinear processes of plate tectonics such as the formation of supercontinents, mid-ocean ridge heat flow, earthquakes, and the formation of ore deposits (Cheng 2016). He has authored or co-authored over 300 refereed journal papers or book chapters and delivered over 100 invited papers or keynote lectures. In 1997 he obtained the President's Award of the International Association for Mathematical Geosciences (IAMG) given to scientists aged 35 or less. In 2008, the IAMG awarded him the William Christian Krumbein medal, its highest award for lifetime achievements in mathematical geoscience, service to the profession and contributions to the IAMG of which he became the 2012–2016 President. From 2016 to 2020, Professor Cheng was President of the International Union of Geological Sciences with over a million members in its affiliated organizations.

Bibliography

- Cheng Q (1994) Multifractal modeling and spatial analysis with GIS: gold mineral potential estimation in the Mitchell-Sulphurets area, northwestern British Columbia. Unpublished doctoral dissertation. University of Ottawa, Ottawa
- Cheng Q (1999) Spatial and scaling modelling for geochemical anomaly separation. *J Geochem Explor* 65:175–194
- Cheng Q (2005) A new model for incorporating spatial association and singularity in interpolation of exploratory data. In: Leuangthong D, Deutsch CV (eds) *Geostatistics Banff 2004*. Springer, Dordrecht, pp 1017–1025

- Cheng Q (2007) Mapping singularities with stream sediment geochemical data for prediction of undiscovered mineral deposits in Gejiu, Yunnan Province, China. *Ore Geol Rev* 32:314–324
- Cheng Q (2016) Fractal density and singularity analysis of heat flow over ocean ridges. *Sci Rep* 6:1–10

Circular Error Probability

Frits Agterberg
Central Canada Division, Geomathematics Section,
Geological Survey of Canada, Ottawa, ON, Canada

Definition

Circular error probability applies to the directions of linear features in two-dimensional space that are subject to uncertainty. Statistical analysis of directional data differs from ordinary statistical analysis based on the normal distribution. Linear features are either directed or undirected. For example, paleocurrents can result in features from which it can be seen in which direction the current was flowing or they result in symmetrical features from which the direction of flow cannot be determined. A sample of n features has values that can be written as x_i ($i = 1, 2, \dots, n$). The probability $P(x_i)$ of occurrence of x_i ranges from 0° to 360° (or from 0 to 2π). The system is closed because $x_i \geq 360^\circ = x_i - 360^\circ$. Estimates of average directions are affected by closure if relatively many measurements deviate significantly from the arithmetic mean; e.g., when the standard deviation of the sample significantly exceeds 30° .

Introduction

Various geological attributes may be approximated by lines or planes. Measuring them results in "angular data" consisting of azimuths for lines in the horizontal plane or azimuths and dips for lines in 3D space. Although a plane usually is represented by its strike and dip, it is fully determined by the line perpendicular to it, and statistical analyses of data sets for lines and planes are analogous. Examples of angular data are strike and dip of bedding, banding, and planes of schistosity, cleavage, fractures, or faults. Azimuth readings with or without dip are widely used for sedimentary features such as axes of elongated pebbles, ripple marks, foresets of cross-bedding and indicators of turbidity flow directions (sole markings). Then there are the B-lineations in tectonites. Problems at the microscopic level include that of finding the preferred orientation of crystals in a matrix (e.g., quartz axes in petrofabrics).

Initially, circular data in geoscience were treated like ordinary data not subject to the closure constraint (see, e.g., analysis of variance treatment by Potter and Olson (1954). However, increasingly 2D (and 3D) directional features were treated by methods that incorporated the closure constraint of vectorial data (Watson and Williams 1956; Stephens 1962). However, Agterberg and Briggs (1963) pointed out that using the normal distribution provides a good approximation if the standard deviation is less than 30 or if the range of observations does not exceed 114° (also see Agterberg 1974, Fig. 102). The following types of mean can be estimated:

AM (arithmetic mean): $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

VM (vector mean): $\hat{\gamma} = \arctan \frac{\sum_{i=1}^n \sin x_i}{\sum_{i=1}^n \cos x_i}$

SVM (semicircular vector mean): $\hat{\gamma}_2 = \frac{1}{2} \arctan \frac{\sum_{i=1}^n \sin 2x_i}{\sum_{i=1}^n \cos 2x_i}$

VM and SVM apply to directed and undirected features, respectively.

There is a close connection between 2D circular and 3D spherical statistics. The features taken, for example, in this chapter are observed in 2D outcrops in the field. In the first example, they are approximately in 2D, because the rock layers containing them are subhorizontal. In the second example, they are 3D but can be regarded as 2D because they resulted from geological processes that took place at significantly different times, during the Hercynian and Alpine orogenies, respectively. In a so-called marked point process, the marks are values at the points (c.f. Stoyan and Stoyan 1996 or Cressie 1991). Thus, on maps showing azimuths and dips of

features at observation points for the second example the dips could be regarded as marks of the azimuths, or vice versa.

The body of publications dealing with statistics of orientation data is large. Introductory papers on unit vectors in the plane include Rao and Sengupta (1970). Statistical theory for the treatment of unit vectors in 2D (Fisher 1993) and 3D (Fisher et al. 1987) is well developed. The comprehensive textbook by Mardia (1972) covers both 2D and 3D applications. Pewsey and Garcia-Portugués (2020) is a more recent review pointing out that <https://git.ub.com/egarpor/DirectionalStatsbib> contains over 1700 references to publications on statistics of 2D and 3D directional features.

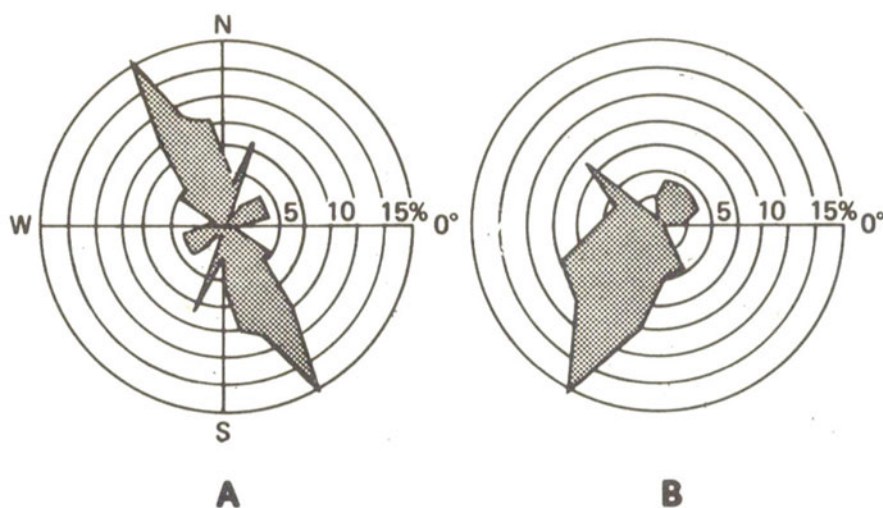
Doubling the Angles

It is useful to plot angular data sets under study in a diagram before statistical analysis is undertaken. Azimuth readings can be plotted on various types of rose diagrams (Potter and Pettijohn 1963). If the lines are directed, the azimuths can be plotted from the center of a circle and data within the same class intervals may be aggregated. If the lines are undirected, a rose diagram with axial symmetry can be used such as the one shown in Fig. 1. Then the method of Krumbein (1939) can be used in which the average is taken after doubling the angles and dividing the resulting mean angle by two. In Agterberg (1974), problems of this type and their solutions are discussed in more detail. Krumbein's solution is identical to fitting a major axis to the points of intersection of the original measurements with a circle. Axial symmetry is preserved when the major axis is obtained.

The three types of mean unit vector listed previously correspond to the following three types of frequency distribution:

Circular Error Probability,

Fig. 1 Axial symmetric rose diagram of the direction of pebbles in glacial drift in Sweden according to Köster (1964). (Source: Agterberg 1974, Fig. 99)



AM: normal (Gaussian) with frequency distribution $f_n(\theta) =$

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\theta^2}{2\sigma^2}\right)$$

VM: circular normal (von Mises) with $f(\theta) =$

$$\frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cdot \cos \theta)$$

SVM: semicircular normal with $f_2(\theta) =$

$$\frac{1}{\pi I_0(\kappa_2)} \exp(\kappa_2 \cdot \cos \theta)$$

In these expressions, θ (previously denoted as x) represents the angle measured clockwise from the pole (usually the north direction); $I_0(\dots)$ is the Bessel function of the first kind of pure imaginary argument.

Bjorne Formation Paleodelta Example

The Bjorne Formation is a predominantly sandy unit of Early Triassic age. It was developed at the margin of the Sverdrup Islands of the Canadian Arctic Archipelago. On northwestern Melville Island, the Bjorne Formation consists of three separate members which can be distinguished, mainly on the basis of clay content which is the lowest in the upper member, *C*. The total thickness of the Bjorne Formation does not exceed 165 m on Melville Island. The formation forms a prograding fan-shaped delta. The paleocurrent directions are indicated by such features as the dip azimuths of planar foresets and axes of spoon-shaped troughs (Agterberg et al. 1967). The average current direction for 43 localities of Member *C* is shown in Fig. 2a. These localities occur in a narrow belt where the sandstone member is exposed at the surface. The azimuth of the paleocurrents changes along the belt. The attitude of sandstone layers is subhorizontal in the entire study area.

The variation pattern (Fig. 3a) shows many local irregularities but is characterized by a linear trend. In the (U, V) plane for the coordinates (see Fig. 3b), a linear trend surface can be written as $x = F(u, v) = 292.133 + 27.859u + 19.932v$ (degrees). At each point on the map, x represents the tangent of a curve for the paleocurrent trend that passes through that point, or $\frac{dv}{du} = -\tan x$. It is readily shown that: $du = \frac{dx}{a_1 - a_2 \tan x}$ where $a_1 = 27.859$ and $a_2 = 19.932$. Integration of both sides gives:

$$u = C + \frac{1}{a_1^2 + a_2^2} [a_1 x - a_2 \log_e(a_1 \cos x + y \sin x)]$$

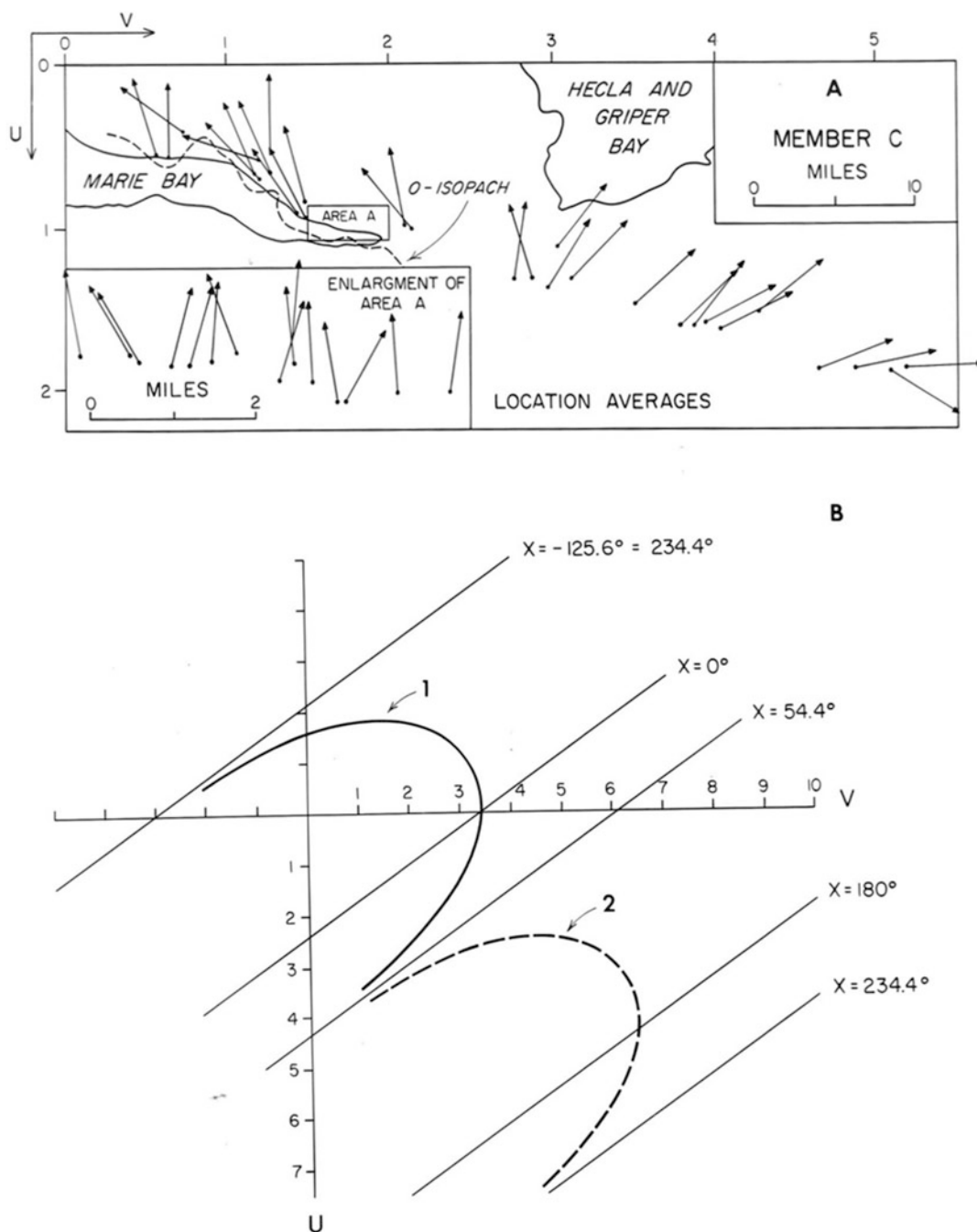
with $y = a_2$ if $\tan x > \frac{a_1}{a_2}$ and $y = -a_2$ if $\tan x < \frac{a_1}{a_2}$. The result is applicable when x is expressed in degrees. When $x \geq 360^\circ$, the quantity 360° may be subtracted from x , because azimuths are periodic with period 360° . The constant C is arbitrary. A value can be assigned to it by inserting specific values for u and v into the equation. For example, if

$u = 0$ and $x = 0$, $C = -0.7018$. It can be used to calculate a set of values for u forming a sequence of values for x ; the corresponding values for v then follow from the original equation for the linear trend surface. The resulting curves (1) and (2) are shown in Fig. 3b. If the value of C is changed, the curves (1) and (2) become displaced in the $x = 54.4^\circ$ direction. Thus, a set of curves is created that represents the paleocurrent direction for all points in the study area.

Suppose that the paleocurrents were flowing in directions perpendicular to the average topographic contours at the time of sedimentation of the sand. If curve (1) of Fig. 3b is moved in the 144.4° direction over a distance that corresponds to 90° in x , the result represents a set of directions which are perpendicular to the paleocurrent trends. Four of these curves which may represent the shape of the delta are shown in Fig. 4c. These contours, which are labeled *a*, *b*, *c*, and *d* satisfy the preceding equation for u with different values of C and with x replaced by $(x + 90^\circ)$ or $(x - 90^\circ)$. Definite values cannot be assigned to these contours because x represents a direction and not a vector with both direction and magnitude.

In trend surface analysis, the linear trend surface (also see Fig. 4a) has explained sum of squares $ESS = 78\%$. The complete quadratic and cubic surfaces have ESS of 80% and 84%, respectively. Analysis of variance for the step from linear to quadratic surface resulted in $\widehat{F}(3, 37) = \frac{(80-78)/3}{(100-80)/37} = 1.04$. This would correspond to $F_{0.60}(3, 37)$ showing that the improvement in fit is not statistically significant. It is tacitly assumed that the residuals are not autocorrelated as suggested by their scatter around the straight line of Fig. 3. Consequently, the linear trend surface as shown in Fig. 4a is acceptable in this situation. The 95% confidence interval for this linear surface is shown in Fig. 4b. Suppose that X is the matrix representing the input data. This is a so-called half-confidence interval with values equal to $[(p+1) \cdot F \cdot s^2(\bar{Y})]$ with $F = F_{0.95}(3, 40) = 2.84$ and $s^2(\bar{Y}) = s^2 \dot{X}_k (\dot{X}X)^{-1} X_k$ with residual variance $s^2 = 380$ square degrees.

All flow lines in Fig. 4a converge to a single line-shaped source. They have a single asymptote in common which suggests the location of a river. An independent method for locating the source of the sand consists of mapping the grade of the largest clasts contained in the sandstone at a given place. Four grades of large clasts could be mapped in the area of study. They are: (1) pebbles, granule, coarse sand; (2) cobbles, pebbles; (3) boulders, cobbles; and (4) boulders (max. 60 cm). The size of the clasts is larger where the velocity of the currents was higher. Approximate grade contours for classes 1–4 are shown in Fig. 4c for comparison. This pattern corresponds to the contours constructed for the paleodelta.



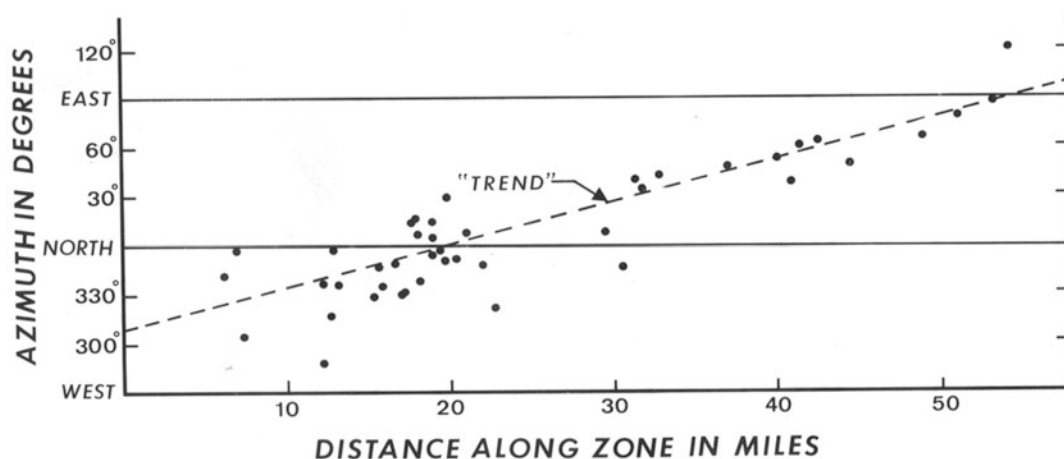
Circular Error Probability, Fig. 2 Variation of preferred paleocurrent direction along straight line coinciding with surface outcrop of Member C, Bjorne Formation (Triassic, Melville Island, Northwest

Territories, Canada); systematic change in azimuth is represented by trend line. (Source: Agterberg 1974, Fig. 12)

Directed and Undirected Unit Vectors

Directional features play an important role in structural geology. Basic principles were originally reviewed and developed by Sander (1948) who associated a three-dimensional (A, B, C) Cartesian coordinate system with folds and other structures. A-axis and C-axis were defined to be parallel to the directions of compression and expansion, respectively, with the B-axis

perpendicular to the AC-plane representing the main plane of motion of the rock particles. Two examples of Hercynian minor folds with clearly developed B-axes are shown in Fig. 5 (from Agterberg 1961) for quartzphyllites belonging to the crystalline basement of the Dolomites in northern Italy. Measurements on B-axes from these quartzphyllites were analyzed previously (Agterberg 1961, 1974, 2012) and are again used here.



Circular Error Probability, Fig. 3 Reconstruction of approximate preferred paleocurrent direction and shape of paleodelta from preferred current directions, Member C, Bjorne Formation, Melville Island. (After Agterberg et al. 1967). (a) Preferred directions of paleocurrent indicators

at measurement stations; 0-isopach applies to lower part of Member C only. (b) Graphical representation of solution of differential equation for paleocurrent trends. Inferred approximate direction of paleoriver near its mouth is N54.4 °E. (Source: Agterberg 1974, Fig. 13)

Stereographic projection using either the Wulff net or the equal-area Schmidt net (see, e.g., Agterberg 1974) continue to be useful tools for representing sets of directional features from different outcrops within the same neighbourhood or for directions derived from crystals in thin sections of rocks. Contouring on the net is often applied to find maxima representing preferred orientations, which also can be estimated using methods developed by mathematical statisticians, especially for relatively small sample sizes (see, e.g., Fisher et al. 1987).

Reiche (1938) used the vector mean to find the preferred orientation of directional features. Using a paleomagnetic data set, Fisher (1953) further developed this method for estimating the mean orientation from a sample of directed unit vectors. The frequency distribution of unit vectors may satisfy the so-called Fisher distribution $f(\theta, \phi) = \frac{\kappa}{4\pi \cdot \sinh \kappa} \exp(\kappa \cdot \cos \theta)$, which is the spherical equivalent of a two-dimensional normal distribution.

TRANSALP Profile Example

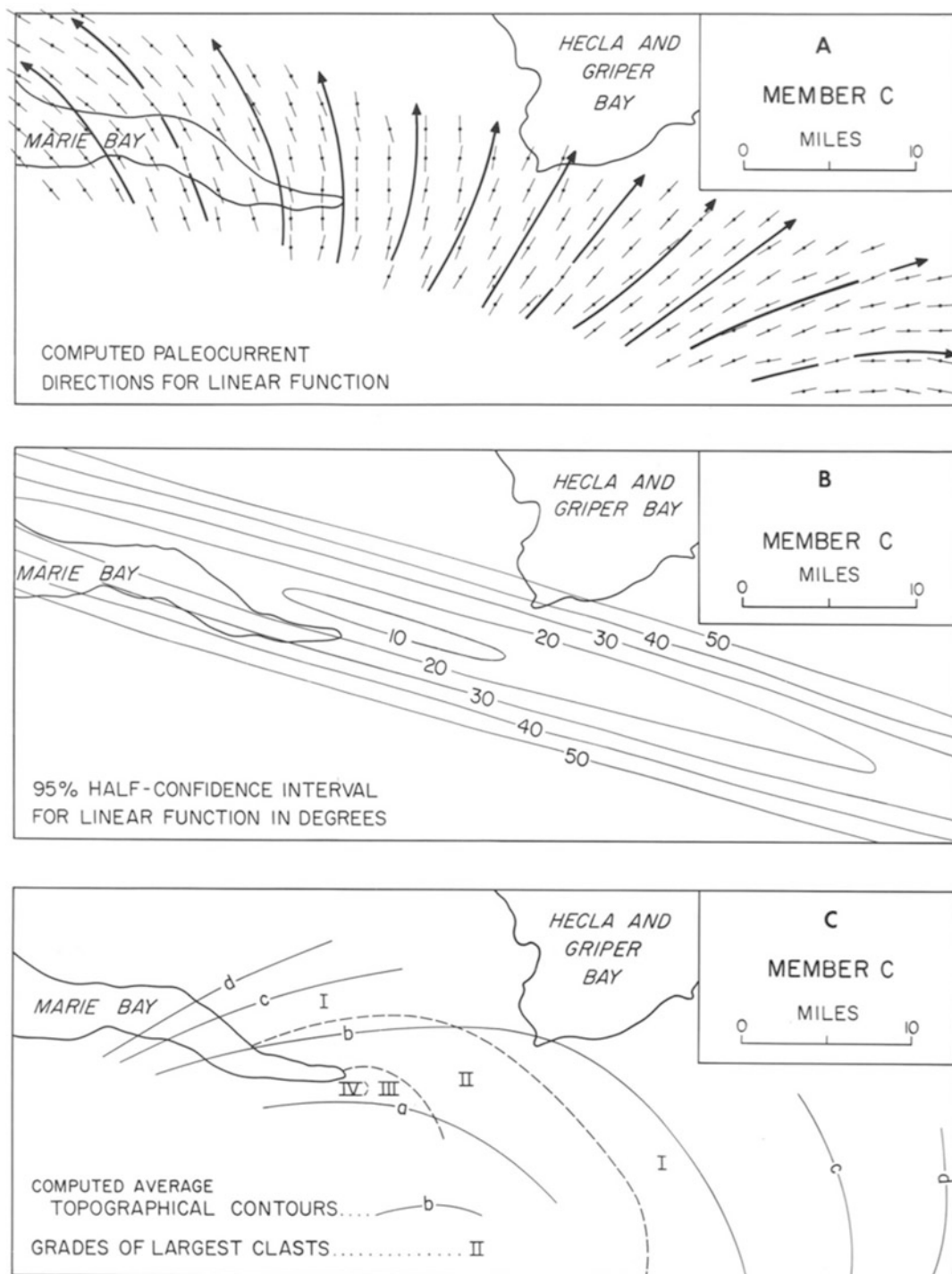
Figure 6 (based on map by Agterberg 1961) shows mean B-axes and schistosity planes in quartzphyllites belonging to the crystalline basement of the Dolomites in an area around the Gaderbach, which was part of the seismic Vibroseis and explosive transects of the TRANSALP Transect. The TRANSALP profile was oriented approximately along a north-south line across the Eastern Alps from Munich to Belluno with locally its position determined by irregular shapes of the topography (see, e.g., Gebrande et al. 2006).

Structurally, the quartzphyllites are of two types: in some outcrops the Hercynian schistosity planes are not strongly

folded and their average orientation can be measured. However, elsewhere, like in the examples of Fig. 6 and in most of the Pustertal to the east of Bruneck, there are relatively many minor folds (as illustrated in Fig. 5) of which the orientation can be measured. However, then it may not be possible to measure representative strike and dip of schistosity at outcrop scale. In the parts of the area where the strike and dip of schistosity planes can be determined, there usually are B-lineations for microfolds on the schistosity planes that also can be measured. Consequently, the number of possible measurements on B-axes usually exceeds the number of representative measurements on schistosity planes.

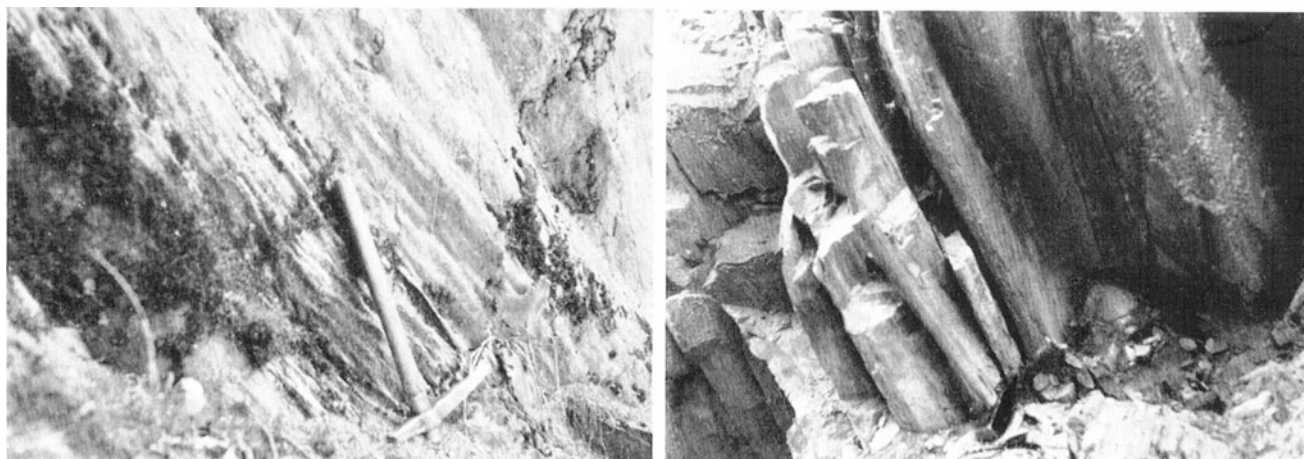
Figure 6 shows that the average strike of schistosity is fairly constant in this area but the average azimuth and dip of the B-axes is more variable. There is a trend from intermediate eastward dip of B-axes in the southern parts of the region to subvertical attitude in the north. This systematic change can also be seen on Schmidt net plots for the six measurement samples along the Gaderbach (Agterberg 1961, Appendix III). In Fig. 7, the B-axes measured along the Gaderbach are projected onto a North-South line approximately parallel to the average orientation of the TRANSALP profile.

In a general way, the results of the preceding linear unit vector field analysis confirm the earlier conclusions on North-South regional change in average B-lineation attitude. Original estimates of mean B-axis orientations of which some are shown in Fig. 6 were obtained by separately averaging the azimuths and dips. Table 1 shows results from another test. In Fig. 7 such average values for the six measurement samples (Method 1) are compared with ordinary unit vector means (Method 2) at mean distances south of the (Periadriatic) Pusteria Lineament near Bruneck. Differences between the two sets of mean azimuth and dip values are at most a few degrees indicating that regionally, on the average, there is a



Circular Error Probability, Fig. 4 (a) Linear trend for data of Fig. 3a. Computed azimuths are shown by line segments for points on regular grid, both inside and outside exposed parts of Member C; flow lines are based on curves 1 and 2 of (c). (b) Contour map of 95% half-confidence interval for A; confidence is greater for area supported by observation points. (c) Estimated topographic contours for delta obtained by shifting

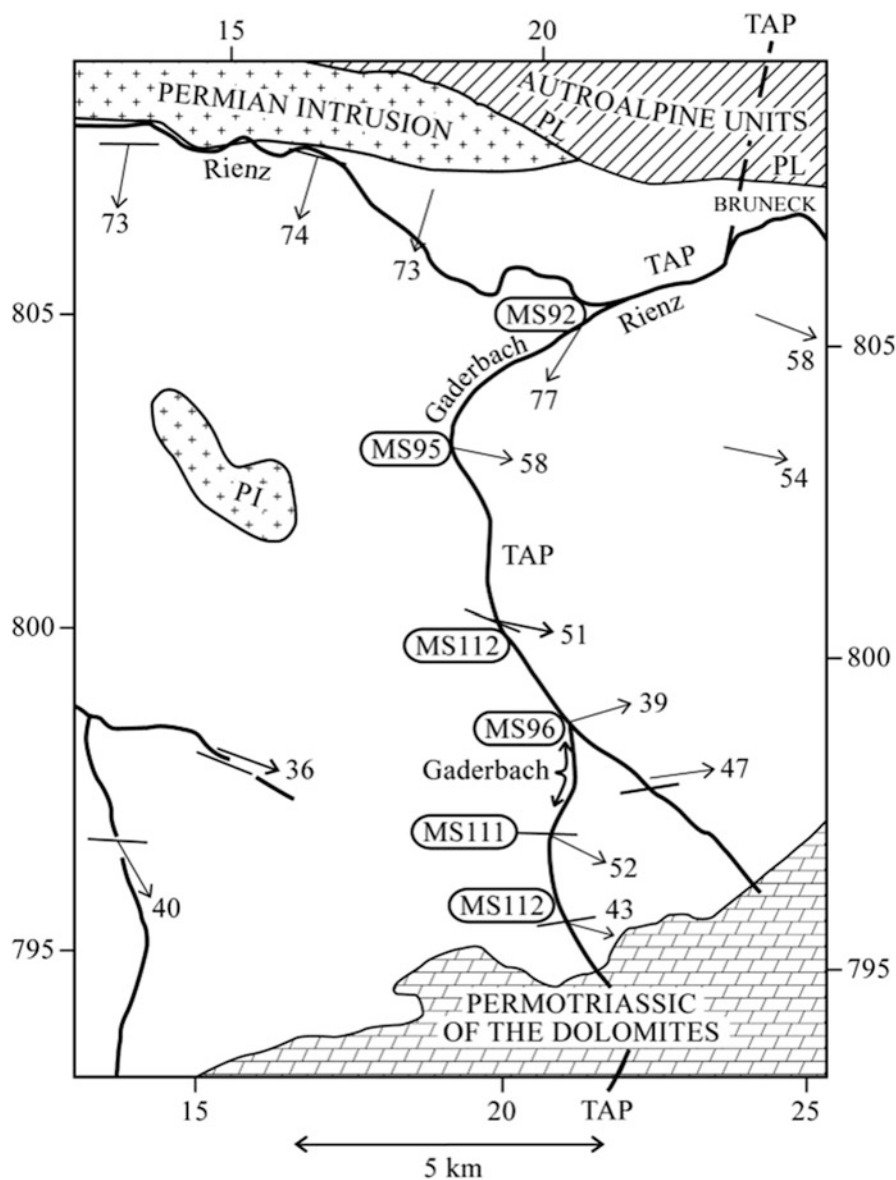
curve 1 of Fig. 3b, 90° in the southeastern direction (= perpendicular to isoazimuth lines), and then moving it into four arbitrary positions by changing the constant C. Contoured grades of largest clasts provide independent information on shape of delta. (Source: Agterberg 1974, Fig. 41)

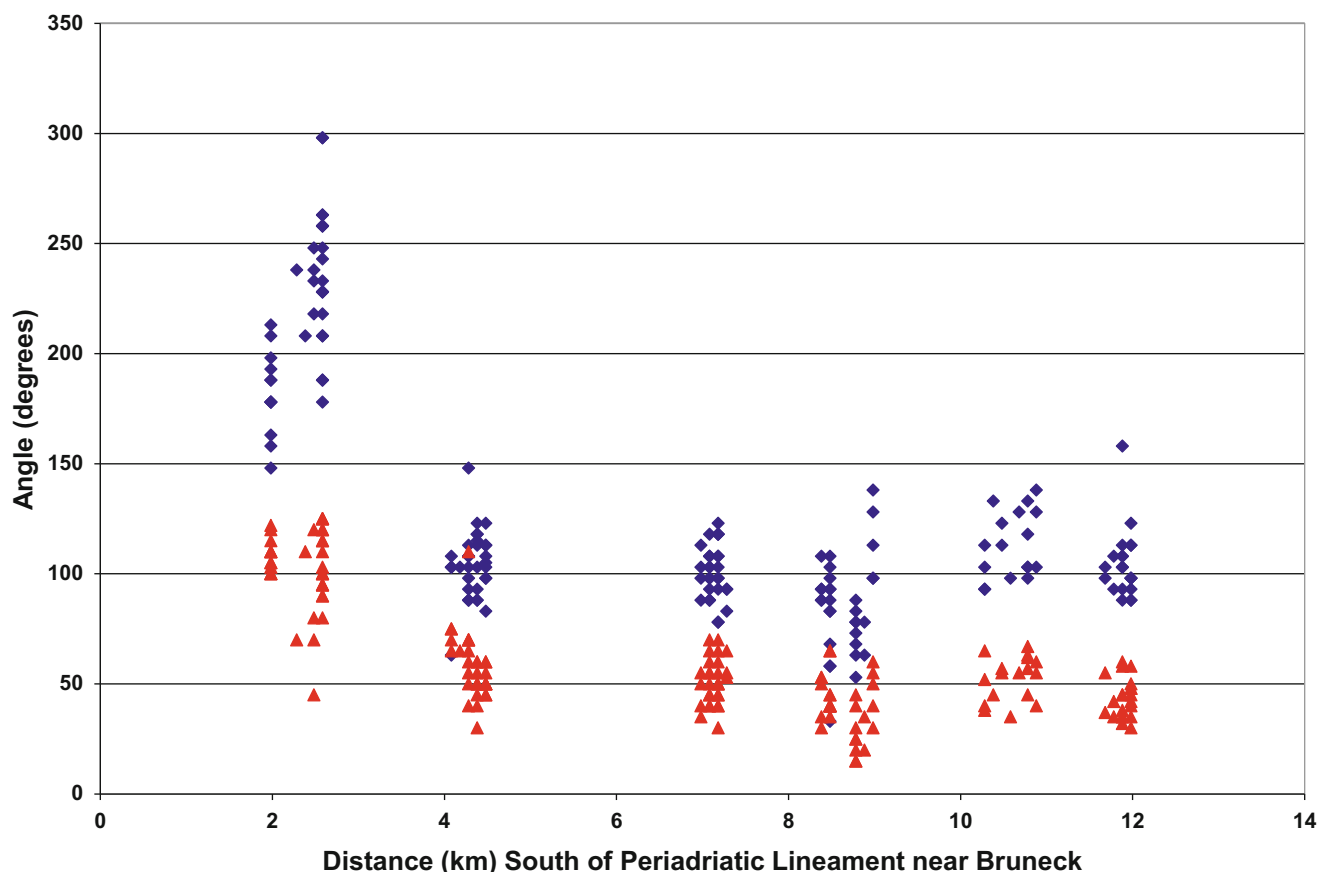


Circular Error Probability, Fig. 5 Two examples of minor folds in the Pustertal quartzphyllite belt. Left side: East-dipping minor folds near Welsberg (Monguelfo). Right side: ditto along TRANSALP profile near

subvertical contact with Permotriassic (grid coordinates 206–767; part of MS112, see Fig. 8). (Source: Agterberg 2012, Fig. 1)

Circular Error Probability, Fig. 6 Schematic structure map of surroundings of TRANSALP profile (TAP) in tectonites of crystalline basement of the Dolomites near Bruneck. Arrows represent average azimuths of B-axes for measurement samples. Average dips of B-axes are also shown. Average *s*-plane attitudes are shown if they could be determined with sufficient precision. MS: Measurement Sample as in Agterberg (1961, Appendix II). PI: Permian Intrusion; PL: Periadriatic Lineament. Coordinates of grid are as on 1:25,000 Italian topographic maps. (Source: Agterberg 2012, Fig. 5)





Circular Error Probability, Fig. 7 Azimuths (diamonds) and dips (triangles) of B-axes from six measurement samples along Gaderbach shown in Fig. 6. Note that along most of this section, the B-axes dip

nearly 50° East. Closest to the Periadriatic Lineament, their dip becomes nearly 90° SSW (also see Table 1). (Source: Agterberg 2012, Fig. 6)

Circular Error Probability, Table 1 Comparison of average orientations obtained by two methods. *MS*: measurement sample; distance in km south of Insubric Line; *Method 1*: average azimuth/dip; *Method 2*: Fisher azimuth/dip; Azimuth and dip are given in degrees

MS #	Distance	Method 1	Method 2
92	2.38	217/77	220/78
95	4.38	104/58	103/58
112	7.08	101/51	100/52
96	8.68	86/39	85/41
111	10.58	113/52	113/53
112	11.88	104/43	102/44

dip rotation from about 45° dip to the east in the south to about 80° dip in the north accompanied by an azimuth rotation of about 120° from eastward to WNW-ward.

Widespread occurrence of steeply dipping B-axes near the Periadriatic Lineament and along the southern border of the Permian (Brixen granodiorite) intrusion, however, is confirmed by results derived from other measurement samples. The rapid change in average azimuth that accompanies the B-axis steepening is probably real but it should be kept in mind that the azimuth of vertical B-axes becomes

indeterminate. The northward steepening of attitudes of *s*-planes in the northern quartzphyllite belt to the West of Bruneck from gentle southward dip near Brixen is probably due to neo-Alpine shortening. The northward steepening of the B-axes near Bruneck, which on average are contained within the *s*-planes, could be due to Neo-Alpine sinistral movements along the Periadriatic Lineament and south of the Permian (Brixen granodiorite) intrusion.

Summary and Conclusions

Circular error probability is associated with two-dimensional directional features that are observed in different types of rock. Examples in sedimentary rocks include cross-bedding, ripple marks and features associated with turbidity currents. Crystals in igneous rocks may display preferred orientations. Microfolds and minor folds in structural geology provide other examples. Different procedures must be used for directed and undirected features, especially in sedimentology. Elongated pebbles in Sweden were used to illustrate one type of statistical method to treat undirected features. There is a

close association between spherical and circular error probability. Directional cross-bedding in sandstones of the Early Triassic Bjorne Formation in the Canadian Arctic was used to reconstruct the shape of a paleodelta with approximate location of the river associated with it. In structural geology the preferred orientations of minor folds in the Hercynian basement of the Dolomites in Northern Italy were taken as an example. Although these features are three-dimensional, their azimuths and dips can be analyzed separately by using a marked 2D model. The Hercynian minor folds were subjected to rotations during the Alpine orogeny.

Cross-References

► Trend Surface Analysis

Bibliography

- Agterberg FP (1961) Tectonics of the crystalline basement of the Dolomites in North Italy. Kemink, Utrecht, p 232
- Agterberg FP (1974) Geomathematics – mathematical background and geo-science applications. Elsevier, Amsterdam, p 596
- Agterberg FP (2012) Unit vector field fitting in structural geology. *Zeitschr D Geol Wis* 40(4/6):197–211
- Agterberg FP, Briggs G (1963) Statistical analysis of ripple marks in Atokan and Desmoinesian rocks in the Arkoma Basin of east-central Oklahoma. *J Sediment Petrol* 33:393–410
- Agterberg FP, Hills LV, Trettin HP (1967) Paleocurrent trend analysis of a delta in the Bjorne Formation (Lower Triassic) of northeastern Melville Island, Arctic Archipelago. *J Sediment Petrol* 37:852–862
- Cressie NAC (1991) Statistics for spatial data. Wiley, New York, p 900
- Fisher RA (1953) Dispersion on a sphere. *Proc R Soc Lond Ser A* 217:295–305
- Fisher NI (1993) Statistical analysis of circular data. Cambridge University Press, Cambridge, UK, p 277
- Fisher NI, Lewis T, Embleton BJJ (1987) Statistical analysis of spherical data. Cambridge University Press, Cambridge, UK/New York, p 316
- Gebrande H, Castellarin A, Lüschen E, Millahn K, Neubauer F, Nicolich R (2006) TRANSALP – a transect through a young collisional orogen: introduction. *Tectonophysics* 414:1–7
- Köster E (1964) Granulometrische und Morphometrische Messmethoden an Mineralkörnern, Steinen und sonstigen Stoffen. Enke, Stuttgart, p 336
- Krumbein WC (1939) Preferred orientations of pebbles in sedimentary deposits. *J Geol* 47:673–706
- Mardia KV (1972) Statistics of directional data, Probability and mathematical statistics. Academic, London, p 358
- Pewsey A, Garcia-Portugués E (2020) Recent advances in directional statistics. *Test* 30:1–61
- Potter PE, Olson JS (1954) Variance components of cross-bedding direction in some basal Pennsylvanian sandstones of the Eastern Interior Basin. *J Geol* 62:50–73
- Potter PE, Pettijohn FJ (1963) Paleocurrents and basin analysis. Academic, New York, p 296
- Rao JS, Sengupta S (1970) An optimum hierarchical sampling procedure for cross-bedding data. *J Geol* 78(5):533–544
- Reiche P (1938) An analysis of cross-lamination of the Coconino sandstone. *J Geol* 46:905–932
- Sander B (1948) Einführung in die Gefügekunde der geologischen Körper, vol I. Springer, Vienna, p 215
- Stephens MA (1962) The statistics of directions – the Fisher and von Mises distributions. Unpublished PhD thesis, University of Toronto, Toronto, p 211
- Stoyan D, Stoyan H (1996) Fractals, random shapes and point fields. Wiley, Chichester, p 389
- Watson GS, Williams EJ (1956) On the construction of significance tests on the circle and sphere. *Biometrika* 43:344–352

Cloud Computing and Cloud Service

Liping Di and Ziheng Sun

Center for Spatial Information Science and Systems, George Mason University, Fairfax, VA, USA

Synonyms

API: application programming interface; AWS: Amazon Web Service; CPU: central processing unit; DevOps: development and operations; EC2: Elastic Computing Cloud; GPU: graphical processing unit; IRIS: Incorporated Research Institutions for Seismology; IT: information technology; PC: personal computer; RAID: redundant array of independent disks; REST: representational state transfer; S3: Simple Storage Service; SSH: Secure Shell

Definitions

- Cloud computing:** A technique that virtualizes the IT resources and delivers them to users over the Internet with pay-as-you-go pricing and on-demand basis.
- Cloud service:** The commodity services delivered by cloud computing vendors to end-point users, including the storage, server, application, and networking.

Background

The exploding volume of Earth observations is driving scientists to seek for better schemes of computing and storage. Large-scale and high-performance computing is more necessary in operational use today than 10 years ago. Many institutes and commercial companies are working intensively to provide such computing capabilities and make them available

to scientists. Cloud has become one of the major computing resource providers in geosciences (Hashem et al. 2015). This entry reviews the existing cloud platforms, cloud-native services, and the schemes of using cloud resources in geoscientific research. This entry offers insights on how to correctly leverage clouds in geoscientific research and how the clouds and geosciences should thrive together in the future.

What Is Cloud Computing?

The leading cloud provider, Amazon Web Service, explains it in this way: “*Cloud computing is the on-demand delivery of IT resources over the Internet with pay-as-you-go pricing. Instead of buying, owning, and maintaining physical data centers and servers, you can access technology services, such as computing power, storage, and databases, on an as-needed basis from a cloud provider*” (<https://aws.amazon.com/what-is-cloud-computing/>). The definition shows two major reasons that are backing the requirements for cloud computing: server costs and on-demand needs. For example, if a geoscientist wants to serve a valuable dataset to the public, in old days the scientist will have to buy a server and plant it in a data center and pay the data center for the costs on electricity, network, air conditioning, room space, racks, and on-site staff (Dillon et al. 2010). Meanwhile, all the risks are on the scientist, such as disk failures, network interruption, vulnerability fix, Internet attack, etc. In a while, the scientist will find the research team in an awkward situation working significantly on disk replacing, network testing, and vulnerability fixing. As the owner of the server, they need to pay a lot to survive the server online or off-line. Even worse, many servers in universities’ data centers are never running on 100% of capability. Most of the computing hours are wasted with servers running idle, which is a big waste of the research resources and results in a serious unbalance between research hours and server maintenance hours.

To reverse the unbalance, cloud computing was proposed, implemented, and instantly gained popularity in server administrators. Compared to the old routine of server-hosting businesses, *cloud computing provides a new paradigm by adding a virtualization layer on top of the physical hardware layer*. The *virtualization layer* is powered by the virtual machine technique which can create multiple new virtual machines from one physical machine. The virtual machines use the host’s CPU, memory, and disks, but running separately. From the user perspective, the virtual machines have no differences from a real physical machine. Users can do everything they did on a physical machine on the virtual machines. In cloud computing, each virtual machine is called an instance. In cloud computing, users only interact with instances and do not have access to the physical hardware level. Cloud data centers host tens of thousands of servers,

link them into cloud clusters, sell the computing hours to the public by creating separated virtualized instances, and provide web-based interfaces or remote-messaging channels (such as SSH and RESTful API) for the users to manipulate the instances remotely (users normally cannot access the data centers physically). The bills will be calculated based on the occupied computing hours and the instance type. The rates (cost per hour) vary across the types (configuration of CPU, GPU, memory, storage, and network) and the location of the data centers. For example, the rate for AWS EC2 m5.xlarge at US East (N.Va) data center is \$0.192 per hour, the rate for the same instance at Asia Pacific (Sydney) data center is \$0.24 per hour (<https://aws.amazon.com/emr/pricing/>). Overall, cloud computing is a scheme in which the computing resources are pooled together and disseminated in the form of virtual machines over the Internet to reduce the labor and financial burdens on both the owners and users of the IT resources.

Thinking of the computing resources as commodities, the cloud providers are sellers and the users are consumers. **Cloud services** denote all the service commodities provided by cloud platforms. To better commercialize the clouds, cloud providers have made many innovative developments to customize their services to fit in users’ scenarios and further reduce the workloads on the user side. This customization categorizes cloud computing into several types (Armbrust et al. 2010):

- **Infrastructure as a Service (IaaS)** (Bhardwaj et al. 2010) offers complete instance machines including all computing resources including data storage, servers, applications, and networking. Users are responsible for managing all four kinds of resources. Examples: AWS EC2, Google Compute Engine, Digital Ocean, Microsoft Azure, and Rackspace.
- **Platform as a Service (PaaS)** (Pahl 2015) offers a programming language execution environment associated with a database. It encapsulates the environment where users can build, compile, and run their programs without worrying about the underlying infrastructure. In this service model, users only manage software and data, and the other resources are managed by the vendors. Examples: AWS Elastic Beanstalk, Engine Yard, and Heroku.
- **Software as a Service (SaaS)** (Cusumano 2010) offers pay per use of application software to users. Users do not need to install software on PC and access the software via a web browser or lightweight client applications. Examples: Google ecosystem Apps (Gmail, Google Docs), Microsoft Office365, Salesforce, Dropbox, Cisco WebEx, Citrix GoToMeeting, and Workday.

There are many more types, such as Artificial Intelligence as a Service, Hadoop as a Service, Storage as a Service, Network as a Service, Backend as a Service, Blockchain as a Service, Database as a Service, Knowledge as a Service,

etc. (Tan et al. 2016). Users should consult cloud experts to analyze which cloud computing model is the most suitable for their use cases. An appropriate selection would save significantly in costs while gaining more from the investments.

Using Cloud in Geosciences

Geoscientists on Cloud

The paradigm of cloud computing makes it possible for scientists to only buy necessary computing hours and storage space instead of owning entire real servers, and thereafter get rid of all the hassles in server maintenance. With cloud computing, the scientist could worry less about the tedious technical issues, and pay for the resources that are used on demand and avoid the constant waste of idle servers. It is an evolution in scientific research efficiency and boosts the utilization ratio of the computing resources. Besides, it is easier to scale up the scientific models to a larger extent on big data, which was impossible before for many scientists in the old routine of computing schemes. Similar to the other sharing economy schemes, cloud computing takes advantage of its vast number of users to average the huge hardware maintenance costs and lower the bills on each user.

Geoscientific Use Cases

Geosciences have numerous study cases that meet the scenario requirements of cloud computing, e.g., earthquakes, landslides, land cover land use, volcano, rainfall, groundwater, snow, ice, climate change, hurricane, agriculture, wildfire, air quality, etc. Each research domain has undergone many years of observation activities and accumulated a lot of datasets. For example, IRIS (incorporated research institutions for seismology) Data Management Center (DMC) has operated a public repository of seismological data of three decades. IRIS offers a wide and growing variety of services that Earth scientists in over 150 countries worldwide reply on through web services. IRIS DMC archive is nearing one petabyte in volume. The funding of IRIS is mainly coming from the National Science Foundation, and the annual cost is around \$30 million (<https://ds.iris.edu/ds/newsletter/vol20/no2/498/geoscicloud-exploring-the-potential-for-hosting-a-geoscience-data-center-in-the-cloud/>). A significant amount is spent on maintaining the data repository to serve all the downloading requests from all over the world. Cloud computing is considered as a promising solution to significantly lower the data-hosting costs and put more funding on data-collection and data-mining activities.

Cloud-Based Data Storage

Most use cases of cloud in geosciences focus on storage, such as storing spatial data that can be accessed remotely. For big spatial data, cloud storage provides a more elastic, efficient,

robust storage service than conventional storage plans. The disks on the servers are consumable products that will certainly fail after reading and writing data onto the disks for a certain number of times (e.g., tens of million times). The standard warranty of a normal hard disk is about 5 years. RAID (Redundant Array of Independent Disks) is a common practice to ensure data security when some disks reach the end of life. RAID creates a pool of disks and stores the data scattered among all the disks instead of relying solely on one disk. Once a disk is detected with alert status, the data on it will be automatically transferred to other healthy disks. Some RAID (e.g., RAID-5) will store all the data duplicated, so when some disk regions are damaged, it can recover the data from the other copy. However, RAID will run through all the disks in every I/O operation which significantly decreases the average life of the disks. RAID needs to replace the disks in alert status as quickly as possible, otherwise, all the data will be lost if the percentage of failed disks reaches a certain ratio. Scientists might be exhausted to do the routine RAID check. Cloud data centers have on-site engineers who can do the disk check and replacement. Data is safer in the cloud than in individual small data centers that are less attended.

Cloud-Based Data Analysis

With the continuous efforts by the generations of scientists working on deploying sensors and building data repositories, geoscientists have many direct channels to get the required data to support their research. The fact today is that *geoscientists do not lack data in most times*. The real missing piece is the capability to allow scientists to manipulate the data in a time-wise and cost-wise manner. Artificial intelligence and machine learning algorithms are being used in geospatial analysis and run against spatial data to automate the processing and reveal insights (Sun et al. 2020). Spatial analysis, such as network analysis, clustering, pattern extraction, classification, and predictive modeling, poses more intelligence for information extraction. However, intelligent algorithms come with higher complexity which requires a more efficient solution to process the big amount of data. For example, it takes about 10 min for a PC to do a buffer process on the road network of Washington D.C. but takes weeks to buffer the road network of the entire world. The time cost increases at an exponential rate over the volume of the input datasets. Big data processing is one of the major applications of clouds. Cloud can create multiple instances that work simultaneously on the same task. Each instance processes one portion of the dataset (Map), and a reduce process will summarize the results from the instances into one result (Reduce). This parallel processing scheme is called MapReduce. Typical Map Reduce software includes the most widely used Apache Hadoop, Apache Spark, and Google Earth Engine (GEE) (Gorelick et al. 2017). For example, GEE is a cloud-based platform for planetary-scale geospatial analysis that brings

Google massive computational capabilities to process tremendous amounts of remote-sensing data products to help scientists to study deforestation, drought, disaster, disease, food security, ocean, water management, climate, and environment protection. These cloud-based big data-processing platforms have been heavily utilized and driven a series of significant discoveries in geoscience communities recently.

Challenges and Opportunities

Turning Geoscientific Applications to Cloud Native

Cloud native is an approach to building and running applications that are small, independent, and loosely coupled. Cloud-native applications are purposefully built for the cloud model to take advantage of the full benefits of clouds to enable resilience and flexibility of the application. Cloud-native app development typically includes DevOps, Agile methodology, microservices, cloud platforms, containers such as Kubernetes and Docker, and continuous delivery. However, the existing applications of geosciences are not cloud native. Most of them need wrapping and reprogramming, which may cause a series of compatibility issues between the traditional application environment and the container environment. Geoscientific applications are large programs that are not suitable to be packaged as microservices and need a breakdown. Geoscientific workflows should be itemized into several small-granule processes and run on the clouds with scalability and elasticity (Sun et al. 2019).

Managing Cloud Expenses

In most cases, cloud computing could save money. Instead of a onetime payment, the cloud cost is recurring based on usage. For applications with big data and a lot of users, more cloud resources will be used to cope with the huge data volume and the increasing requests. Correspondingly, the cost will rise as well and might make the entire bill unaffordable. For short-lived or small-scale projects, it is expensive to transfer all the data into public clouds. Using clouds needs to spend optimization which aims to generate the largest possible saving with predictive analytics and actionable resource purchasing recommendations. It is financially wise to automatically identify idle resources, unused instances, and shut them down to save costs. The cloud platforms usually provide cost optimization options by financial analytics and governance policies for automatically releasing the unused resources and decreasing the costs.

Data Security and Privacy

This is the top concern of most users on the cloud. Uploading data into the cloud feels like turning in the control over the data to a third party. This is true from a certain perspective, as

the data is physically transferred from users' machines to some machines owned by another institute. But from an engineering perspective, the cloud-based storage is more secure than local disks. Besides the common credential protection (password), the state-of-the-art cloud storage uses professional security methods such as firewalls, traffic monitoring, event logging, encryption, and physical security to keep the data private even though they are on a publicly accessible server. Another benefit of using the cloud is easy updating to fix vulnerabilities. Users do not need to worry about occasionally upgrading their PC systems to fix the newly discovered vulnerabilities. It is also smoother to apply the new security technology provided by cloud vendors on the fly. On the other side, users must follow the vendors' security guidelines, and negligent use could compromise even the best protection.

Governance

In industries, cloud computing governance creates business-driven policies and principles that establish the appropriate degree of investments and control around the life-cycle process for cloud services (http://www.opengroup.org/cloud/gov_snapshot/p3.htm). It ensures that all the sponsored activities are aligned with the project objective. For geosciences, one fundamental measure in governance is whether the activity directly contributes to problem-solving. The priority should be categorized based on the job significance in the entire workflow. In addition, the governance of the cloud is also an implementation of project management policies. The management models and standards for team members and the end-point users are going to be enforced in the form of a cloud computing governance framework. Coordinating the team members to work simultaneously on the cloud is another challenge facing the geoscientists and also an important research area for cloud vendors to innovate and advance their cloud services.

Private Cloud

In reality, many scientists will not use the public clouds out of various reasons (cost and trust); however, they would like to use the cloud features within their institutes. A private cloud would be the answer. Building a cloud is similar to building a cluster or supercomputer, all of which need multiple servers (more than one), including one management server and several client/slave servers. All the servers are connected via high-speed networks to pass messages among them. Operating systems of the member servers are mostly Linux-based, even in Microsoft Azure, because of its better support for virtualization hypervisor (software used for creating virtual machines). The cloud platform software includes OpenStack, CloudStack, Eucalyptus, and OpenNebula. Among them, OpenStack is the choice of most industrial companies, e.g.,

RedHat, IBM, Rackspace, Ubuntu, SUSE, etc. The installation of OpenStack is complicated. For scientific teams, it is better to consult cloud professionals before installing the servers.

Summary

This entry introduces the concept of cloud computing and the situation of using clouds in geosciences. Cloud computing has been proven as an effective solution to lift the bar for scientists to access and process the big geoscientific datasets. More and more use cases of cloud computing in geosciences are expected soon. There are several associated challenges with the transition from conventional computing paradigm to cloud, including transforming the old geoscientific models into cloud-native applications, managing cloud expenses over time, ensuring data security and privacy within the cloud data centers, coordinating the team efforts to work simultaneously on the cloud, and creating private clouds for scientists to integrate computing resources and serve the users within the institutions. Cloud vendors could take on these challenges to advance their cloud services to further extend their collaboration with geoscientists and assist them to solve critical scientific problems that are insolvable before.

Bibliography

- Armbrust M, Fox A, Griffith R, Joseph AD, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A, Stoica I (2010) A view of cloud computing. *Commun ACM* 53:50–58
- Bhardwaj S, Jain L, Jain S (2010) Cloud computing: a study of infrastructure as a service (IAAS). *Int J Eng Inf Technol* 2:60–63
- Cusumano M (2010) Cloud computing and SaaS as new computing platforms. *Commun ACM* 53:27–29
- Dillon T, Wu C, Chang E (2010) Cloud computing: issues and challenges. In: *Proceedings of 24th IEEE international conference on Advanced Information Networking and Applications (AINA)*, pp 27–33
- Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R (2017) Google Earth Engine: planetary-scale geospatial analysis for everyone. *Remote Sens Environ* 202:18–27
- Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Khan SU (2015) The rise of “big data” on cloud computing: review and open research issues. *Inf Syst* 47:98–115
- Pahl C (2015) Containerization and the paas cloud. *IEEE Cloud Comput* 2:24–31
- Sun Z, Di L, Cash B, Gaigalas J (2019) Advanced cyberinfrastructure for intercomparison and validation of climate models. *Environ Model Softw* 123:104559
- Sun Z, Di L, Burgess A, Tullis JA, Magill AB (2020) Geoweaver: advanced cyberinfrastructure for managing hybrid geoscientific AI workflows. *ISPRS Int J Geo Inf* 9:119
- Tan X, Di L, Deng M, Huang F, Ye X, Sha Z, Sun Z, Gong W, Shao Y, Huang C (2016) Agent-as-a-service-based geospatial service aggregation in the cloud: a case study of flood response. *Environ Model Softw* 84:210–225

Cluster Analysis and Classification

Antonella Buccianti and Caterina Gozzi
Department of Earth Sciences, University of Florence,
Florence, Italy

Definition

Cluster analysis and classification – a family of data analysis methods that attempt to find natural groups clearly separated from each other, thus giving indications about potential classification rules to be exploited in discriminant analysis procedures.

Historical Material

Our numerical world produces masses of data every day in earth sciences as well as in several other fields of investigation. The analysis of this big available data offers the possibility to exploit unknown paths leading to important advances in knowledge. However, the management of high-dimensional data, with the aim to extract as much information as possible, needs adequate methods and procedures. Cluster analysis and classification is one of the most interesting fields of statistical analysis both as explorative tool of the data grouping structures and model-based techniques for classification.

The aim of cluster analysis is to find meaningful groups in data internally coherent and well separated from others. The members of the groups joint some common features in the multivariate space, and this peculiarity is enhanced by using appropriate similarity/dissimilarity measures. The grouping of cases in cohesive groups dates back to Aristotle in his *History of Animals* in 342 B.C., while in natural and taxonomical sciences, a typical example is referred to Linnaeus (1735) for plant and animals. The systematic use of numerical methods in searching for groups in quantitative data refers to anthropological studies (Czekanowski 1909) where the measures of similarity between objects began to be the focus. In the 1950s, several advancements were proposed by considering the way of linking cases in a hierarchical procedure of aggregation whose final graphical result was the dendrogram (a tree diagram that shows the hierarchical relationships between objects). In the 1960s, a turning point is represented by the publication of the book of Sokal and Sneath (1963), and some recent interesting developments concern the model-based clustering where each group is modeled by its own probability distribution, thus obtaining a finite mixture model (Bouveyron et al. 2019). The application of cluster analysis for compositional data has also been discussed

considering the constrained working sample space (Aitchison 1986) and the implications to manage distances inside it (Templ et al. 2008). Compositional data, in fact, describe parts of the same whole and are commonly presented as vectors of proportions, percentages, or concentrations. Since they are expressed as real numbers, people are tempted to interpret or analyze them as real multivariate data, leading to paradoxes and misinterpretations.

When applied to geological and environmental data, cluster analysis can be used to group the variables, with the aim to highlight the correlation structure of the data matrix or to cluster cases into homogeneous subsets for subsequent analysis (e.g., discriminant analysis or mapping): these two different ways are known in literature as R-mode and Q-mode, respectively.

Essential Concepts and Applications

In cluster analysis, there are basically two different paths noting if the case is allocated to just one cluster (hard clustering) or if it is distributed among several clusters by considering a certain degree of association (fuzzy clustering, Dumitrescu et al. 2000).

The input of most of the hierarchical clustering algorithms is a distance matrix (distances between observations) $n \times n$ generated from an input original matrix $n \times p$ (n = number of cases, p = number of variables). Many different distance measures can be considered as a measure of similarity between cases in the multivariate space. An interesting deepening can be found in Shirkhorshidi et al. (2015).

The Minkowski family includes the Euclidean and Manhattan distances, which are particular cases of the Minkowski distance. The latter is defined by:

$$d_{\min} = \left(\sum_{i=1}^n |x_i - y_i|^m \right)^{1/m} \quad (1)$$

with $m \geq 1$ and x_i and y_i two vectors in n -dimensional space. The distance performs well when the data clusters are isolated or compacted, otherwise the large-scale attributes dominate the others. The Manhattan distance is a special case of the Minkowski one at $m = 1$ and like the previous one is sensitive to outliers. The Euclidean distance is a special case of the Minkowski one when $m = 2$ and similarly to this one performs well in the same conditions. It is very popular in cluster analysis, but it has some drawbacks. In fact, even if two data vectors have no attribute values in common, it might occur that they have a smaller distance than the other pair of data vectors containing the same attribute values. Another problem with Euclidean distance, as a family of the Minkowski metric, is that the largest-scaled feature could

dominate the others. Data normalization is a common solution to overcome this problem.

Average distance is a modified version of the Euclidean distance developed to try to solve the previous problems. For two data points x and y in a n -dimensional space, the average distance is defined as:

$$d_{\text{ave}} = \left(\frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2 \right)^{1/2} \quad (2)$$

The weighted Euclidean distance is instead defined as:

$$d_{\text{we}} = \left(\sum_{i=1}^n w_i (x_i - y_i)^2 \right)^{1/2} \quad (3)$$

where w_i is the weight given to the i th component to be used if the information of the relative importance according to each attribute is available.

Chord distance was proposed to overcome the problems of the scale of measurements and the critical issues of the Euclidean one. It is defined as:

$$d_{\text{chord}} = \left(2 - 2 \frac{\sum_{i=1}^n x_i y_i}{\|x\|_2 \|y\|_2} \right)^{1/2}, \quad (4)$$

where $\|x\|_2$ is the L^2 norm $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$.

Mahalanobis distance, in contrast with the previous ones, is a data-driven measure, and it is used to extract hyper-ellipsoidal clusters tone down distortion caused by linear correlation among features. It is defined by:

$$d_{\text{mah}} = \sqrt{(x - y) S^{-1} (x - y)^T} \quad (5)$$

where S is the covariance matrix of the data set.

The cosine similarity measure is given by:

$$\cos_{xy} = \frac{\sum_{i=1}^n x_i y_i}{\|x\|_2 \|y\|_2} \quad (6)$$

where $\|y\|_2$ is the Euclidean norm of vector $y = (y_1, y_2, \dots, y_n)$ defined as $\|y\|_2 = \sqrt{y_1^2 + y_2^2 + \dots + y_n^2}$. This measure is invariant to rotation but not to linear transformations. It is also independent of the vector length.

When the target of the analysis is the clustering of variables, a similarity matrix $p \times p$ will be generated by the $n \times p$ original one, by choosing a correlation measure.

The Pearson correlation coefficient is widely used, even if it is very sensible to outliers; it is defined by:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^n (x_i - \mu_x)^2} \sqrt{\sum_{i=1}^n (y_i - \mu_y)^2}} \quad (7)$$

where μ_x and μ_y are the means for x and y , respectively.

Agglomerative techniques start considering that each sample forms its own cluster then enlarging the clusters stepwise with new contributions, thus linking the most similar groups at each step. The process ends when there is one single cluster containing all the cases. On the other hand, divisive clustering starts with one cluster containing all samples and then successively splits this into groups in subsequent steps. This reverse procedure generally requires intensive calculus.

To link clusters in the agglomerative procedure, several different methods exist. The best known are average linkage, complete linkage, and single linkage (Kaufman and Rousseeuw 2005). The average linkage method considers the averages of all pairs of distances between the cases of two clusters. The two clusters with the minimum average distance are combined into one new cluster. Complete linkage

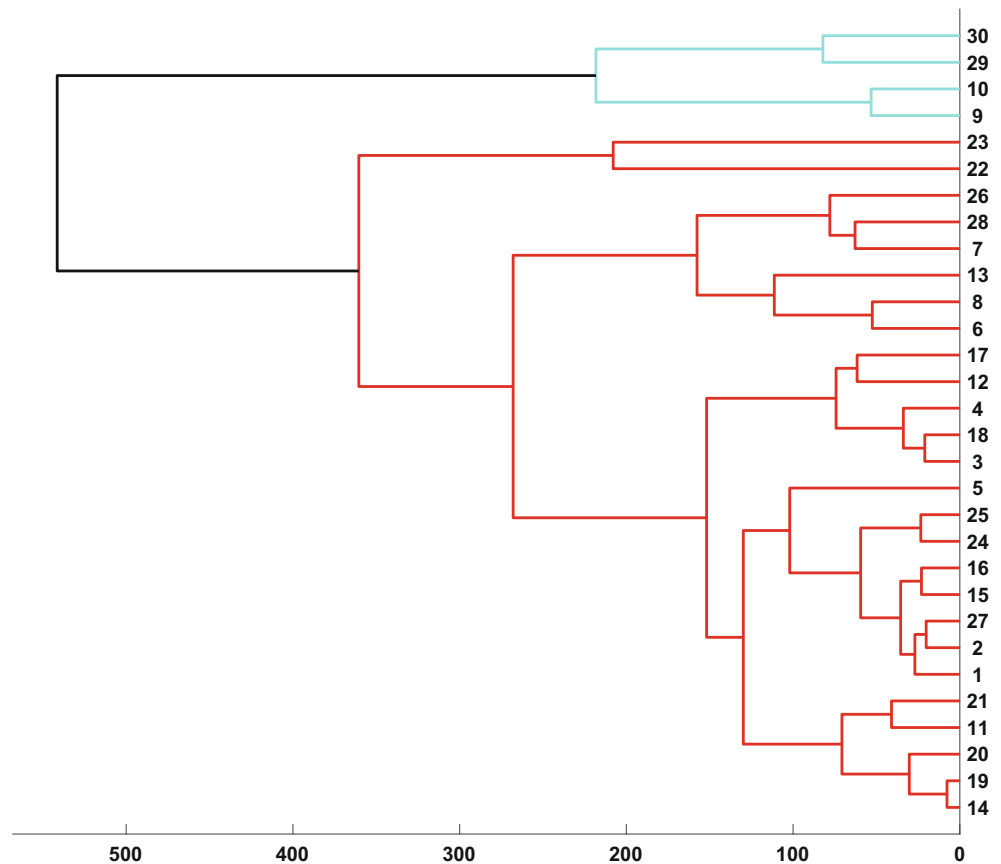
(or farthest-neighbor method) looks for the maximum distance between the samples of two clusters. The clusters with the smallest maximum distance are combined. Single linkage (or nearest-neighbor method) considers the minimum distance between all samples of two clusters. The clusters with the smallest minimum distance are thus linked.

The dendrograms obtained with the previous linking methods will show some differences. The single linkage will give cluster chains, complete linkage very homogeneous clusters in the early stages of agglomeration, and small resulting clusters, average linkage a compromise solution between the two mentioned methods. The Ward method (Kaufman and Rousseeuw 2005) is an alternative to the single link procedure able to create compact clusters, even if computationally intensive. Instead of measuring the distance directly, it monitors the variance of clusters during the merging process, keeping the growth as small as possible, by using the within-group sum of squares as a measure of homogeneity.

In the dendrogram, the horizontal line indicates the link between cases in clusters, whereas the vertical axis is related to the similarity, that is, to the chosen measure of distance/correlation. The axes can be inverted without problems. The linking of two groups at a large height indicates strong dissimilarity and vice versa. If the dendrogram is cut at a given

Cluster Analysis and Classification,

Fig. 1 Dendrogram for the chemical composition of water data sampled in a sedimentary aquifer (Minkowski distance and complete linking method)



height, the assignation of the cases to a distinct cluster can be obtained and, eventually, their spatial distribution can be visualized.

To demonstrate the procedure, a set of samples defining the chemical composition of an aquifer mainly developed in a sedimentary context has been chosen. The $n \times p$ basic matrix was given by $n = 30$ cases and $p = 7$ variables, representing the main chemical components: Na^+ , K^+ , Ca^{2+} , Mg^{2+} , HCO_3^- , Cl^- , and SO_4^{2-} , all expressed in mg/L. The dendrograms based on the cited anions and cations are reported in Figs. 1, 2, and 3 for the similarity matrix $n \times n$ considering the same similarity metric (Minkowski) but different linkage methods (complete, single, and average). The numbers on the vertical and horizontal axes are related to the label of the cases and to the measure of closeness of either individual data points or clusters.

The figures highlight the differences in using different linkage methods due to the hierarchical way used to construct the clusters and to associate cases to them, step by step, as previously reported. However, some groups of data appear to be well characterized and homogeneous, while others (e.g., 22 and 23, 9, 10, 29 and 30) need more inspection to verify the origin of their isolated behavior, as reported in Table 1. At this stage of the investigation, the possibility to map data will help

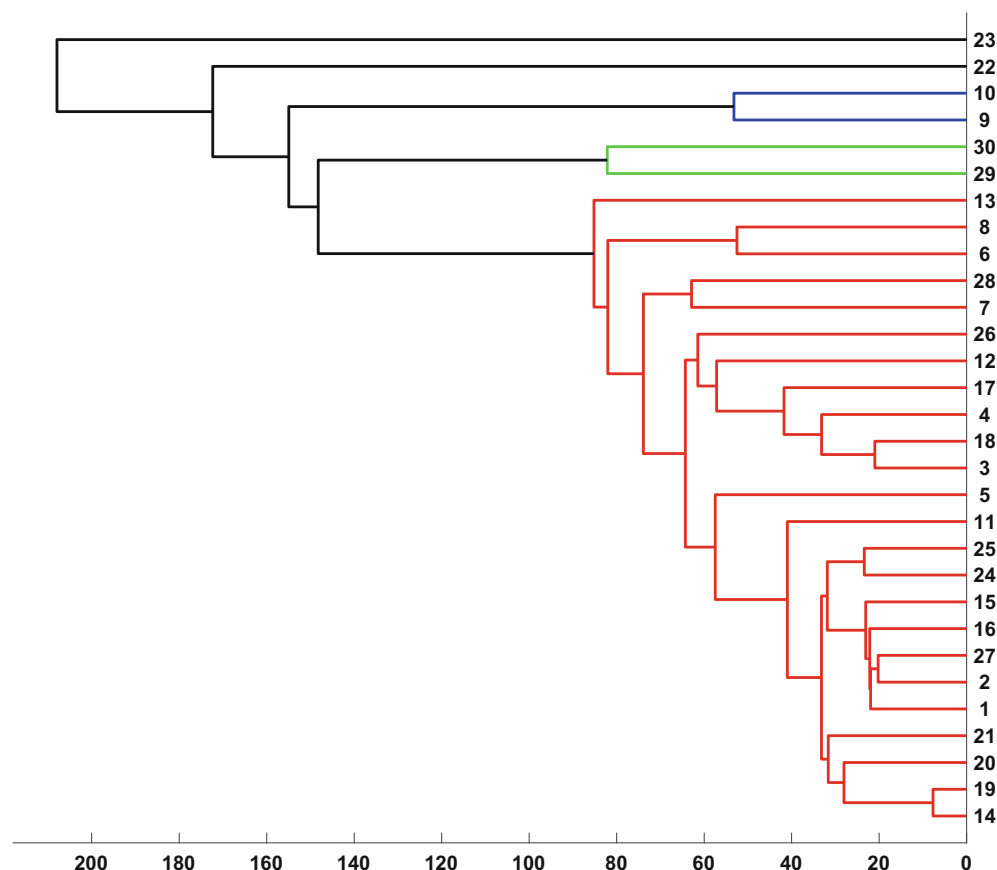
to associate the presence of the clusters with possible environmental drivers such as lithology, depth of the well, climate, giving interesting information about processes, and forcing variables.

In the context of geoscience, cluster analysis may be very useful to unravel the complexity of nature and its laws, opening paths for classification rules of unknown cases, thus representing the explorative basic tool of discriminant analysis.

When the aim of the investigation is the analysis of the relationships among variables, the Pearson correlation matrix $p \times p$ can represent the basis for cluster analysis. Results are reported in Fig. 4 by using the average linkage method. No appreciable differences were obtained changing the similarity measure (e.g., the sample Spearman rank correlation between variables) or the linking method. The cluster analysis enhances the relationships characterizing water chemistry revealing a strong link among Na^{2+} and HCO_3^- and a more isolated behavior of K^+ , perhaps related to its complex biogeochemical cycle and possible anthropical contribution due to the extensive use of fertilizers in agriculture. The identification of groups of variables helps in evaluating which type of water-rock interaction processes can be considered in the investigated sedimentary context; for example, the

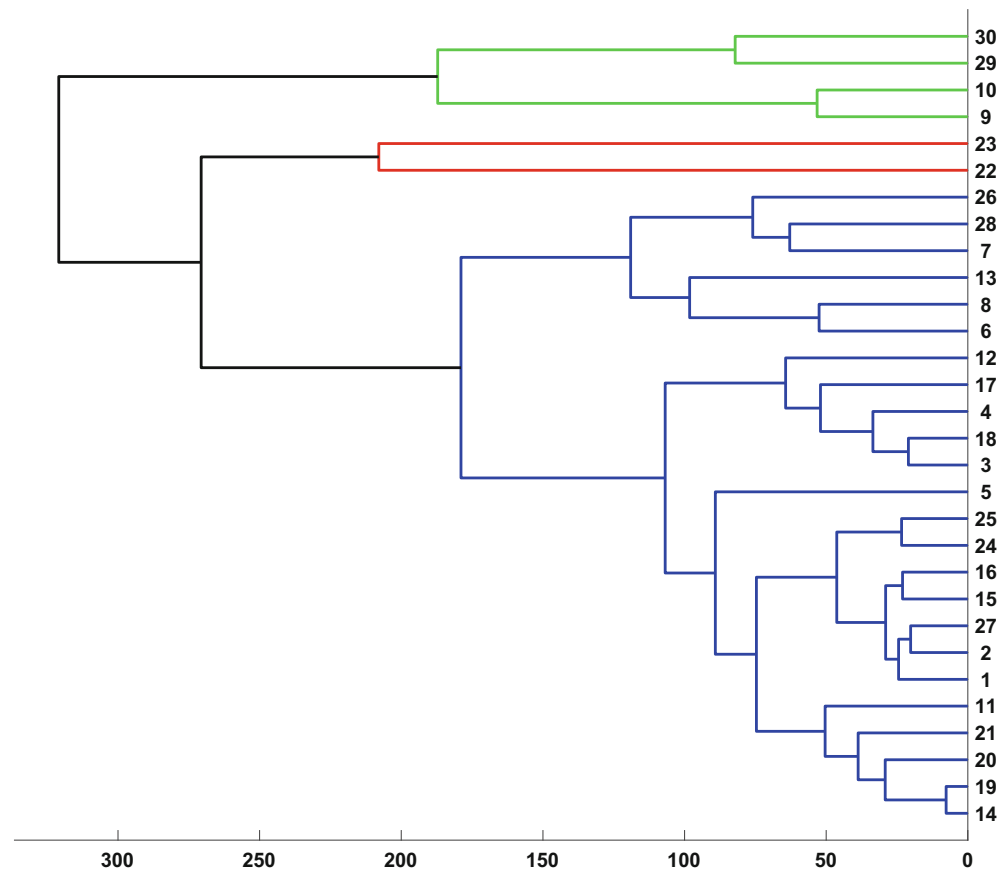
Cluster Analysis and Classification,

Fig. 2 Dendrogram for the chemical composition of water data sampled in a sedimentary aquifer (Minkowski distance and single linking method)



Cluster Analysis and Classification,

Fig. 3 Dendrogram for the chemical composition of water data sampled in a sedimentary aquifer (Minkowski distance and average linking method)



relationship among SO_4^{2-} and Ca^{2+} invites to check for the presence of sulfatic rocks while that between HCO_3^- and Na^+ to explore the role of cation exchange in the presence of clays.

In contrast to hierarchical clustering methods, partitioning methods require to a priori know the number of clusters (Kaufman and Rousseeuw 2005). A popular partitioning algorithm is the k-means algorithm that works minimizing the average of the squared distances between cases and their cluster centroids. Starting from k initial cluster centroids, the algorithm assigns the observations to their closest centroids using a given distance measure. Then, the cluster centroids are recomputed and iteratively the reallocation of the data points to the closest centroids performed. In this framework, Manhattan distance is used for k-medians when the centroids are defined by the medians of each cluster rather than their means. Several clustering methods have been proposed by Kaufman and Rousseeuw (2005) and are contained in several software packages (SPSS, Matlab, R) as the PAM (Partitioning Around Medoids) method that minimizes the average distances to the cluster medians, like the k-medians, and the CLARA (Clustering Large Applications) based on random sampling to reduce computing time and RAM storage problems.

Two different concepts of validity criteria need to be considered to evaluate the performance of a clustering procedure. External criteria compare the obtained partition with a partition that is known a priori and several indices can be used to do it (Haldiki et al. 2002). On the other hand, internal criteria are cluster validity measures that evaluate the clustering results of an algorithm by using only quantities and features inherent in the data set based on within and between cluster sums of squares, analogous to the multivariate analysis of variance (MANOVA), as the average silhouette width of Kaufman and Rousseeuw (2005).

Data Problems in the Context of Cluster Analysis

In multi-element analysis of geological materials, elements occur in very different concentrations so that, when methods exploring the variance-covariance structure are used, the variable with the greatest variance will have the greatest effect in the outcome. Consequently, variables expressed in different units should not be mixed or, as an alternative, the use of adequate transformation and standardization techniques (i.e., centering and scaling the data by using also a robust

Cluster Analysis and Classification, Table 1 Chemical composition in mg/L of the analyzed samples. The colored rows are related to cases that in cluster analysis have given isolated groups (22 and 23, 9, 10, 29 and 30, Figs. 1, 2, and 3)

Cases	Ca ²⁺	Mg ²⁺	Na ⁺	K ⁺	HCO ₃ ⁻	Cl ⁻	SO ₄ ²⁻
1	122	10.6	29.5	1.78	367	19.1	44.3
2	124	10.0	25.0	1.10	388	17.0	50.0
3	97.7	32.2	81.0	1.46	434	78.0	38.3
4	101	31.0	76.5	0.91	466	79.0	45.0
5	177	14.9	34.4	4.31	390	40.1	115
6	40.3	12.4	161	0.79	545	22.3	9.10
7	91.4	28.7	105	0.83	542	40.8	53.3
8	55.0	14.5	141	0.63	511	52.5	0.70
9	24.6	8.70	255	0.57	651	83.3	0.70
10	24.6	7.80	256	0.65	704	79.0	0.48
11	123	18.9	36.8	1.49	472	30.5	30.6
12	67.7	17.2	114	0.55	420	53.9	54.6
13	19.9	5.80	183	0.98	447	64.5	0.02
14	152	11.2	23.8	1.92	445	20.9	42.1
15	142	11.0	27.2	11.3	375	31.6	42.3
16	136	9.50	25.5	13.0	396	28.7	48.0
17	93.2	25.5	68.7	1.67	396	69.1	39.9
18	96.5	25.0	64.5	0.85	438	68.0	38.0
19	150	13.3	26.2	1.51	440	23.8	39.1
20	156	13.0	18.3	1.51	417	22.0	51.4
21	142	14.4	38.8	3.33	459	23.4	63.4
22	45.2	10.5	39.4	1.49	214	12.8	35.4
23	38.0	8.60	156	1.70	230	184	37.0
24	161	9.00	24.7	1.39	356	37.2	59.0
25	165	10.0	25.0	1.20	378	31.0	62.0
26	98.5	48.9	97.6	0.72	512	104	61.5
27	127	8.35	16.7	0.85	389	16.3	31.9
28	76.9	30.9	149	0.89	547	71.9	81.3
29	102	52.0	211	0.99	586	195	102
30	73.0	46.0	253	0.83	621	149	129

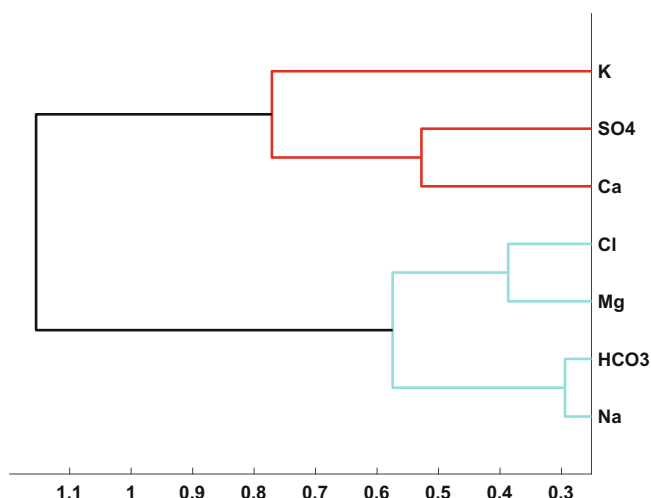
version of mean and standard deviation, such as median and MAD for skewed variables) is needed.

The presence of outliers can have severe effects on cluster analysis affecting the distance measures and distorting the underlying data structure. Thus, it is advisable to remove the outliers prior to the application of the clustering procedure or to adopt methods able to handle them (Kaufman and Rousseeuw 2005). From this point of view, the use of explorative univariate and bivariate tools (e.g., histograms, box plots, binary diagrams) is fundamental before proceeding with every multivariate methodology to probe the data structure.

The method of substitution of censored data can have a big effect on the results of the clustering procedure, particularly when a high percentage of cases are substituted with an identical value (e.g., the detection limit or its percentage). The elimination of variables characterized by a large amount

of data below the detection limit often represents the only solution. However, with the aim of not losing the information, these variables can be transformed into dichotomic variables (values below and above the detection limit transformed in 0 and 1, respectively) so that a contingency table can be realized cross-correlating the clusters obtained by the analysis without them and the variables with dichotomic behavior.

Compositional (closed) data can also be a problem in cluster analysis since they pertain to the simplex sample space and not the real one (Aitchison 1986), thus compromising the correct use of every similarity/distance measure. One of the transformations of the log-ratio family, in particular the log-centered or the isometric one, should be applied or the use of the Aitchison distance promoted (Templ et al. 2008).



Cluster Analysis and Classification, Fig. 4 Dendrogram for the chemical composition of water data sampled in a sedimentary aquifer. The linking method was given by the average linkage and the similarity measure by the Pearson correlation coefficient

Conclusions

Cluster analysis is widely used in geological sciences as an exploratory methodology to verify the presence of natural groups in data. From this point of view, cluster analysis is not a statistical inference technique, where parameters from a sample are assessed as possibly being representative of a population. Instead, it is an objective methodology for quantifying the structural characteristics of a set of observations. However, two critical issues must be considered, the representativeness of the sample and multicollinearity. In the first case, the researcher would be confident that the obtained sample is truly representative of the population and the presence of multimodality or outlier cases would be carefully monitored in single variables (univariate analysis) and in the multivariate distribution. In the second case, the attention must be posed to the extent to which a variable can be explained by the other variables in the analysis. In this context, the use of clustering techniques with compositional data is very critical due to the constrained interrelationships among all the variables.

As an alternative, model-based (Mclust) clustering can be used. It is not based on distances between the cases but on models describing the shape of possible clusters. The algorithm selects the cluster models (e.g., spherical or elliptical cluster shape) and determines the cluster memberships of all cases for solution over a range of different numbers of clusters. The estimation is done using the expectation maximization (EM) algorithm (Bouveyron et al. 2019).

Cross-References

- Compositional Data
- Correlation and Scaling
- Correlation Coefficient
- Dendrogram
- Discriminant Analysis
- Exploratory Data Analysis
- Fuzzy C-means Clustering

Bibliography

- Aitchison J (1986) The statistical analysis of compositional data. Chapman and Hall, London. 416 pp (reprinted in 2003 by Blackburn Press)
- Bouveyron C, Celeux G, Brendan Myrphy T, Raftery AE (2019) Model-based clustering and classification data science with applications in R. Cambridge Series in statistical and probabilistic mathematics. Cambridge University Press. 427 pp
- Czekanowski J (1909) Zur differential-diagnose der Neadertalgruppe. Korrespondenz-Blatt der Deutschen Gesellschaft für Anthropologie, Ethnologie, und Urgeschichte 40:44–47
- Dumitrescu D, Larrinerini B, Jain LC (2000) Fuzzy sets and their application to clustering and training. CRC Press, Boca Raton. 622 pp
- Haldiki M, Batistakis Y, Vazirgiannis M (2002) Cluster validity methods. SIGMOD Record 31:409–445
- Kaufman L, Rousseeuw PJ (2005) Finding groups in data. Wiley, New York. 368 pp
- Linnaeus C (1735) Systema Naturae, 1st edn. Theodorum Haak, Leiden, p 2
- Shirkhorshidi AS, Aghabozorgi A, Wah TY (2015) A comparison study on similarity and dissimilarity measures in clustering continuous data. PLoS One 10(12). <https://doi.org/10.1371/journal.pone.0144059>
- Sokal RR, Sneath PHA (1963) Principles of numerical taxonomy. W. H. Freeman, San Francisco, p 2
- Templ M, Filzmoser P, Reimann C (2008) Cluster analysis applied to regional geochemical data: problems and possibilities. Appl Geochem 23:2198–2213

Compositional Data

Vera Pawlowsky-Glahn¹ and Juan José Egozcue²

¹Department of Computer Science, Applied Mathematics and Statistics, U. de Girona, Girona, Spain

²Department of Civil and Environmental Engineering, U. Politècnica de Catalunya, Barcelona, Spain

Synonyms

Closed data; CoDa; Compositional data

Definitions

Definition 1 Compositional data (CoDa) are vectors with strictly positive components that carry relative information. Consequently, the quantities of interest are ratios between components.

CoDa represent parts of some whole, like, e.g., the geochemical composition of a rock sample, which states the proportion of each of the different elements present in the sample. The proportion can be expressed in percentages, as is usually done for major element oxides, or in ppm or ppb, as is ordinarily the case for trace elements. The change from percentages to ppm is attained by multiplying the observed values by a constant, and it is assumed that they still carry the same compositional information. The vectors are then proportional, as they are different representations of the same composition. They are equivalent compositions and constitute an equivalence class. A typical representation is as vectors of strictly positive components with constant sum, which is very convenient from the mathematical point of view. For a graphical representation, see Fig. 1. There are many examples in the geosciences, like the mineral composition of a rock, or the geochemical composition of a sample. The representation of such data in ternary diagrams has a long tradition in this field.

Definition 2 Composition. A composition is a vector of D components taking values in the positive orthant of real space, $\mathcal{R}_+^D$, that carry relative information. It is custom to call

them *parts*, a term suggested by Aitchison (1982) to avoid confusion with real variables. In order to represent a composition $\mathbf{x} = [x_1, x_2, \dots, x_D]$ in the simplex all parts are divided by their sum and multiplied by κ ,

$$\mathcal{C}\mathbf{x} = \left[\frac{\kappa x_1}{\sum_{i=1}^D x_i}, \frac{\kappa x_2}{\sum_{i=1}^D x_i}, \dots, \frac{\kappa x_D}{\sum_{i=1}^D x_i} \right], \quad \kappa > 0, \quad (1)$$

where the components (parts) of $\mathcal{C}\mathbf{x}$ add up to κ . $\mathcal{C}\mathbf{x}$ is called closure of $\mathbf{x}$ to the constant κ ; it is a composition equivalent to $\mathbf{x}$. Usually $\kappa = 1$ or $\kappa = 100$.

A random composition, $\mathbf{X} = [X_1, X_2, \dots, X_D]$, is a composition whose parts are random variables. In order to determine the sample space of the parts it is convenient to consider $\mathbf{Y} = \mathcal{C}\mathbf{X}$. The parts of $\mathbf{Y}$, Y_i , are positive and add to the closure constant $\kappa > 0$. Then, the values of Y_i are in the D -part simplex satisfying $\sum_{i=1}^D Y_i = \kappa$.

The notation in capitals, like $\mathbf{X}$, is used for random compositions and the lower case, as $\mathbf{x}$, for realizations of the random composition or for fixed values. The notation with capitals is also used for denoting generic parts; for instance, in a geochemical analysis, $X_1 \equiv \text{Ca}$, $X_2 \equiv \text{Na}$ refers to the parts Ca and Na, respectively, in a generic sense.

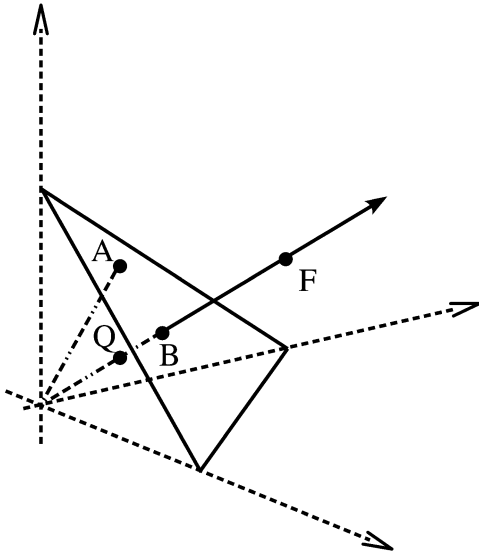
Although the D -part simplex is only a subset of $\mathcal{R}^D$ space, quite often, it is assumed that operations and metrics of $\mathcal{R}^D$, the ordinary Euclidean geometry, is valid for the analysis of random compositions in the simplex. This assumption underlays almost all statistical methods and leads to spurious results when applied to CoDa.

Definition 3 Subcomposition. A subcomposition of a D -part composition $\mathbf{X}_D$ is a subvector of S parts, $S < D$,

$$\mathbf{X}_S = [X_1, X_2, \dots, X_S],$$

which support is the positive orthant of real space, $\mathcal{R}_+^S$, carrying relative information.

It is assumed that the same rules stated above for compositions hold for subcompositions. Thus, they represent parts of a (partial) whole and proportional vectors represent the same subcomposition. They are equivalence classes and any suitable representant of the class can be chosen. It is common to use a closed representation, the closure constant being usually 1, 100 or 10^6 . One frequent example in the geosciences is the representation of a three-part subcomposition of a geochemical analysis in a ternary diagram.



Compositional Data, Fig. 1 Compositional equivalence classes in 3D. The ray from the origin through B is the class which contains the equivalent points Q and F

Introduction

Historically, CoDa have been defined as vectors of positive components constraint by a constant sum or closure constant, usually 1, 100 or 10^6 . Nowadays, understanding CoDa as equivalence classes represented by one element of that class has broadened the concept. Available methods, designed to be scale invariant, allow us to analyze the data in such a way that the closure constant, if any, is not important.

Problems with the analysis of CoDa were first detected by Karl Pearson (1897), who coined the term *spurious correlation*. Awareness of these problems was kept alive by many authors in the geosciences. It was John Aitchison (1982) who put forward the log-ratio approach, opening the way to deal consistently with CoDa. He set the basic framework that allowed later to define the algebraic-geometric structure of the sample space of CoDa. This was published simultaneously by Billheimer et al. (2001) and by Pawlowsky-Glahn and Egozcue (2001). It was the clue to understanding that there is no solution to these problems under the assumption that compositional data are vectors in real space that obey the usual Euclidean geometry.

It can be shown mathematically that for raw data represented in the simplex (Eq. 2), i.e., closed or normalized to a constant, the covariance matrix necessarily is singular, has some non-zero and some negative entries, and is subcompositionally incoherent. To find subcompositions for which the sign of one or more covariances changes, it suffices to take the closed representant of the subcomposition consisting of those parts which have positive covariances. This shows that covariances of compositions are subject to nonstochastic controls, and this fact leads to the conclusion that they are spurious, and thus nonsensical.

Problems with raw CoDa were described extensively by Felix Chayes (1971). They appear because they are assumed to be points in D -dimensional real space endowed with the usual Euclidean geometry. This assumption does not guarantee subcompositional coherence, and the following issues appear: (1) the mathematical necessity of at least one non-zero covariance; (2) the bias towards negative covariances; (3) the singularity of the covariance matrix; and (4) the non-stochastic control, and thus description and interpretation, of the dependence between the variables under study.

Principles

John Aitchison (1982) introduced three principles, here called *Aitchison principles*, which should govern any compositional analysis.

Aitchison Principle 1. Scale Invariance

Two proportional vectors give the same compositional information. They represent the same composition

$$\begin{aligned}\mathbf{X}_D &= [X_1, X_2, \dots, X_D] \equiv \kappa \cdot \mathbf{X}_D \\ &= [\kappa \cdot X_1, \kappa \cdot X_2, \dots, \kappa \cdot X_D].\end{aligned}$$

Aitchison Principle 2. Subcompositional Coherence

A result for $X_i, X_j \in \mathbf{X}_S$ will not lead to a contradiction with a result obtained for $X_i, X_j \in \mathbf{X}_D$.

Aitchison Principle 3. Permutation Invariance

Results should be invariant if the order of the parts in the composition is changed.

The principles above were stated by Aitchison (1982) before the vector space structure of the sample space of CoDa was recognized. They were led by the idea that methods for CoDa should be consistent whichever path is followed. Considering CoDa as equivalence classes corresponds to the Aitchison Principle 1, while the Aitchison geometry of the simplex makes the requirements of the Aitchison Principles 2 and 3 unnecessary, as long as one works on orthonormal (Cartesian) coordinates.

The Sample Space of Compositional Data

The set of compositions represented by vectors subject to a constant sum is the D -part simplex,

$$\mathcal{S}^D = \left\{ [x_1, x_2, \dots, x_D] \in \mathcal{R}_+^D \mid x_i > 0, i = 1, 2, \dots, D; \sum_{i=1}^D x_i = \kappa \right\}, \quad (2)$$

with κ an arbitrary constant, usually 1, 100 or 10^6 . The simplex is usually taken as a representative of the sample space of CoDa. The closure operation (1) allows one to select the representative of an equivalence class in $\mathcal{S}^D$. The principle of scale invariance, mentioned above, is reflected in the selection of the simplex (Eq. 2); it excludes zero or negative values. This leads to the zero problem in CoDa, which is extensively addressed in Palarea-Albaladejo and Martín-Fernández (2015).

An important concept in CoDa analysis is subcompositional coherence. As mentioned above, a subcomposition is a vector that includes only some parts of the original composition. A subcomposition is in general normalized to a constant sum, like when three parts of a larger composition are plotted in a ternary diagram. Subcompositional coherence is then the requirement that

statements about parts in the original composition hold when a subcomposition is considered.

The Aitchison Geometry for CoDa

For $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$, $\alpha \in \mathcal{R}$, the following operations can be defined:

- Perturbation: $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1 y_1, x_2 y_2, \dots, x_D y_D]$;
- Powering: $\alpha \odot \mathbf{x} = \mathcal{C}[x_1^\alpha, x_2^\alpha, \dots, x_D^\alpha]$;
- Inner product: $\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}$;

These operations generate a $(D - 1)$ -dimensional Euclidean vector space structure known as Aitchison geometry (Pawlowsky-Glahn and Egozcue 2001). The inner product has an associated norm

$$\|\mathbf{x}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D \left(\ln \frac{x_i}{x_j} \right)^2},$$

and an associated distance:

$$d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2}.$$

Perturbation, powering, and distance were already defined by Aitchison (1982). Perturbation is particularly interesting, as it can be interpreted in a very intuitive way as *linear filtering*. Also, any change in composition expressed as increasing or decreasing percentages of the parts is a perturbation as, for instance, a decrease of 3% implies multiplication by 0.97 of that part.

Linear Functionals on Compositions

In most CoDa analysis some real functions of compositions, called functionals, may have a special interest depending on the research questions. However, these functionals should be scale invariant to be consistent with the Aitchison principles. A functional ϕ is scale invariant if, for any composition $\mathbf{x}$ and any scalar $\mathbf{k}$, $\phi(k\mathbf{x}) = \phi(\mathbf{x})$. If the function is not scale invariant, equivalent compositions closed to a different constant κ give different results. For instance, the log ratios $\log(\text{Ca/Mg})$ and $\log(\text{Cl}/(\text{Ca} + \text{Mg}))$ in a water analysis do not depend on the units in which they are expressed, be it in mg/L or $\mu\text{g/L}$. In many cases, a linear behavior of functionals can be a convenient condition for functionals. This is $\phi(\alpha \odot (\mathbf{x} \oplus \mathbf{y})) = \alpha(\phi(\mathbf{x}) + \phi(\mathbf{y}))$. This linearity property assures that a change of units acts on the functional in the same way for mean values and that variability only changes with the square of the scaling factor α . Changing from mg/L to atoms/L in Ca and Mg consists of dividing both concentrations in

mg/L by the respective atomic weights which are, approximately, (40., 24.), respectively. This division of both elements and any other in the composition is a perturbation. If $\mathbf{x} = (\dots, \text{Ca}, \text{Mg}, \dots)$ and $\mathbf{y} = \mathcal{C}(\dots, 1/40., 1/24., \dots)$, then

$$\log \frac{\text{Ca mg/L}}{\text{Mg mg/L}} = \log \frac{\text{Ca atom/L}}{\text{Mg atom/L}} + \log \frac{y_{\text{Ca}}}{y_{\text{Mg}}},$$

$$y_{\text{Ca}} = \frac{1}{40.}, \quad y_{\text{Mg}} = \frac{1}{24.}.$$

Taking the values of $\log(\text{Ca/Mg})$ in mg/L across a sample, the average of the functional values changes summing the log ratio $\log(y_{\text{Ca}}/y_{\text{Mg}})$, and the variance remains unaltered. This recalls that a perturbation, in this case a change of units, is a shift in the simplex.

It is straightforward to check that the amalgamated log ratio $\log(\text{Cl}/(\text{Ca} + \text{Mg}))$ is nonlinear. Contrarily, consider the functional

$$\phi(\mathbf{x}) = \sum_{i=1}^D a_i \log x_i, \quad \sum_{i=1}^D a_i = 0,$$

called *log contrast*. All linear scale invariant log ratios have this form. Functionals like $\log(x_1/x_2)$, $\log(x_1/g_m(\mathbf{x}))$, and $\log(g_m(x_1, x_2)/g_m(x_3, \dots, x_D))$, where $g_m(\cdot)$ is the geometric mean of the arguments, are log contrasts.

Among log contrasts, the class of *balances* are of the form

$$B(\mathbf{x}_+/\mathbf{x}_-) = \sqrt{\frac{n_+ n_-}{n_+ + n_-}} \log \frac{g_m(\mathbf{x}_+)}{g_m(\mathbf{x}_-)}, \quad (3)$$

where $\mathbf{x}_+$ and $\mathbf{x}_-$ denote two groups of parts included in $\mathbf{x}$ and n_+ , n_- are the number of parts in the two groups, respectively. Sometimes, the square root in (3) is suppressed; the remaining log ratio of geometric means is a *non-normalized balance*.

The Principle of Working in Coordinates

It is well known that every $(D - 1)$ -dimensional Euclidean space is isometric to the $(D - 1)$ -dimensional real Euclidean vector space $\mathcal{R}^{D-1}$. This implies that *orthonormal coordinates* in $\mathcal{S}^D$ are elements of $\mathcal{R}^{D-1}$ that obey the usual Euclidean geometry; and, consequently, all available methods developed in $\mathcal{R}^{D-1}$ under the implicit or explicit assumption that the coordinates are orthonormal hold on orthonormal coordinates in $\mathcal{S}^D$. To apply this fact in practice, it suffices to realize that most statistical methods developed in $\mathcal{R}^{D-1}$ assume cartesian coordinates. However, some non-orthonormal coordinates have properties that made them historically relevant.

Non-Orthogonal Log-Ratio Representations

The following two non-orthogonal representations were introduced by Aitchison (1982):

Centered Log-Ratio (clr) Representation

For $\mathbf{x} \in \mathcal{S}^D$ and $g_m(\mathbf{x})$ the geometric mean of the parts in $\mathbf{x}$, it is defined as

$$\text{clr}(\mathbf{x}) = \left[\ln \frac{x_1}{g_m(\mathbf{x})}, \ln \frac{x_2}{g_m(\mathbf{x})}, \dots, \ln \frac{x_D}{g_m(\mathbf{x})} \right],$$

$$g_m(\mathbf{x}) = \left(\prod_{i=1}^D x_i \right)^{1/D}.$$

This representation is very useful for computational issues, as it has important properties:

- For $v_i = \ln \frac{x_i}{g(\mathbf{x})}$ it holds $\sum_{i=1}^D v_i = 0$.
- The clr is a bijection between $\mathcal{S}^D$ and the hyperplane $\mathcal{R}_0^D \subset \mathcal{R}^D$, that is, the plane in $\mathcal{R}^D$ such that for any $\mathbf{v} \in \mathcal{R}^D$, $\sum_{i=1}^D v_i = 0$.
- The clr inverse is

$$\mathbf{x} = \text{clr}^{-1}(\mathbf{v}) = \mathcal{C} \exp[v_1, v_2, \dots, v_D].$$

- The clr is an isomorphism, i.e., $\text{clr}(\mathbf{x} \oplus \mathbf{y}) = \text{clr}(\mathbf{x}) + \text{clr}(\mathbf{y})$ and $\text{clr}(\lambda \odot \mathbf{x}) = \lambda \cdot \text{clr}(\mathbf{x})$.
- The clr is isometric, i.e., $d_a(\mathbf{x}, \mathbf{y}) = d_e(\text{clr}(\mathbf{x}), \text{clr}(\mathbf{y}))$, where d_a stands for the Aitchison distance defined above, and d_e for the usual Euclidean distance.

Additive Log-Ratio (alr) Representation

For $\mathbf{x} \in \mathcal{S}^D$, it is defined as

$$\text{alr}(\mathbf{x}) = \left[\ln \frac{x_1}{x_D}, \ln \frac{x_2}{x_D}, \dots, \ln \frac{x_{D-1}}{x_D} \right],$$

where the part in the denominator can be any part in the composition. This representation has mainly historical interest, as it corresponds to a representation in oblique coordinates, which are not always easy to handle in computing Aitchison distances and orthogonal projections.

Both representations, clr and alr, are too easily associated with the compositional parts appearing in the numerators of the log ratio, thus suggesting a kind of identification. This association is a source of confusion and misinterpretation. In fact, the log ratios in both alr and clr involve more than a single part, thus favoring the false impression that the log-

ratio CoDa approach can deal with single parts when the relevant quantities are always scale-invariant ratios.

Orthogonal Log-Ratio Representations

They are based on the Euclidean vector space theory. The orthogonal projection of a vector on an orthonormal basis of the space generates orthonormal coordinates which fully represent the vector. The same ideas apply to the simplex where the vectors are now compositions.

Definition 4 Orthonormal basis. Compositions $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}$ in $\mathcal{S}^D$ constitute an orthonormal basis if

$$\langle \mathbf{e}_i, \mathbf{e}_j \rangle_a = \langle \text{clr}(\mathbf{e}_i), \text{clr}(\mathbf{e}_j) \rangle = \delta_{ij},$$

where the Kronecker delta δ_{ij} is equal to 0 when $i \neq j$ and equal to 1 for $i = j$.

Definition 5 Basis contrast matrix. It is the $\text{clr}(D-1, D)$ matrix of the basis:

$$\Psi = \begin{pmatrix} \text{clr}(\mathbf{e}_1) \\ \text{clr}(\mathbf{e}_2) \\ \vdots \\ \text{clr}(\mathbf{e}_{D-1}) \end{pmatrix},$$

$$\Psi\Psi' = I_{D-1}, \quad \Psi'\Psi = I_D - (1/D)\mathbf{1}_D'\mathbf{1}_D$$

Coordinates of a composition $\mathbf{x}$ are obtained computing the Aitchison inner product of $\mathbf{x}$ with each of the elements of an orthonormal basis, $x_i^* = \langle \mathbf{x}, \mathbf{e}_i \rangle_a$, and the expression of the original composition in terms of orthonormal coordinates leads to $\mathbf{x} = \oplus_{i=1}^{D-1} x_i^* \odot \mathbf{e}_i$. The function (olr) that assigns orthonormal coordinates to a composition is one to one.

Originally, olr coordinates were known as *isometric log-ratio coordinates* (ilr) (Egozcue et al. 2003), but this concept was confusing, as, e.g., the clr coordinates are isometric but not orthonormal, while any orthonormal system of coordinates in $\mathcal{S}^D$ endowed with the Aitchison geometry is isometric (see discussion Egozcue and Pawłowsky-Glahn 2019a, b).

Properties of an Orthonormal Representation

The olr: $\mathcal{S}^D \rightarrow \mathbb{R}^{D-1}$ defined through an orthonormal basis $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}$ in $\mathcal{S}^D$ as $\text{olr}(\mathbf{x}) = \mathbf{x}^* = \text{clr}(\mathbf{x}) \cdot \Psi'$, with inverse $\text{olr}^{-1}(\mathbf{x}^*) = \mathbf{x} = \mathcal{C}(\exp(\mathbf{x}^*\Psi))$, is an **isometry**, that is,

- $\text{olr}(\alpha \odot \mathbf{x}_1 \oplus \beta \odot \mathbf{x}_2) = \alpha \cdot \text{olr}(\mathbf{x}_1) + \beta \cdot \text{olr}(\mathbf{x}_2) = \alpha \cdot \mathbf{x}_1^* + \beta \cdot \mathbf{x}_2^*$
- $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_a = \langle \text{olr}(\mathbf{x}_1), \text{olr}(\mathbf{x}_2) \rangle_e = \langle \mathbf{x}_1^*, \mathbf{x}_2^* \rangle_e$
- $\|\mathbf{x}\|_a = \|\text{olr}(\mathbf{x})\|, d_a(\mathbf{x}_1, \mathbf{x}_2) = d_e(\text{olr}(\mathbf{x}_1), \text{olr}(\mathbf{x}_2))$

where the subscripts e indicate that the operator is the ordinary Euclidean one.

There are different ways to build orthonormal coordinates within the Aitchison geometry for CoDa. One of them is based on principal component analysis of a dataset in its clr representation. The nonstandardized scores are olr coordinates and the loadings are a contrast matrix (Pawlowsky-Glahn et al. 2015). Consequently, the coordinates obtained in this way are data driven. Some further details are discussed in section “Exploratory CoDa Tools: Variation Matrix, Biplot, and CoDa-Dendrogram.”

Sequential Binary Partitions

A straightforward and intuitive way to define an orthonormal basis is through a sequential binary partition (SBP) of a generic composition. The construction is conveniently based on expert knowledge, but it can be done automatically or using specific algorithms, like the principal balances explained below. The procedure to define SBP (Pawlowsky-Glahn et al. 2015) is best explained with an example. Consider a composition $\mathbf{x} \in \mathcal{S}^5$. To define an SBP and to obtain the coordinates in the corresponding orthonormal basis proceed as illustrated in Table 1: The olr-coordinates assigned to each step of the partition are balances as defined in Eq. (3). For instance, the coordinate in step 3 in Table 1 would be $B(x_1/x_3, x_4)$ following the notation in that Equation.

Zeros

Most compositional datasets report zeros in some parts and cases. These zero values are incompatible with their transformation into log ratios, since infinite values would appear in the transformed dataset. Therefore, a preprocessing of the raw data is frequently required. However, the treatment of zeros needs a careful examination of their nature, and possibly some additional assumptions and decisions which are not based on the data themselves. This is a short description of the procedures to treat zeros. The interested reader is redirected to Palarea-Albaladejo and Martín-Fernández (2015) and references therein.

A first step is to decide the origin and nature of the zeros in the dataset.

At least, three classes can be distinguished: essential zeros, rounded zeros, and zeros from counts. *Essential zeros* are those which are due to the definition of the parts implying that, in certain circumstances, one or more parts are completely absent. An example is the use of land including for instance classes of forest and some types of agriculture. If a parcel is completely covered by forest, then there is no agriculture and each type of agriculture would be reported as 0%. These zeros are essential zeros. The treatment of essential zeros mainly consists in translating the presence-absence of zeros into a factor defining a subpopulation.

Rounded zeros and zero count can be processed in, at least, two ways: (A) the substitution of reported zeros by an imputed value, which is then processed as a regular compositional datum or, alternatively, (B) the zeros are thought as observations in a model in which these zero values are acceptable. Then, the analysis of such a model, consisting in the estimation of parameters, requires to know the likelihood of the parameters given the data including the zeros. The B methods avoid the explicit imputation of the zero values by estimating the parameters of the model.

Commonly, in A cases, a detection threshold is available or can be inferred to help the imputation. There are two classes of imputation, those where information only of the single part, which contains the zero, and of the corresponding detection limit is used; and those taking into account the relationship with other parts of the composition. If these imputation techniques are applied to counts, the imputation consists of a pseudo-count, frequently a value between 0 and 1.

A simple case is that of data in percentages, where also the detection limit is given in percentages. When a zero is present the multiplicative strategy of replacement consists of assigning a fraction of the detection limit to the reported zero, and then replacing all other observed data so that the resulting composition adds to 100.

A typical example of including the zeros in the estimation of parameters is the zero counts in a multinomial sampling. The goal is frequently to estimate the multinomial probabilities which are assumed positive. The multinomial likelihood

Compositional Data, Table 1 Example of SBP for a five-part composition. Four partition steps are necessary to complete the SBP; each one divides the active subcomposition into two groups of parts, marked

Step	x_1	x_2	x_3	x_4	x_5	Coordinate
1	+ 1	− 1	+ 1	+ 1	− 1	$y_1 = \sqrt{\frac{3 \cdot 2}{3+2}} \ln \frac{(x_1 \cdot x_3 \cdot x_4)^{1/3}}{(x_2 \cdot x_5)^{1/2}}$
2	0	+ 1	0	0	− 1	$y_2 = \sqrt{\frac{1 \cdot 1}{1+1}} \ln \frac{x_2}{x_5}$
3	+ 1	0	− 1	− 1	0	$y_3 = \sqrt{\frac{1 \cdot 2}{1+2}} \ln \frac{x_1}{(x_3 \cdot x_4)^{1/2}}$
4	0	0	+ 1	− 1	0	$y_4 = \sqrt{\frac{1 \cdot 1}{1+1}} \ln \frac{x_3}{x_4}$

with +1 or −1, respectively; nonactive parts are labeled with 0. The expression of balance coordinates corresponding to each step are in the right column

can include in a natural way zeros in the observation. A Bayesian estimation of multinomial probabilities provides an elegant solution of the problem. This class of approach typically requires the assumption of an underlying model for the observations.

The R-package *zCompositions* implements most of these imputation techniques (Palarea-Albaladejo and Martín-Fernández 2015, and references therein). However, the treatment of zeros in CoDa is not only a computational problem, but it requires a detailed examination of the character of the zeros and even a complete model for the whole dataset.

Exploratory CoDa Tools: Variation Matrix, Biplot, and CoDa-Dendrogram

Following the principle of working in coordinates (section “The Principle of Working in Coordinates”), compositional datasets can be represented in coordinates, preferably orthogonal, which are real variables. As a consequence, all exploratory tools devised for real variables can be applied to compositional log-ratio coordinates. However, there are at least three exploratory techniques that, although inspired in exploration of real variables, have characteristics that deserve particular attention. They are the variation matrix, the CoDa-biplot, and the CoDa-dendrogram. In order to illustrate the use of these tools, the Varanasi dataset described below is used.

Varanasi Dataset

The Varanasi dataset (VDS) (Olea et al. 2017) consists of concentration of ions dissolved in ground water sampled at 95 wells in the surroundings of Varanasi (India) on the two margins of the Ganga river. The following concentrations are reported: Fe($\mu\text{g/L}$), As($\mu\text{g/L}$), Ca(mg/L), Mg(mg/L), Na (mg/L), K(mg/L), HCO_3 (mg/L), SO_4 (mg/L), Cl(mg/L), NO_3 (mg/L), and F(mg/L). The data also reports the location

of the sample wells. Wells numbered 1 to 44 are placed on the left margin, and the remaining wells, from 45 to 95, on the right margin of the Ganga river.

The data matrix is denoted by $\mathbf{X}$; its columns are denoted by X_j , $j = 1, 2, \dots, D$ ($D = 11$ in VDS) and the rows containing a composition are denoted $\mathbf{x}_i$, where $i = 1, 2, \dots, n$ ($n = 95$ in VDS).

A first step in an exploration of a compositional dataset is often the computation of quantiles of its columns. This step is not important in a compositional data study except for detecting errors and/or the presence of zeros, and to have a first impression of the scale and units of the parts. In VDS, there are no zeros, but the initial values of Fe and As were given in $\mu\text{g/L}$, while the remaining elements are in mg/L . Although the compositional analysis can be performed on these nonhomogeneous units, Fe and As were divided by 1000 for translating them into mg/L . Note that such transformation is a perturbation.

Center and Variability

The *center* of a composition plays the role of the mean for real data. It is estimated as a perturbation-average along the data matrix

$$\text{cen}[\mathbf{X}] = \frac{1}{n} \odot \bigoplus_{i=1}^n \mathbf{x}_i,$$

where the big $\oplus$ is a repeated perturbation. The symbol *cen* denotes *sample center* and it is an estimator of the theoretical center. Table 2 shows the sample center of the VDS in mg/L . Those values are the geometric means of the columns of VDS provided that all compositional data (rows of $\mathbf{X}$) are given in homogeneous units. When all the data in the sample are closed to a constant κ , the resulting geometric means by columns should be closed to the same κ .

In statistical analysis of real data, variability of the sample is very often described by the sample covariance matrix or its

Compositional Data, Table 2 Sample center (first row in mg/L) and normalized variation matrix $[\tau_{ij}]$ of Varanasi data corresponding to the samples on the left margin of Ganga river. Normalized variations less than 0.2 are in boldface and marked with an asterisk

	Ca	Mg	Na	K	HCO3	SO4	Cl	NO3	F	Fe	As
cenXL	43.62	43.99	57.94	4.89	328.15	36.68	71.90	9.96	0.64	0.45	2.30e-3
Ca	0.00	0.90	0.72	1.18	0.42	1.33	0.70	1.88	0.55	0.44	0.53
Mg	0.90	0.00	0.54	1.06	0.30	0.67	0.29	3.05	0.30	0.78	0.43
Na	0.72	0.54	0.00	0.92	0.36	0.58	0.35	2.41	0.49	0.58	0.60
K	1.18	1.06	0.92	0.00	0.87	1.11	0.97	2.97	0.93	1.03	0.87
HCO3	0.42	0.30	0.36	0.87	0.00	0.91	0.44	2.54	*0.11	0.47	*0.16
SO4	1.33	0.67	0.58	1.11	0.91	0.00	0.60	3.43	0.91	1.11	1.11
Cl	0.70	0.29	0.35	0.97	0.44	0.60	0.00	2.73	0.49	0.76	0.58
NO3	1.88	3.05	2.41	2.97	2.54	3.43	2.73	0.00	2.54	1.31	2.44
F	0.55	0.30	0.49	0.93	*0.11	0.91	0.49	2.54	0.00	0.47	0.29
Fe	0.44	0.78	0.58	1.03	0.47	1.11	0.76	1.31	0.47	0.00	0.48
As	0.53	0.43	0.60	0.87	*0.16	1.11	0.58	2.44	0.29	0.48	0.00

standardization, the (Pearson) correlation matrix. When dealing with compositional data these matrices are spurious (Aitchison 1986). They are useless and confusing for a further analysis. Alternatively, variability of a compositional dataset can be described by means of the variation matrix (Aitchison 1986). The *variation matrix* is a (D, D) matrix T whose entries are

$$t_{ij} = \text{var} \left[\log \frac{X_i}{X_j} \right], \quad i, j = 1, 2, \dots, D.$$

The entries $t_{ii} = 0$, as there is no variability in $\log(X_i/X_i)$. When t_{ij} is small, near to zero, it suggests linear association between the parts X_i and X_j (Egozcue et al. 2018). In fact, if $X_i \simeq aX_j$, for some $a \in \mathcal{R}$, then $t_{ij} \simeq 0$, that is, they are proportional or nearly so.

Large values of t_{ij} point out the main sources of variability in the sample. In Aitchison (1986), the sample *total variance* is defined as the average entry of T , that is,

$$\text{totvar}[\mathbf{X}] = \frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D t_{ij}.$$

The total variance can be also expressed using the clr and olr representations of the sample. It can be shown that

$$\text{totvar}[\mathbf{X}] = \sum_{i=1}^D \text{var}[\text{clr}_i(\mathbf{X})] = \sum_{i=1}^{D-1} \text{var}[\text{olr}_i(\mathbf{X})], \quad (4)$$

where the last term points out that the sample total variance is the trace of the sample covariance matrix of the olr coordinates, thus matching the terminology in statistical analysis of real data.

The values in T strongly depend on the number of parts, D , and the total variance. In order to interpret and compare the values in T , some normalizations have been proposed. For instance, if the total variance were spread over all non-null entries of T , the values would be $2D\text{totvar}[\mathbf{X}]/(D(D-1))$. Comparing this value with t_{ij} , a normalized variation is

$$\tau_{ij} = \frac{D-1}{2\text{totvar}[\mathbf{X}]} \cdot t_{ij}, \quad \sum_{i=1}^D \sum_{j=1}^D \tau_{ij} = (D-1)D$$

where values of $\tau_{ij} \geq 1$ suggest no linear association and $\tau_{ij} < 1$ indicates possible association (Pawlowsky-Glahn et al. 2015). In practice, $\tau_{ij} > 0.2$ corresponds to very poor linear associations. It is worth noting that the approximate linear association between two parts depends on the subcomposition considered (Egozcue et al. 2018). Table 2 shows the sample center and the normalized variation matrix for the data from the left margin of the Ganga river. The ion bicarbonate is suggested to be linearly associated with F and As. These evidences of association are not apparent in the right margin.

Compositional Biplot

The singular value decomposition (SVD) is an algebraic technique for decomposition of data matrices. It is routinely used in principal component analysis. When dealing with compositional data the SVD is applied to the centered $\text{clr}(\mathbf{X})$, that is to $\text{clr}(\mathbf{X} \ominus \text{cen}[\mathbf{X}])$, where $\text{cen}[\mathbf{X}]$ is perturbation-subtracted from each row of $\mathbf{X}$. The SVD results in

$$\text{clr}(\mathbf{X} \ominus \text{cen}[\mathbf{X}]) = U \Lambda V^T,$$

where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_{D-1}, 0)$ contains the singular values ordered from larger to smaller; U is an (n, D) matrix and V is a (D, D) matrix. The last singular value is 0 due to the fact that the rows of the decomposed matrix add up to 0, as pointed out above, after the definition of the clr. This means that the last 0 in Λ , and the last columns of U and V , can be suppressed. After this suppression and maintaining the same notation, Λ , U , and V are $(D-1, D-1)$, $(n, D-1)$, $(D, D-1)$ matrices, respectively (Pawlowsky-Glahn et al. 2015). The main property of V is that it is orthogonal, i.e., $V^T V = I_{D-1}$. Therefore, $\Psi = V^T$ is a contrast matrix and $U\Lambda = \text{olr}(\mathbf{X} \ominus \text{cen}[\mathbf{X}])$. The columns of V (rows of Ψ) contain the clr's of the elements of an orthonormal basis of the simplex, and the rows of $U\Lambda$ are the (centered) olr coordinates of the compositional dataset. Also $U^T U = I_{D-1}$ and, consequently, the singular values are proportional to the standard deviations of the olr coordinates. Furthermore, $\sum_{i=1}^{D-1} \lambda_i^2 = n \text{totvar}[\mathbf{X}]$, where the centering of $\mathbf{X}$ can be suppressed, since centering does not affect the total variance. These properties allow the simultaneous plotting of the centered clr variables and of the compositions in the dataset, represented by their olr coordinates. Commonly, two olr coordinates are chosen for the plot, say the first and the second coordinates. For such a projection the proportion of total variance shown in the plot is $(\lambda_1^2 + \lambda_2^2)/n \text{totvar}[\mathbf{X}]$. Superimposed to the olr coordinates, the clr of the vectors of the basis (columns of V) can be plotted as arrows. As these vectors are unitary, the length in the plot indicates the orthogonal projection on the coordinates chosen for the biplot. This kind of plot is called *form compositional biplot*. It is appropriate to look for features of the data and to their (Aitchison) interdistances projected onto a two-dimensional plot.

However, the most frequent biplot is the *covariance compositional biplot*. The compositions in the dataset are represented by the standardized coordinates in U (called scores in a standard principal component analysis); simultaneously, the arrows to the columns of ΛV are plotted, so that the length of the arrows are proportional to the standard deviation of the clr variables, provided the projection represents a good deal of the total variance. The main interpretative tools are the lengths and directions of the links between the heads of two arrows. These lengths, up to a good projection,

are proportional to the standard deviations of the log ratios between the parts labeling the corresponding arrows. That is, they are approximately proportional to the square root of the elements in the variation matrix.

Figure 2 shows covariance biplots for VDS. The corresponding form biplots are quite similar and are not shown. The left panel corresponds to the projection on the first and second principal coordinates, but this projection only represents 50% of the total variance. The right panel shows the projection on the first and third principal coordinates, this latter contributing an additional 15% of the total variance. The sampling points on the left margin of the river are plotted in red and those on the right margin in green. In the two projections the left margin points appear less disperse than those of the right margin. However, the left margin sample appears mixed within the right margin in both projections. In fact, the total variance of the joint sample is 7.61 while the total variance for the left and right margin samples are, respectively, 5.14 and 8.28.

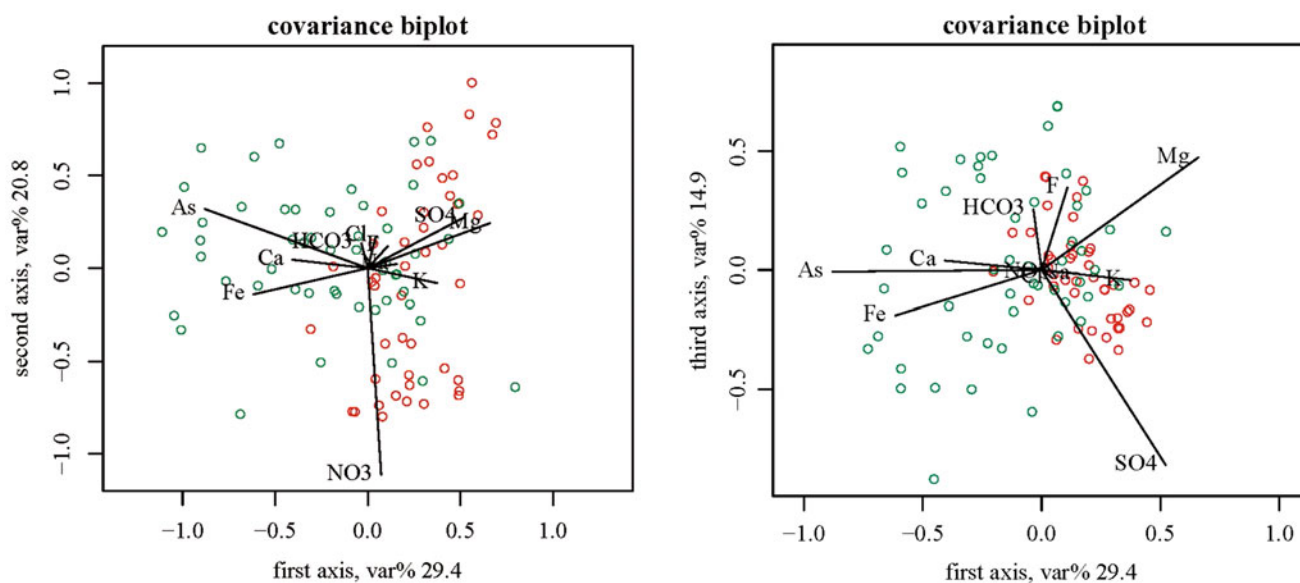
In both projections shown in Fig. 2, the links between the arrows from the apexes of As, Fe, and Ca to those of Mg, K, and SO_4 are approximately parallel to the first principal axis. This means that the fraction of total variance of the first principal coordinate comes from log-ratio variances between elements from one from each group. Also, this suggests that the variance of the balance $B(\text{Mg}, \text{K}, \text{SO}_4/\text{As}, \text{Fe}, \text{Ca})$ can be similar (but less than) the variance of the first principal coordinate. Additionally both biplots suggest that the two margins of the river are distinguishable in mean by this first principal coordinate or by its proxy, the mentioned balance. Many observable features in the biplots are also identifiable in the variation matrices in a more quantitative way. For this reason

the CoDa-biplot can be considered one of the most powerful tools in the exploration of CoDa. More detailed explanations can be found in Pawlowsky-Glahn et al. (2015) and references therein.

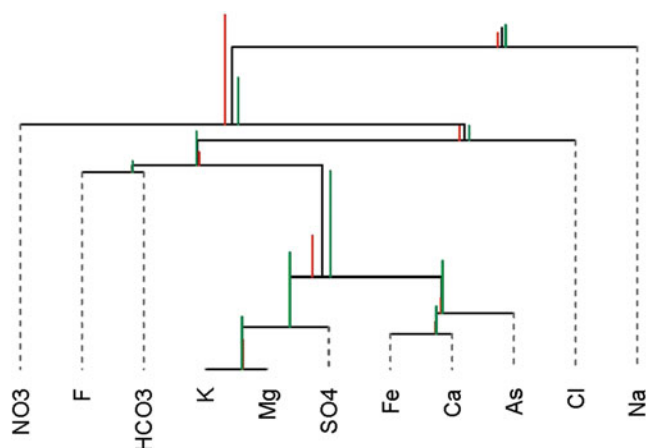
Compositional Dendrogram

The principle of working on coordinates proposes the statistical analysis of olr coordinates as real coordinates. This is valid for exploratory purposes as well. Consequently, the standard description of sample mean, quantiles, and variances of coordinates should be accompanied of a definition of the basis that define the coordinates. In the case in which the basis is defined by an SBP and the coordinates are balances, the basis can be represented by a tree structure or dendrogram. An exploration based on mean and variances of the balance coordinates can be shown on the dendrogram structure, thus providing a powerful interpretation of the balance coordinates. Figure 3 is a CoDa-dendrogram for the VDS. A partition of the composition has been chosen using the principal balances technique (Martín-Fernández et al. 2018) which tries to mimic the principal component analysis using balances. The elements in a CoDa-dendrogram are:

- (1) The tree which reproduces the SBP selected.
- (2) The horizontal bars, which are equally scaled in the interval $(-7, 7)$ in Fig. 3, representing an interval for the corresponding balance in the tree. Their length in the plot is meaningless and only corresponds to the spacing of the labels. The anchoring of vertical bars on the horizontal ones points out the mean of the balance in the mentioned scale.



Compositional Data, Fig. 2 Covariance biplots for VDS. Left margin, red points; right margin, green points. Projection on first-second (left panel) and first-third (right panel) principal coordinates



Compositional Data, Fig. 3 CoDa dendrogram for VDS. The SBP corresponds to principal balances. Black tree: joint sample; red tree: left margin sample; green tree: right margin sample

- (3) The length of vertical full lines is proportional to the variance associated with the balance, thus reproducing the decomposition of the total variance by olr coordinates in Eq. (4).

The CoDa-dendrogram can be completed with some boxplots following the scale of the horizontal bars (not shown in Fig. 3). When there are two or more subpopulations, like in the present case the left and right margins of the Ganga river, the trees corresponding to each subpopulation can be superimposed. In Fig. 3, the black tree corresponds to the joint sample, whereas the red and the green trees represent the data in the left and right margins, respectively. This allows the comparison of mean values and variances of the subpopulations, a kind of visual ANOVA for each coordinate.

If the goal is to identify features that distinguish the left and right margins of the river, Fig. 3 reveals that the sample mean of balance $B(\text{Mg}, \text{K}, \text{SO}_4/\text{As}, \text{Fe}, \text{Ca})$ shows the largest difference for the two margins. Simultaneously, the CoDa-dendrogram indicates that the variance of this balance in the right margin is larger than in the left margin, a fact also shown in Fig. 2. Also, the variance of balance $B(\text{F}/\text{HCO}_3)$ is similar in both margins, and the variance is smaller than that of other balances, thus suggesting a linear association between F and HCO_3 . It should be remarked that the CoDa-dendrogram is strongly dependent on the SBP selected, but is extremely useful when the goal of the exploration is clearly stated and expert knowledge about the problem is available.

Software for CoDa

Several software packages are available for the analysis of compositional data following a log-ratio approach, among others the following:

1. **CoDaPack**: This is a user-friendly, stand-alone, freeware, developed by the Research Group on Compositional Data Analysis from the Computer Science, Applied Mathematics, and Statistics Department of the University of Girona (UdG). It can be downloaded from the web. This software was developed as support for short courses and is not a complete statistical package. Nevertheless, the most recent version allows a direct link to R, opening a wide range of opportunities (Comas-Cufí and Thió-Henestrosa 2011).
2. **R-package “compositions”**: This is a very complete package including most procedures available for CoDa. It is conceived for scientists familiar with the mathematical foundations of CoDa and R-programming. An introduction to this package and CoDa procedures is van den Boogaart and Tolosana-Delgado (2013).
3. **R-package “robCompositions”**: Originally implementing robust statistical procedures in CoDa, it grew to incorporate other CoDa procedures. Many of its features are described in Filzmoser et al. (2018).
4. **R-package “zCompositions”**: Directed to the analysis of different kinds of zeros and missing data appearing in CoDa datasets. Useful in preprocessing of most CoDa problems (Palarea-Albaladejo and Martín-Fernández 2015).

Conclusions

CoDa represents the paradigm of what can go wrong if a methodology is applied to data that do not comply with the assumptions on which the methodology is based. A consistent and coherent solution for CoDa came from recognizing the Euclidean-type structure of the sample space, nowadays known as Aitchison geometry. In the framework of this geometry coordinates can be defined that allow for the application of standard methods. It is still a very active and open field of research, as many research questions call for the development of appropriate models which comply with the log-ratio approach.

Books on Compositional Data Analysis

The Aitchison (1986) book is the first book on compositional data analysis. It is seminal for most of procedures in CoDa within the log-ratio approach. After the recognition of the Aitchison geometry of the simplex and of the conception of compositions as equivalence classes, a number of books covering all theory, applications, and software have been published:

- Buccianti et al. (2006), made of contributed chapters. Most chapters are applications of compositional data to geological cases, but it also contains theoretical elements.

- Pawlowsky-Glahn and Buccianti (2011), made of contributed chapters. It includes chapters collecting advances in theoretical issues related to CoDa analysis and related methodologies. Also a variety of applications are collected. It represents the state of the art in 2011.
- van den Boogaart and Tolosana-Delgado (2013), mainly oriented to introducing the R-package *compositions* and its use, also contains valuable descriptions of the CoDa theory.
- Pawlowsky-Glahn et al. (2015), initially conceived as lecture notes, is a textbook introducing most methods dealing with CoDa.
- Filzmoser et al. (2018) is a guide through CoDa statistical applications with theoretical hints and R-programming linked to the R-package *rob-Compositions*.
- Greenacre (2018) presents some CoDa methodologies illustrated with examples developed with software in R.

Cross-References

- Aitchison, John
- Chayes, Felix
- Principal Component Analysis

Bibliography

- Aitchison J (1982) The statistical analysis of compositional data (with discussion). *J R Stat Soc Ser B (Stat Methodol)* 44(2):139–177
- Aitchison J (1986) The statistical analysis of compositional data. Monographs on statistics and applied probability. Chapman & Hall, London. (Reprinted in 2003 with additional material by The Blackburn Press). 416 p
- Billheimer D, Guttorp P, Fagan W (2001) Statistical interpretation of species composition. *J Am Stat Assoc* 96(456):1205–1214
- Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V (2006) Compositional data analysis in the geosciences: from theory to practice, volume 264 of special publications. Geological Society, London
- Chayes F (1971) Ratio correlation. University of Chicago Press, Chicago, 99 p
- Comas-Cufí M, Thió-Henestrosa S (2011) Codapack 2.0: a stand-alone, multi-platform compositional software. In: Egozcue JJ, Tolosana-Delgado R, Ortego MI (eds) Proceedings of the 4th international workshop on compositional data analysis (2011). CIMNE, Barcelona. ISBN 978-84-87867-76-7
- Egozcue JJ, Pawlowsky-Glahn V (2019a) Compositional data: the sample space and its structure (with discussion). *Test* 28(3):599–638. <https://doi.org/10.1007/s11749-019-00670-6>
- Egozcue JJ, Pawlowsky-Glahn V (2019b). Rejoinder on: “compositional data: the sample space and its structure”. *Test* 28(3):658–663. <https://doi.org/10.1007/s11749-019-00674-2>
- Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. *Math Geol* 35(3):279–300
- Egozcue JJ, Pawlowsky-Glahn V, Gloor GB (2018) Linear association in compositional data analysis. *Aust J Stat* 47(1):3–31
- Filzmoser P, Hron K, Templ M (2018) Applied compositional analysis. With worked examples in R. Springer Nature, Cham, 280pp
- Greenacre M (2018) Compositional data analysis in practice. Chapman and Hall/CRC, Cham, 122 pp
- Martín-Fernández JA, Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2018) Advances in principal balances for compositional data. *Math Geosci* 50(3):273–298
- Olea RA, Raju NJ, Egozcue JJ, Pawlowsky-Glahn V, Singh S (2017) Advancements in hydrochemistry mapping: methods and application to groundwater arsenic and iron concentrations in Varanasi, Uttar Pradesh, India. *Stochastic Environ Res Risk Assess (SERRA)* 32(1): 241–259
- Palarea-Albaladejo J, Martín-Fernández JA (2015) zCompositions – R package for multivariate imputation of left-censored data under a compositional approach. *Chemom Intell Lab Syst* 143:85–96
- Pawlowsky-Glahn V, Buccianti A (eds) (2011) Compositional data analysis: theory and applications. Wiley, Hoboken, 378 p
- Pawlowsky-Glahn V, Egozcue JJ (2001) Geometric approach to statistical analysis on the simplex. *Stochastic Environ Res Risk Assess (SERRA)* 15(5):384–398
- Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015) Modeling and analysis of compositional data. *Statistics in practice*. Wiley, Chichester, 272 pp
- Pearson K (1897) Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc R Soc Lond LX*:489–502
- van den Boogaart KG, Tolosana-Delgado R (2013) Analysing compositional data with R. Springer, Berlin, p 258

Computational Geoscience

Eric Grunsky
Department of Earth and Environmental Sciences, University of Waterloo, Waterloo, ON, Canada

Definition

Process – A phenomenon that is defined within a coordinate space with a valid metric. Processes can be defined by the geochemical composition of a material (e.g., mineral), combinations of minerals (rocks), or measures of rock assemblage properties including magnetic properties, electroconductivity, and density. Coherence in the form of geospatial continuity or an ordered and distinctive framework defines patterns that can be interpreted in a geochemical/geological context and subsequently tested as processes. Examining geochemical data on an element-by-element basis has been reviewed by McKinley et al. (2016). Multielement geochemical surveys provide insight into geochemical processes through the use of multivariate statistical methods.

Metric – A measure of a process within a coherent coordinate space. Different processes have different coordinate spaces and hence, different metrics. Commonly used metrics are derived from Hilbert spaces that are intrinsically orthonormal. The most commonly known metric is the Euclidian space. Non-orthogonal bases can also be used to discover/display features of a process. Such bases occur when

compositional data are in raw form or transformed using the additive logratio (alr), centered logratio (clr), isometric logratio (ilr), or pivot coordinates (pvc) Aitchison (1986), Egozcue et al. (2003). Despite the non-orthogonality of these bases, meaningful interpretations can be derived from the data and used to validate known processes or identify new processes (Greenacre et al. 2022). The relationships between the elements of geochemical data are controlled by “natural laws” (Aitchison 1986). In the case of inorganic geochemistry that law is stoichiometry, which governs how atoms are combined to form minerals, and thereby defines the structure within the data.

In the case of compositional data, additional measures are required in order to overcome the constant sum (closure) problem. Elements or oxides of elements are generally expressed as parts per million (ppm), parts per billion (ppb), weight percent (wt%, or simply %), or some other form of “proportion.” When data are expressed as proportions, there are two important limitations: first the data are restricted to the positive number space and must sum to a constant (e.g., 1,000,000 ppm, 100%), and second when one value (proportion) changes, one or more of the others must change too to maintain the constant sum. This problem cannot be overcome by selecting sub-compositions so that there is no constant sum. The “constant sum,” or “closure,” problem results in unreliable statistical measures. The use of ratios between elements, oxides, or molecular components that define a composition is essential when making comparisons between elements in systems such as igneous fractionation. The use of logarithms of ratios, or simply logratios, is required when measuring moments such as variance/covariance.

Depending on the data type, other metrics can be used, which can be: nominal, binary, ordinal, count, time, and interval.

Each of these metrics defines different ways of expressing the relationships between the variable (elements) and the observations (analyses). From a Process Discovery perspective, different metrics can assist in the identification of different processes.

The following coordinate spaces are common in the field of data analytics for geochemistry:

- Principal Component Analysis (PCA) – Jolliffe (1986)
- Independent Component Analysis (ICA) – Comon (1994)
- Multidimensional Scaling (MDS) – Cox (2001)
- Minimum-maximum Autocorrelation Factor Analysis (MAF) – Switzer and Green (1984)
- t-distributed Stochastic Neighbor Embedding (t-SNE) – van der Maaten and Hinton (2008)
- Uniform Manifold Approximation and Projection (UMAP) – McInnes et al. (2020)

Other metrics can be used to measure associations between variables and observations.

A two-step approach is recommended for evaluating multielement geochemical data. In the first “process discovery” step, patterns, trends and associations between observations (sample sites) and variables (elements) are extracted. Geospatial associations are also a significant part of process discovery. Patterns and/or processes that demonstrate geospatial coherence likely reflect an important geological/geochemical process.

Following process discovery, “process validation” is the step in which the patterns or associations are statistically tested to determine if these features are valid or merely coincidental associations. Patterns and/or associations that reveal lithological variability in surficial sediment, for instance, can be used to develop training sets from which lithology can be predicted in areas where there is uncertainty in the geological mapping and/or paucity of outcrop. Patterns and associations that are associated with mineral deposit alteration and mineralization may be predicted in the same way. In low-density geochemical surveys, where processes such as those related to alteration and mineralization are generally under-sampled, it may be difficult to carry out the process validation phase relating to these processes.

Censored Data

Quality assurance and quality control of geochemical data require that rigorous procedures be established prior to the collection and subsequent analysis of geochemical data. This includes the use of certified reference standards, randomization of samples, and the application of statistical methods for testing the analytical results. Historical accounts of Thompson and Howarth plots, for analytical results, can be found in Thompson and Howarth (1976a, b).

Geochemical data reported at less than the lower limit of detection (censored data) can bias the estimates of mean and variance; therefore, a replacement value that more accurately reflects an estimate of the true mean is preferred. Replacement values for censored geochemical data can be determined using several methods. The *lrEM* function from the R Computing Environment, *zCompositions* package (Palarea-Albaladejo et al. 2014) is suitable to use for estimating replacement values. Values greater than the upper limit of detection (ULD) can also be adjusted, based on a linear regression model formed by the uncensored data below the ULD.

Process Discovery

The following provides a few ways that processes can be discovered in multielement geochemical survey data.

Multivariate Data Analysis Techniques

The usefulness of multivariate data analysis methods applied to geochemical data has been well documented. An interpretation of the relationships of 30 elements is almost impossible without applying some techniques to simplify the relationships of elements and observations. Multivariate data analysis techniques such as those listed above provide numerical and graphical means by which the relationships of a large number of elements and observations can be studied Grunsky (2010) and Grunsky and Caritat (2019). These techniques typically simplify the variation and relationships of the data in a reduced number of dimensions, which can often be tied to specific geochemical/geological processes.

Incorporation of the spatial association with multielement geochemistry involves the computation of auto- and cross-correlograms or co-variograms. This field of study falls into the realm of geostatistics, which is not covered in this contribution. Multivariate geostatistics, which incorporates both the spatial and inter-element relationships, has been studied by only a few. Grunsky (2012) provides details on several approaches to account for geospatial variation in multivariate data.

The discovery of processes is typically carried out through the use of unsupervised methods within a multivariate framework in which patterns of the variables are detected. Unsupervised methods are also known as “unsupervised learning.” Pattern recognition depends on two essential elements: a reduction of noise in the system (signal/noise ratio) and the application of these methods at an appropriate scale of measurement (metric). The metric of measurement can be either linear or nonlinear. Linear metrics are useful for detecting patterns that are defined by linear processes such as stoichiometric control in mineralogy or linear combinations of minerals that comprise various rock types. Nonlinear processes may reflect processes that cannot be modeled using a linear metric. Geological processes such as mineral sorting through gravitational processes or chaotic processes (storm events) can be nonlinear. The scale of the metric may have an influence on the ability to determine linear/nonlinear processes. A multivariate approach is an effective way to start the process discovery phase. Linear combinations of elements that are controlled by stoichiometry may occur as strong patterns, while random patterns and/or under-sampled processes show weak or uninterpretable patterns.

An essential part of the process discovery phase is a suitable choice of coordinates to overcome the problem of closure. The centered logratio (clr) transformation (Aitchison 1986) is a useful transformation for evaluating geochemical data. Discovery of persistent and meaningful patterns enables the establishment of training sets for specific models to carry out process prediction. Models can include geology, mineral occurrences/prospects/deposits, soil types, and anthropogenic domains.

Statistical measures applied to geochemical data typically reveal linear relationships, which may represent the

stoichiometry of rock-forming minerals and subsequent processes that modify mineral structures, including hydrothermal alteration, weathering, and water–rock interaction. Physical processes such as gravitational sorting can effectively separate minerals according to the energy of the environment and mineral/grain density. Mineral chemistry is governed by stoichiometry and the relationships of the elements that make up minerals are easily described within the simplex, an n -dimensional composition within the positive real number space. It has long been recognized that many geochemical processes can be clearly described using element/oxide ratios that reflect the stoichiometric balances of minerals during formation (e.g., Pearce 1968). Geochemical data, when expressed in elemental or oxide form, can be a proxy for mineralogy. If the mineralogy of a geochemical data set is known, then the proportions of these elements can be used to calculate normative minerals.

Comprehension of the patterns that are revealed in process discovery include a wide range of graphics and tables. The relative relationships between variables and observations can be observed through the different multivariate coordinate systems generated by metrics including:

- Principal Component Analysis (PCA)
- Independent Component Analysis (ICA)
- t -distributed Stochastic Neighbor Embedding
- Minimum/Maximum Autocorrelation Factor analysis (MAF)
- Multidimensional Scaling (MDS)
- Self-Organizing Maps (SOM)
- Neural Networks (NN)
- Random Forests (RF)
- Blind Source Separation
- Hierarchical/nonhierarchical/model-based clustering
- Fractal/Multi-fractal analysis

From a process discovery perspective, patterns generated by these coordinate systems that show geospatial coherence and/or positive/negative associations between the variables, may be useful in describing processes. The commonly used principal component coordinates of clr-transformed data are orthonormal (i.e., statistically independent) and can reflect linear processes associated with stoichiometric constraints.

Fractal Methods

Fractal mathematics is now commonly used in the geosciences. Cheng and Agterberg (1994), Cheng and Zhao (2011) showed how fractal methods can be used to determine thresholds of geochemical distributions on the basis of spatial patterns of abundance. Where the concentration of a particular component per unit area satisfies a fractal or multifractal model, the area of the component follows a power law relationship with the concentration. This can be expressed mathematically as:

$$A(\rho \leq v) \propto \rho^{-\alpha_1} \quad A(\rho > v) \propto \rho^{-\alpha_2}$$

where $A(\rho)$ denotes an area with concentration values greater than a contour (abundance) value greater than ρ . This also implies that $A(\rho)$ is a decreasing function of ρ . If v is considered the threshold value then the empirical model shown above can provide a reasonable fit for some of the elements.

In areas where the distribution of an element represents a continuous single process (i.e., background) then the value of α remains constant. In areas where more than one processes have resulted in a number of superimposed spatial distributions, there may be additional values of α defining the different processes.

The use of power-spectrum methods to evaluate concentration-area plots derived from geochemical data has also been shown by Cheng et al. (2000), whereby the application of filters and patterns can be detected related to background and noise, thus enabling the identification of areas that are potentially related to mineralization.

Visualizing the Coordinates and Elemental Associations of a Multivariate Metric

The transformation of the logratio coordinates can be visualized in the transformed metric space and the geographic domain. Some of the transformations permit the calculation and visualization of the variables with the observations.

An important part of determining variable significance is the measure of variance accounted by each component (dimension).

- Principal Component Biplots
- Minimum-Maximum Autocorrelation Biplots

Exploring data geospatially involves the use of coordinate system, in either a geodetic or orthogonal grid/metric. Coordinate systems generated by methods such as MAF include measures of geospatial association (Mueller et al. 2020). Geospatial signatures derived from variables that have positive/negative associations over a geospatial domain enhance the geospatial aspect of multivariate associations.

The process of principal component analysis involves choosing the appropriate metric (i.e., logcentered transform) from which eigenvalues/eigenvectors are computed using singular value decomposition. The resulting eigenvalues, plotted as a “screeplot,” indicates how much “structure” is in the data. Principal component biplots display the relative relationships between the observations (principal component scores) and the variables (principal component loadings). Biplots can reveal much information about the relative relationships between the variables and the sample population. The resulting principal components can also be plotted in the geospatial domain, whereby the principal component patterns can be visually assessed in terms of the geospatial relationships between geological/geochemical processes. Areas of

contiguous zones of similar principal component scores can show geospatial patterns that define the “geospatial coherence” of the data in a geographic sense.

Geospatial Coherence

Multivariate coordinate systems have coordinates based on metrics of the variables (elements) and the observations (geochemical analyses). Different coordinate systems display different relationships of the data. These coordinate systems provide insight into the discovery of processes. For geospatial geochemical data, geospatial coherence is defined as both local and regional continuity of a response based on geostatistical methods such as semi-variograms.

If a geospatial rendering of a principal component shows no spatial coherence (i.e., no structure, or a lot of “noise”), then it is likely that the image will be difficult to interpret within a geological context. The most effective way to test this is through the generation and modeling of semi-variograms that describe the spatial continuity of a specific class based on prior or posterior probabilities (PPs), which are described below. If meaningful semi-variograms can be created, then geospatial maps of PPs can be generated through interpolation using the kriging process. Maps of PPs may show low overall values but still be spatially coherent. This is also reflected in the classification accuracy matrix that indicates the extent of classification overlap between classes. Geospatial analysis methodology described by the “gstat” package (Pebesma 2004) in R can be used to generate the geostatistical parameters and images of the PCs and PPs from kriging.

If the spatial sampling density appears to be continuous then it may be possible to carry out spatial prediction techniques such as spatial regression modeling and kriging. A major difficulty with the application of spatial statistics to regional geochemical data is that the data seldom exhibit stationarity. Stationarity means that the data has some type of location invariance, that is, the relationship between any sets of points are the same regardless of geographic location. This is seldom the case in regional geochemical datasets. Thus, interpolation techniques such as kriging must be applied cautiously, particularly if the data cover several geochemical domains in which the same element has significantly different spatial characteristics.

Evaluation of the variogram or the autocorrelation plots can provide insight about the spatial continuity of an element. If the autocorrelation decays to zero over a specified range, then this represents the spatial domain of a particular geological process associated with the element. Similarly, for the variogram, the range represents the spatial domain of an element, which reaches its limit when the variance reaches the “sill” value, the regional variance of the element. Theoretically, at the origin, the variance should be zero at lag zero. However, typically, an element may have a significant degree of variability even at short distances from neighboring points. This variance is termed the nugget effect.

Skill, knowledge, and experience are required to effectively use geostatistical techniques. It requires considerable effort and time to effectively model and extract information from geospatial data. The benefits of these efforts are a better understanding of the spatial properties of the data, which permits a more effective predictive estimate of geochemical trends. However, they must be used and interpreted with the awareness about problems with techniques of interpolation and the spatial characteristics of the data.

Recognizing Rare Events – Under-Sampled Processes in Principal Component Analysis

In regional geochemical surveys, the sample design is not optimally suited for the recognition of geological processes that have small geospatial footprints such as a mineral deposit. Thus, sites that are associated with mineral deposits are under-sampled, relative to the dominant processes defined by background lithologies. A principal component analysis may identify these “rare event” processes in the lesser eigenvalues/eigenvectors. Examination of biplots and geospatial plots of these lesser components can help identify mineral exploration targets or zones of contamination.

Process Validation – Testing Models Derived from Process Discovery

Process Validation – Process validation is the statistical measure of the likelihood of a given process as defined by discovery or by definition. When processes are defined by discovery, as described above, or by training sets defined by other means, the application of process validation provides a statistical measure of uniqueness of the process.

Using different multivariate coordinate systems (metrics), geochemical data can be classified using a variety of classification methods. The combination of coordinate systems and measures of proximity for classification results in a range of predictions. In a multivariate normal dataset, the results of the different combinations of coordinate systems and classifications can be similar. In datasets where there are multiple and distinct processes, the results may vary widely and provide different interpretations.

Classification can be carried out as a binary or multi-class process. For many geoscience applications that use multi-element geochemical data, multi-class classification studies are more common.

Bayesian methods produce predictive probabilities that are based on prior probabilities and are used in many classification algorithms.

A non-exhaustive list of classification/prediction methods include:

- Quadratic Discriminant Analysis (QDA)
- Logistic Regression (LR)
- Neural Networks (NN)
- Support Vector Machines (SVM)
- Boosting
- Classification and Regression Trees (CART)

Process validation is the methodology used to verify that a geochemical composition (response) reflects one or more processes. These processes can represent lithology, mineral systems, soil development, ecosystem properties, climate, or tectonic assemblages. Validation can take the form of an estimate of likelihood that a composition can be assigned membership to one of the identified processes. This is typically done through the assignment of class identifier or a measure of probability. Training sets contain the classes used to build a prediction model. The proportion of each class relative to the sum of all of the classes is termed the prior probability. After the application of the prediction model to the training set, and/or the test dataset, for each observation, there is list of proportional values of each class, which are termed the posterior probabilities (PP).

A critical part of process validation is the selection of variables that produce an effective classification. This requires the selection of variables that maximize the differences between the various classes and minimizes the amount of overlap due to noise, unrecognized or under-sampled processes in the data. As stated previously, because geochemical data are compositional in nature, the variables that are selected for classification require transformation to logratio coordinates. The alr or ilr are both effective for the implementation of classification procedures. The clr-transform is not suitable because the covariance matrix of these coordinates is singular. However, analysis of variance (ANOVA) applied to clr-transformed data enables the recognition of the compositional variables (elements) that are most effective at distinguishing between the classes. Choosing an effective alr-transform (choice of suitable denominator) or balances for the ilr-transform can be challenging and requires some knowledge and insight about the nature of the processes being investigated. The choice of a suitable divisor for the alr transform can be based on prior knowledge of the processes that involve specific elements. ANOVA applied to the PCs derived from the clr-transform has been shown to be highly effective at discriminating between the different classes. Typically, the dominant PCs (PC1, PC2, ...) commonly identify dominant processes and the lesser components (PCn, PCn-1, ..., where n is the number of variables) may reflect under-sampled processes or noise. The use of the dominant components can be interpreted as a technique to increase the signal/noise ratio and result in a more effective classification using only a few variables. Classification results can be expressed as direct class assignment or posterior probabilities (PPs) in the form of forced class allocation, or as class typicality. Forced

- Random Forests (RF)
- Linear Discriminant Analysis (LDA)

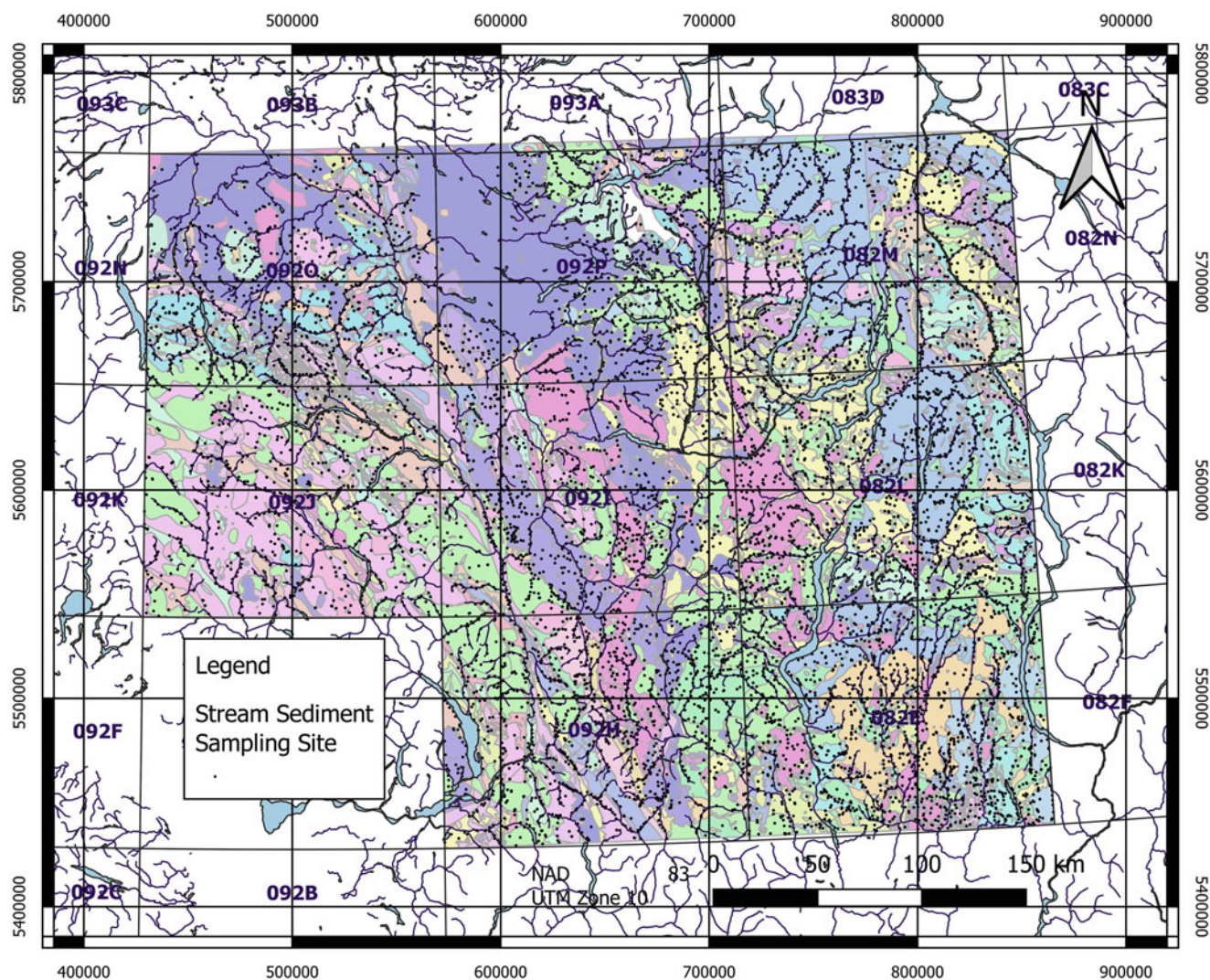
class allocation assigns a PP based on the shortest Mahalanobis distance of a compositional observation from the compositional centroid of each class. Class typicality measures the Mahalanobis distance from each class and assigns a PP based on the F-distribution. This latter approach can result in an observation having a zero PP for all classes, indicating that its composition is not similar to any of the compositions defined by the class compositional centroids.

The application of a procedure such as LDA can make use of cross validation procedures, whereby the classification of the data is repeatedly run based on random partitioning of the data into a number of equal sized subsamples. One subsample is retained for validation and the remaining subsamples are used as training sets. Classification accuracies can be assessed through the generation of tables that show the accuracy and errors measured from the estimated classes against the initial classes in the training sets used for the classification.

Class imbalance occurs when there is a disproportionate number of samples for the different classes being considered in a model. For example, in predictive lithology study, for the creation of a training set, there may be large number of observations that represent granitic rocks and a small number of samples that reflect volcanic rocks. This class imbalance can affect the resulting model from many of the classification methods. Class imbalance can be adjusted using methods such as the Synthetic Minority Oversampling Technique (Hariharan et al. 2017).

Example of Process Discovery and Process Prediction of Mineral Resource Potential Using Geochemical Survey Data

Regional geochemical survey data (Fig. 1) from southern British Columbia, Canada were studied for the potential of



Computational Geoscience, Fig. 1 Location map of stream sediment sampling sites (black dots) overlying a map of the regional geology. (From Grunsky and Arne 2020, Fig. 1)

discovering/prediction specific types of mineral deposits (Grunsky and Arne 2020), based on mineral deposit models that have been developed for the province (BC Geological Survey 1996). The study integrated the regional geochemical survey with a mineral occurrence database (MINFILE) comprised of active mines, past producers, developed prospects, prospects, occurrences, and anomalies (BC Geological Survey 2019) as seen in Fig. 2. Each mineral occurrence was assigned a specified mineral deposit model. The mineral occurrences are not colocated with the regional stream sediment survey data. Thus, based on the distance between a stream sediment sampling site and a mineral occurrence (Fig. 3), the attributes of the mineral occurrence were assigned to the stream sediment sample site. If the distance between the two sites exceeded 3000 m, no assignment was made and the mineral deposit model attribute was classed as “unknown.” Additionally, if the MINFILE record was identified as a “showing” and the distance from the stream sediment site to the MINFILE site was

greater than 1000 m, the stream sediment site mineral deposit model was classed as “unknown.” Some mineral deposit types, listed in Table 1, were not included due to the distance selection criteria.

Mineral deposit models with less than ten MINFILE records were not included. Also, the mineral deposit model (I05 – Polymetallic veins Ag-Pb-Zn ± Au) were not included because the geochemistry of the stream sediment sites associated with I05 overlap with almost every other mineral deposit type. It should be noted that the geochemistry of the mineral deposit models using stream sediment geochemistry is not unique and that compositional overlap is a problem when using geochemistry alone. The resulting deposit models used in the process prediction/validation part of the study are highlighted in bold in Table 1.

Process Discovery

The geochemical survey data were screened for censoring and values were imputed for element values less than the lower

Minfile sites

GroupModels

Stream sediment site

Sample Site

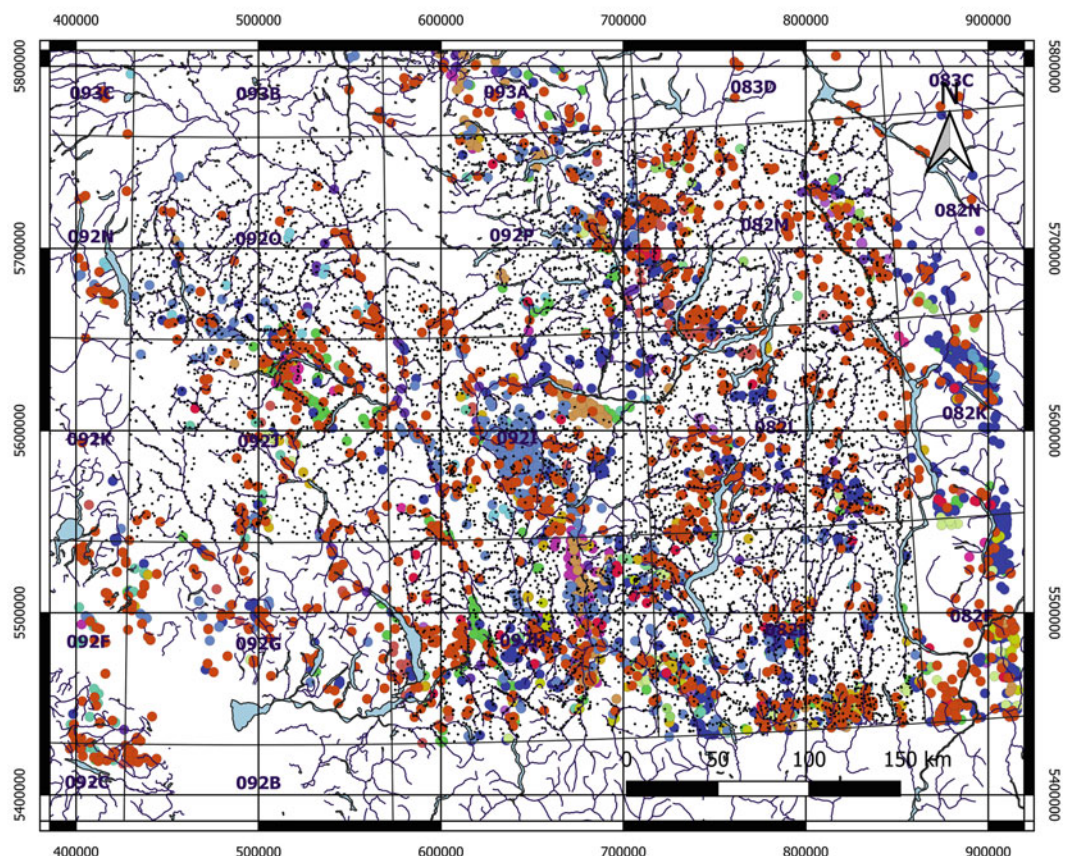
MINFILE Deposit Model

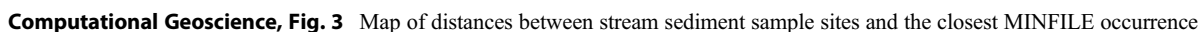
- C01
- D03
- E12
- G04
- G06
- H02
- H05
- I01
- I02
- I05
- I06
- I09
- J01
- K01
- K02
- K04
- K05
- L01
- L03
- L04
- L05
- unknown

— River

□ NTS boundary

Datum: NAD 83
Zone: UTM 10





patterns and relationships of the variables (PCA only) and the observations. Geospatial rendering, via kriging of the transformed variables (e.g., PC1, t-SNE1) yielded maps for the identification of patterns. Geospatially coherent patterns were considered to represent geochemical processes related to bedrock, alteration, mineralization, weathering, and mass transport.

The application of principal component analysis yielded eigenvalues/eigenvectors that showed significant “structure” in the data. Figure 4 shows a screeplot and a table for the first ten eigenvalues, cumulative eigenvalues, and cumulative percentage eigenvalues. The first four eigenvalues represent most of the variability (56.9%) of the data, while the lesser values may represent under-sampled or random processes.

Figure 5 shows a principal component biplot PC1-PC2 with the observations coded by the attributes of geological terrane, bedrock lithologies, and mineral deposit model assignment at each stream sediment site. Mean values for each of the attributes are shown by the larger symbols on separate biplots, which indicate the compositional difference of the attributes across the PC1-PC2 space.

Computational Geoscience, Table 1 Frequency of MINFILE mineral deposit models in the study area

Mineral deposit model	Description	Frequency
C01	Surficial placers	118
D03	Volcanic redbed Cu	115
E05	Mississippi Valley-type Pb-Zn	2
E12	Mississippi Valley-type Pb-Zn	12
E13	Irish-type carbonate-hosted Zn-Pb	5
E15	Blackbird sediment-hosted Cu-Co	1
G04	Besshi massive sulfide Cu-Zn	34
G05	Cyprus massive sulfide Cu (Zn)	6
G06	Noranda/Kuroko massive sulfide Cu-Pb-Zn	102
G07	Subaqueous hot spring Ag-Au	2
H02	Hot spring Hg	13
H03	Hot spring Au-Ag	2
H04	Epithermal Au-Ag-Cu; high sulfidation	7
H05	Epithermal Au-Ag; low sulfidation	44
H08	Alkalic intrusion-associated Au-Ag	10
I01	Au-quartz veins	311
I02	Intrusion-related Au pyrrhotite veins	30
I05	Polymetallic veins Ag-Pb-Zn $\pm$ Au	988
I06	Cu $\pm$ Ag quartz veins	82
I08	Silica-Hg carbonate	4
I09	Stibnite veins and disseminations	24
J01	Polymetallic manto Ag-Pb-Zn	11
J04	Sulfide manto Au	2
K01	Cu skarns	151
K02	Pb-Zn skarns	28
K04	Au skarns	54
K05	W skarns	25
K07	Mo skarns	5
L01	Subvolcanic Cu-Ag-Au (As-Sb)	61
L02	Porphyry-related Au	7
L03	Alkalic porphyry Cu-Au	198
L04	Porphyry Cu $\pm$ Mo $\pm$ Au	384
L05	Porphyry Mo (low F-type)	52
Unknown		1218

Models highlighted in bold were used in the classification/prediction

Figure 6 shows a scatterplot matrix of the nine independent components (ICA), coded by the attributes of lithology. There is a clear distinction between granitic and metamorphic lithologies throughout the scatterplots. The volcanic and clastic rock types are transitional between the two igneous/metamorphic lithologies.

Figure 7a, b show t-SNE coordinates 3 and nine coded by the attributes of underlying bedrock lithology, mineral

deposit model assignment at each stream sediment site. Mean values for each bedrock lithology and mineral deposit model are shown as larger symbols across the t-SNE3–t-SNE9 space.

The use of the stream sediment geochemistry PCA, ICA, and t-SNE coordinates and the MINFILE mineral deposit model attributes form the basis of process discovery from which the multielement associations and zones of geospatial coherence form a training set (provisional model) from which mineral deposit models can be predicted based.

AOV

An analysis of variance (AOV) was carried out on the PCA space using the tagged mineral deposit models to each stream sediment site. One hundred stream sediment sites with mineral deposit models tagged as “unknown” were included in the analysis. Figure 8 shows a plot of ordered F-values for each principal component. Components with high F-values are better at discriminating between the different mineral deposit types based on 41 PCs. PCs 1, 3, 8, and 13 are the dominant components for mineral deposit discrimination. In the case of the 9-dimensional t-SNE coordinates, t-SNE9, t-SNE3, t-SNE5, and t-SNE4 accounted for most of the variability of the data for the geochemistry tagged with a mineral deposit model. An AOV for the ICA coordinate space indicates that ICA2, ICA5, and ICA9 account for most of the mineral deposit class separation.

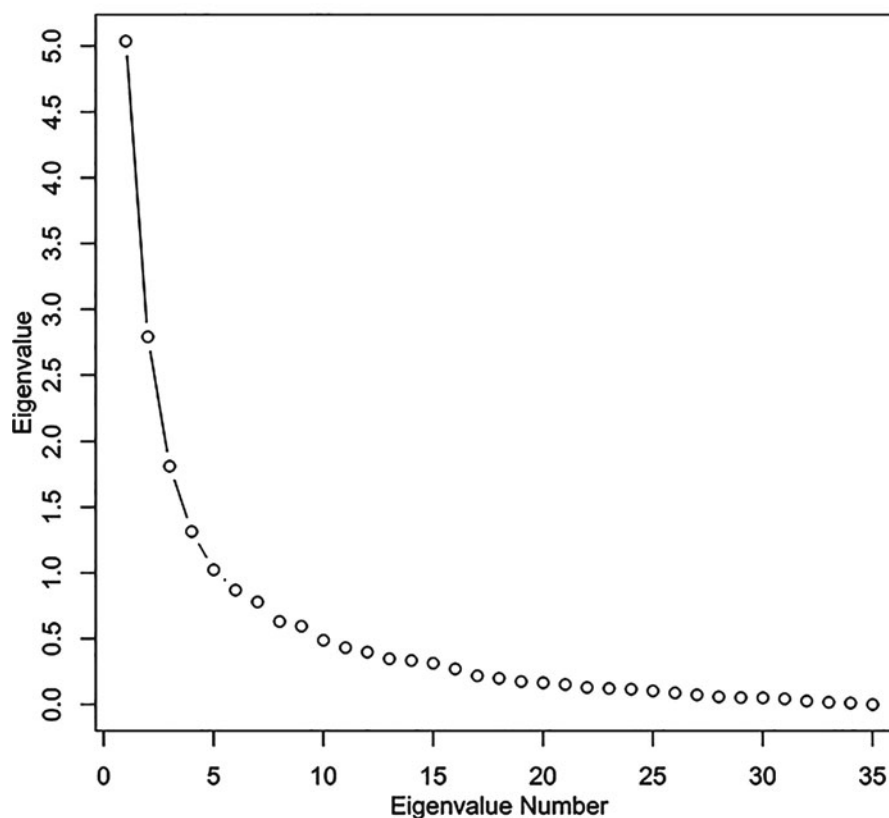
Geospatial Coherence in Process Discovery

Figure 9a, b show kriged images for principal components 1 and 3. The stream sediment sampling sites are shown on the images. The choice of these components for geospatial rendering is based on the analysis of variance of the principal components and the discrimination of the mineral deposit models assigned to the geochemical data. The kriged images of PC1 and PC3 show broad geospatially coherent regions that represent associations of elements and sampling sites as shown in the biplot of Fig. 5. Negative PC1 scores represent granitic/felsic rock types and positive scores indicate mafic and sedimentary rock types. Figure 10a, b show kriged images for t-SNE coordinates 3 and 7. The choice of displaying these maps is based on the application of random forests for the prediction of mineral deposit types.

Figure 11a, b show kriged ICA coordinates 2 and 5 that shows broad regional patterns that reflect underlying lithologies across the area. These two variables were determined the most significant for discriminating between the mineral deposit classes.

The kriged images of selected coordinates of the three different metrics show zones of geospatial continuity

Quest South Stream Sediment Geochemistry



Computational Geoscience, Fig. 4 Screeplot of eigenvalues derived from a principal component analysis of logcentered transformed geochemical data. The chart below lists the eigenvalues, percent contribution of the variance, and the cumulative contribution of the variance for each

indicating that these variables represent geospatially coherent processes, which will be tested using methods of classification and prediction.

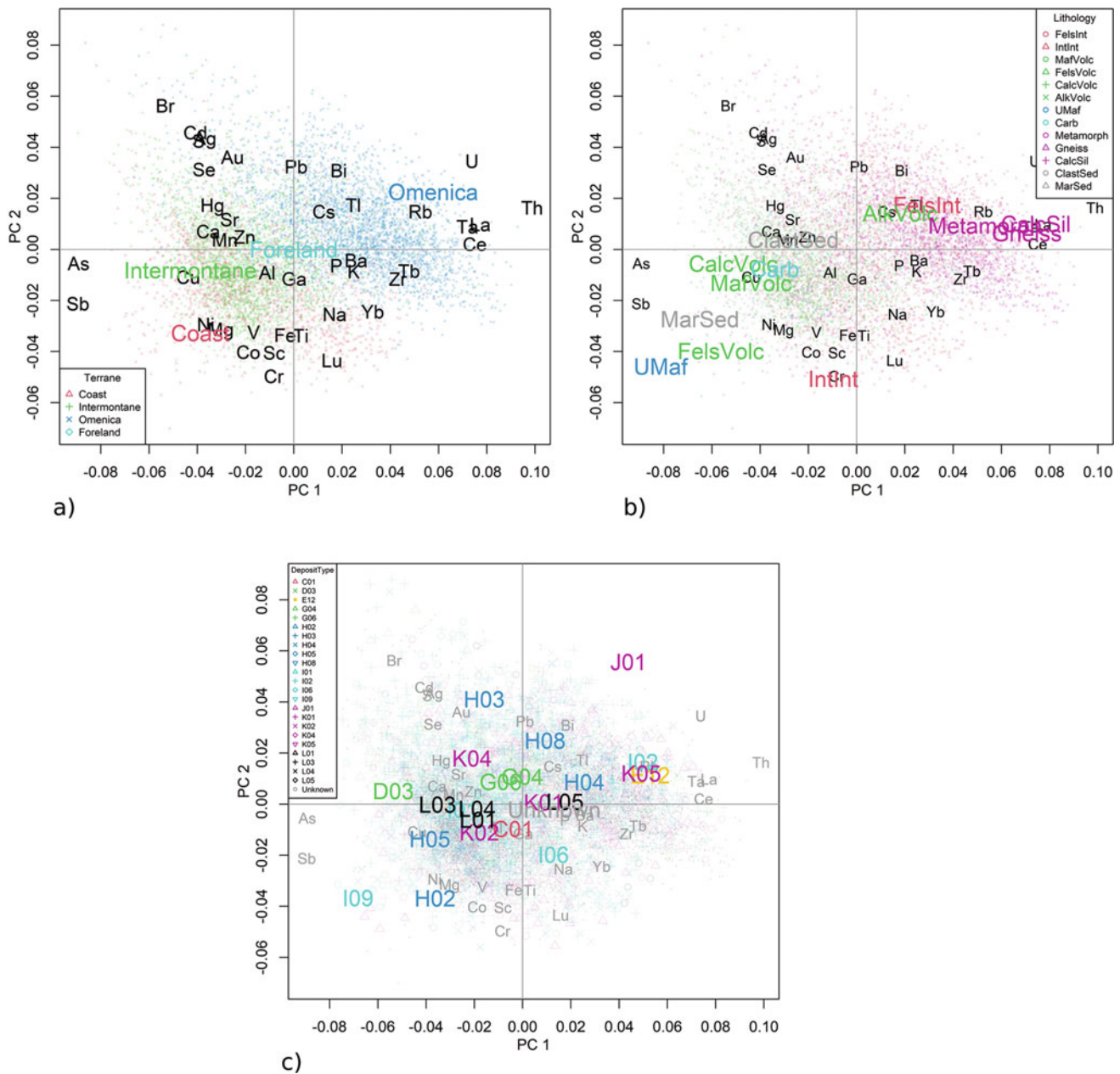
Process Validation

An initial step in process validation requires an investigation on the ability of the training set of stream sediment geochemistry sites using PCA, t-SNE, and ICA metrics, tagged with a mineral deposit model, to predict the mineral deposit type for stream sediment sites where the mineral deposit model is tagged as unknown. Part of the strategy for use of the metrics is to choose principal components, t-SNE, and ICA coordinates that account for most of the observed signals in PCA biplots and geospatial maps of the spaces. This approach increases the signal-to-noise ratio, which increases the ability to predict mineral deposit types. Stream sediment sites tagged

with a mineral deposit model formed the training set, plus the additional 100 sites tagged as “unknown,” were assigned to the training set.

As described above, there are several methods to classify and predict processes. Three examples are shown here: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and random forests (RF). Because geochemical processes can be linear, nonlinear, or a combination of both, it is useful to classify and predict processes with more than one method. This approach provides a measure of the variability of the data and identifies processes that are variable dependent on the method used and an indirect measure of uncertainty.

Table 2 and Fig. 12 show the prediction accuracies for each of the three metrics and three classification methods based on the training sets as described above. The table and figures

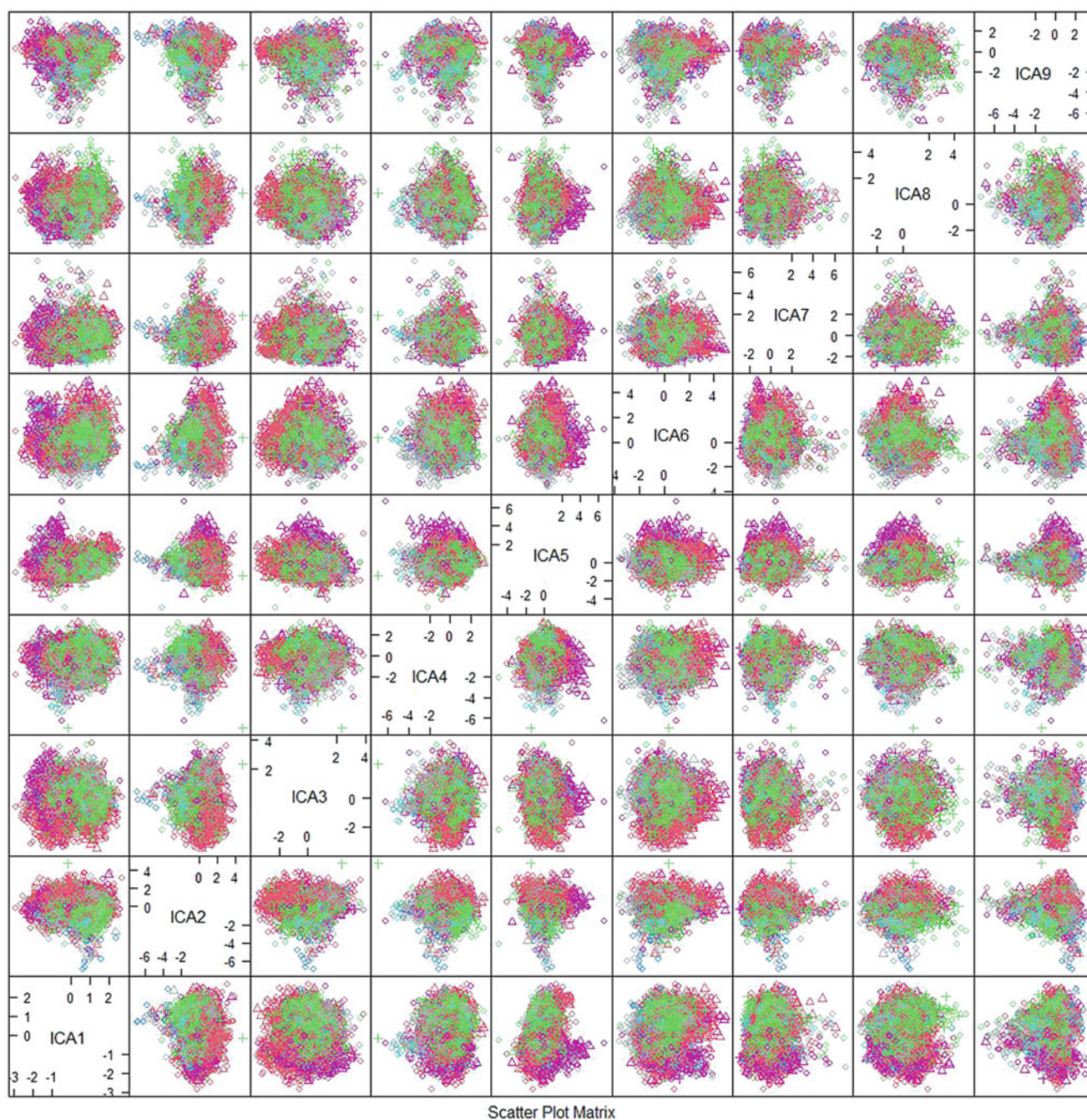


Computational Geoscience, Fig. 5 Three PCA biplots of PC1.v. PC2. (a) Scores coded by geological terrane. Mean values for geological terranes are shown by the terrane label placement in the biplot. (b) Scores coded by rock type. Mean values of the different rock types are shown by

the label placement in the biplot. (c) Scores coded by MINFILE mineral deposit model. Mean values for each of the mineral deposits models are shown by the label placement in the biplot

show that some of the mineral deposit models have higher prediction accuracies than others. The mineral deposit models C01 (placer gold), I01 (Au quartz veins), L03 (alkalic porphyry Cu-Au), L04 (porphyry Cu), and L05 (porphyry Mo) have the highest prediction accuracies for all three metrics. Quadratic discriminant analysis (Fig. 12b) and random forests (Fig. 12c) have higher prediction accuracies than those based

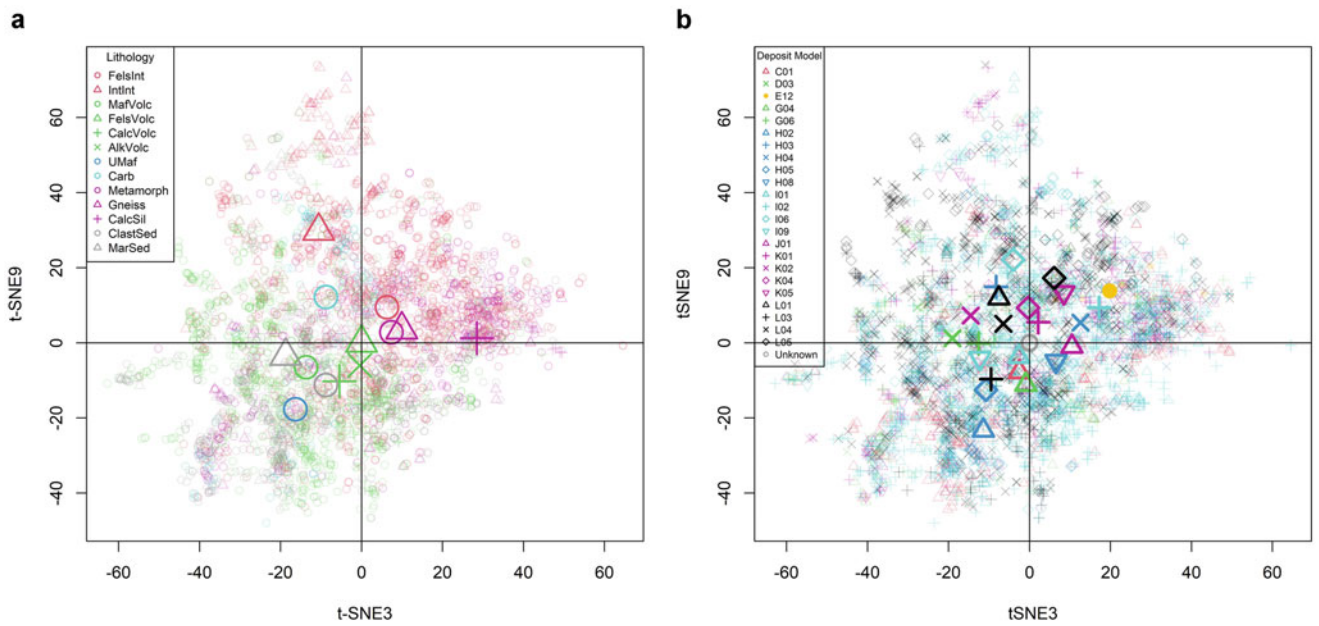
on predictions using linear discriminant analysis (Fig. 12a). Figure 12d shows the overall prediction accuracies for the three metrics and three methods. The figure shows that the method of random forests provides the highest level of prediction accuracy. Quadratic discriminant analysis outperforms linear discriminant analysis with the exception of the ICA metric.



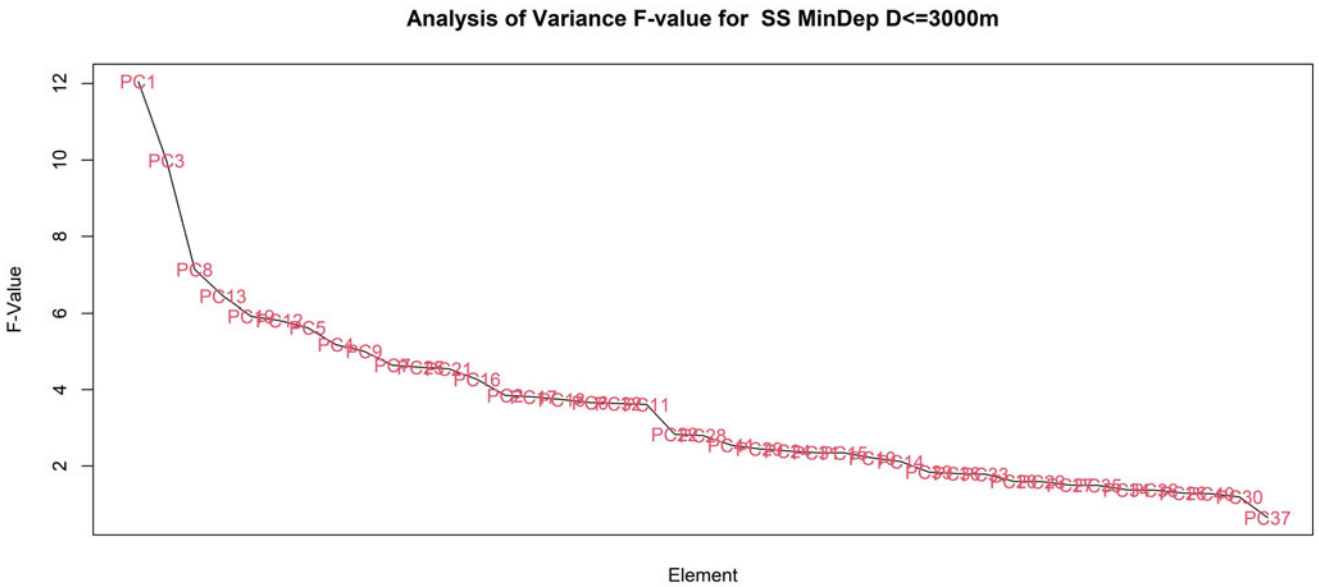
Computational Geoscience, Fig. 6 Scatterplot matrix of the ICA coordinates coded by rock type. Legend is the same as Fig. 5b

Figure 13a–f show kriged images of the posterior probabilities generated for mineral deposit predictions using random forests on the three metrics. Each map shows the MINFILE occurrences, coded as yellow crosses, designated for the mineral deposit models C01 (placer Au) and L04

(porphyry Cu-Au). Each stream sediment site that was predicted as C01 (Fig. 13a, c, e) and L04 (Fig. 13b, d, f) are shown as red dots. The kriged images reflect prediction values that range between 0 and 1. The geospatial coherence of the images and the strength of the kriged prediction surfaces



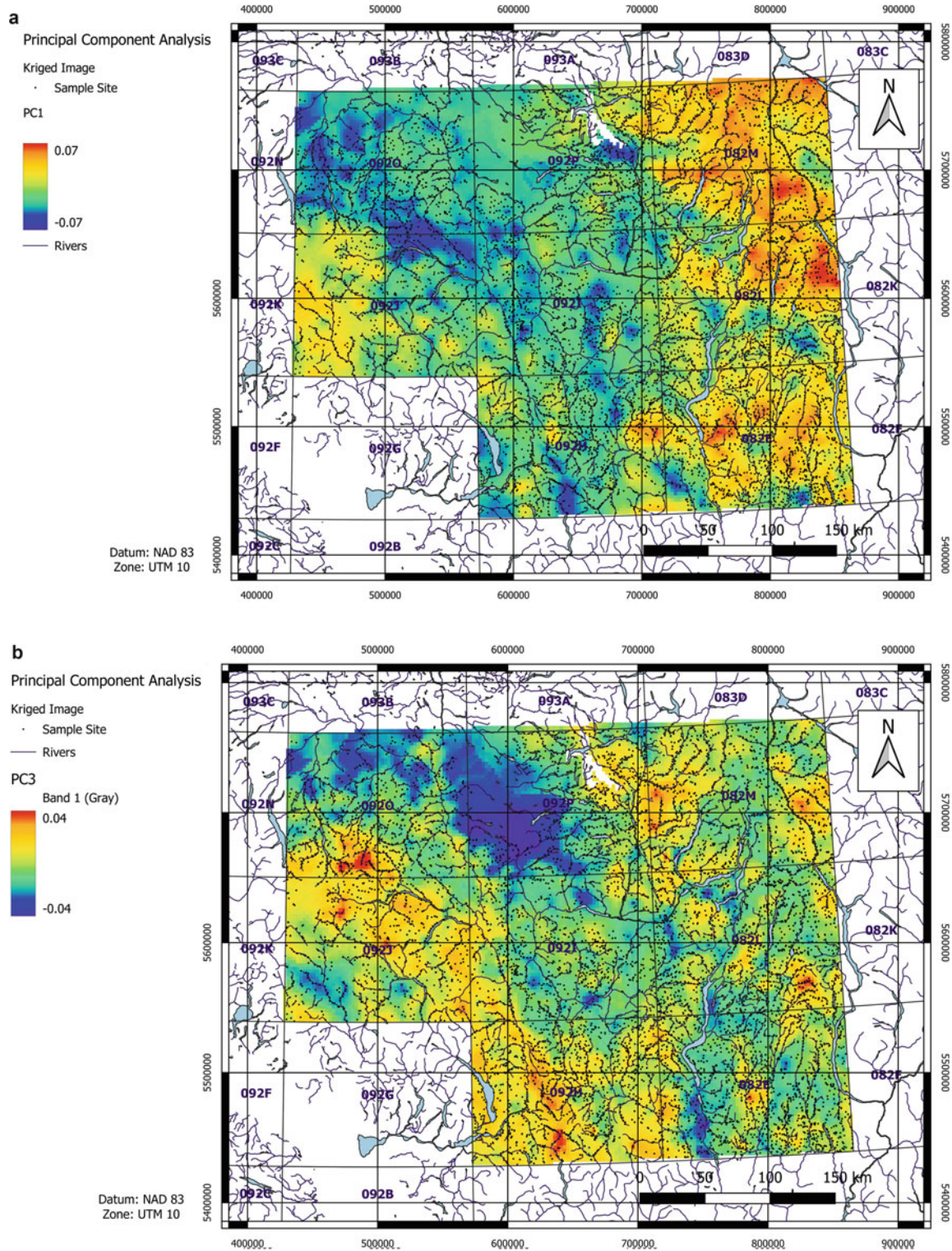
Computational Geoscience, Fig. 7 (a) Scatterplot of t-SNE3.v.t-SNE5 coded by rock type. (b) Scatterplot of t-SNE3.v.t-SNE5 coded by mineral deposit model



Computational Geoscience, Fig. 8 Plot of F-values obtained from an analysis of variance for the mineral deposit models based on the principal components

indicate that the use of the t-SNE metric results in a prediction that is better than that of the ICA or PCA metric. Similarly, the classification for individual stream sediment sites show that fewer sites are predicted for the PCA metric, while the t-SNE

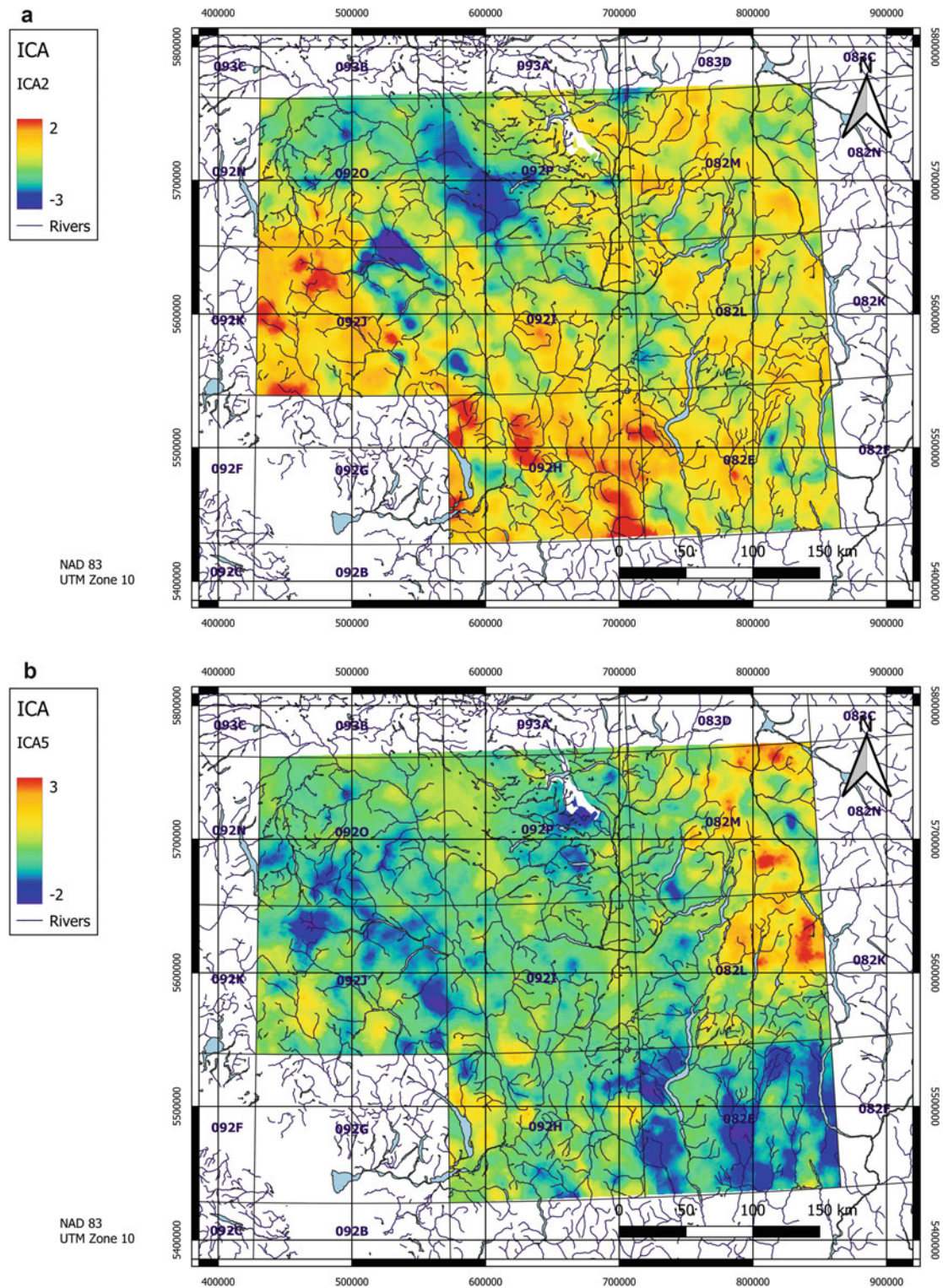
metric shows a much larger number of sites predicted for both the C01 and L04 deposit models. The increased number of predicted sites and the kriged probability surfaces for both models is partly supported by the addition of the known



Computational Geoscience, Fig. 9 (a) Kriged image of principal component 1. Broad regional domains indicating geospatial coherence are evident. The eastern part of the map shows dominantly positive PC1

scores. The central part of the map shows mainly negative PC1 scores. (b) Kriged image of PC3 showing regions of geospatial coherence

Computational Geoscience, Fig. 10 (a) Kriged image of t-SNE component 3. (b) Kriged image of t-SNE component 7. The two figures show broad regions of geospatial coherence



Computational Geoscience, Fig. 11 (a) Kriged image of ICA component 2. (b) Kriged image of ICA component 5. The two figures show broad regions of geospatial coherence

Computational Geoscience, Table 2 Prediction accuracies for three metrics and three classification methods

LDA	C01	D03	G06	H05	I01	I06	K01	K04	L03	L04	L05	Unknown
PCA	22.72	29.41	18.75	14.81	36.55	13.33	2.56	0	56	1.49	37.5	0.57
ICA	48.18	29.41	10.41	3.70	38.70	13.33	10.25	0	8	59.70	12.5	57
tSNE	30.90	5.88	2.08	18.51	0	13.33	0	0	60	2.23	12.5	0
QDA	C01	D03	G06	H05	I01	I06	K01	K04	L03	L04	L05	Unknown
PCA	33.63	17.64	16.66	25.92	36.55	0	25.64	9.09	16	51.49	0	49
ICA	39.09	17.64	18.75	11.11	37.63	6.66	17.94	0	40	58.20	6.25	52
tSNE	31.81	17.64	31.25	40.74	30.10	20	28.20	18.18	48	43.28	31.25	70
RF	C01	D03	G06	H05	I01	I06	K01	K04	L03	L04	L05	Unknown
PCA	59.78	0	20.49	14.36	42.74	0	12.54	0	0	69.24	0	74.81
ICA	52.50	5.57	10.22	14.36	32.02	0	15.05	0	0	67	0	59.76
tSNE	53.41	16.83	49.48	50.94	44.89	18.98	32.77	16.92	27.21	61.76	29.96	77.82
PCA	C01	D03	G06	H05	I01	I06	K01	K04	L03	L04	L05	Unknown
LDA	22.72	29.41	18.75	14.81	36.55	13.33	2.56	0	56	1.49	37.5	0.57
QDA	33.63	17.64	16.66	25.92	36.55	0	25.64	9.09	16	51.49	0	49
RF	59.78	0	20.49	14.36	42.74	0	12.54	0	0	69.24	0	74.81
ICA	C01	D03	G06	H05	I01	I06	K01	K04	L03	L04	L05	Unknown
LDA	48.18	29.41	10.41	3.70	38.70	13.33	10.25	0	8	59.70	12.5	57
QDA	39.09	17.64	18.75	11.11	37.63	6.66	17.94	0	40	58.20	6.25	52
RF	52.50	5.57	10.22	14.36	32.02	0	15.05	0	0	67	0	59.76
tSNE	C01	D03	G06	H05	I01	I06	K01	K04	L03	L04	L05	Unknown
LDA	30.90	5.88	2.08	18.51	0	13.33	0	0	60	2.23	12.5	0
QDA	31.81	17.64	31.25	40.74	30.10	20	28.20	18.18	48	43.28	31.25	70
RF	53.41	16.83	49.48	50.94	44.89	18.98	32.77	16.92	27.21	61.76	29.96	77.82
Overall accuracy												
LDA.PCA	LDA. ICA	LDA. TSNE	QDA. PCA	QDA. ICA	QDA. TSNE	RF. PCA	RF. ICA	RF. TSNE				
16.06	38.90	9.92	34.96	38.11	39.84	45.49	45.49	45.49				

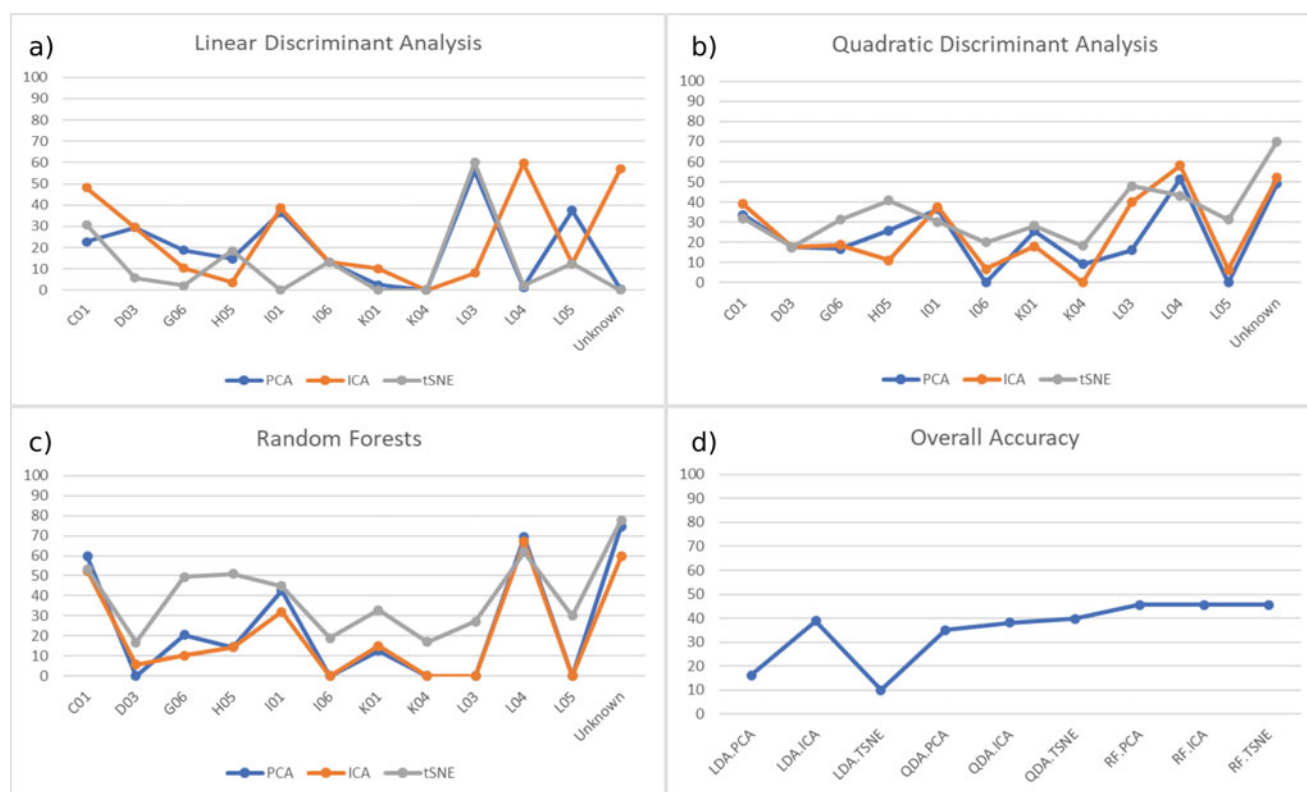
MINFILE occurrences shown in yellow crosses. The areas of increased probability and/or sampling sites predicted as C01 or L04, but with few or no MINFILE occurrences may represent regions of previously unrecognized sites with increased mineral deposit potential, or they reflect compositional overlap between deposit types and the site may be mineralized but belong to a different mineral deposit class.

Within the current scope and context of this study, some fundamental assumptions have been made as detailed in Grunsky and Arne (2020).

The geochemical composition of the stream sediment associated with individual mineral deposit models is uniquely distinct. In some cases, this assumption is not warranted. As described above, the mineral deposit model I05 (polymetallic veins) has characteristics that overlap with many other mineral deposit types, which resulted in a significant amount of confusion when applying process prediction. As a result, mineral deposit models with a significant amount of overlap with other models should be removed.

Not all mineral deposit types can be best represented with stream sediment geochemistry. The size fraction and the analytical methods used may not extract unique information to distinguish a sub-cropping mineral deposit or distinguish between different mineral deposit types, as in the case for the polymetallic vein class (I05). The method of dissolution using aqua regia is useful for sheet silicates and sulfide minerals, but aqua-regia digestion does not dissolve many other forms of silicates. Thus, some unique geochemical aspects of specific mineral deposit types based on silicate mineral assemblages may not be recognized.

The MINFILE model identification is accurate. This may not be the case for some types of mineral systems and, as a result, there will be an increase in confusion of prediction. The identification of the mineral deposit profiles, as specified in the MINFILE field “Deposit Type,” may be incorrect or inconclusive, or the locations may be inaccurate. This can lead to misclassification errors in the subsequent application of machine-learning prediction methods.



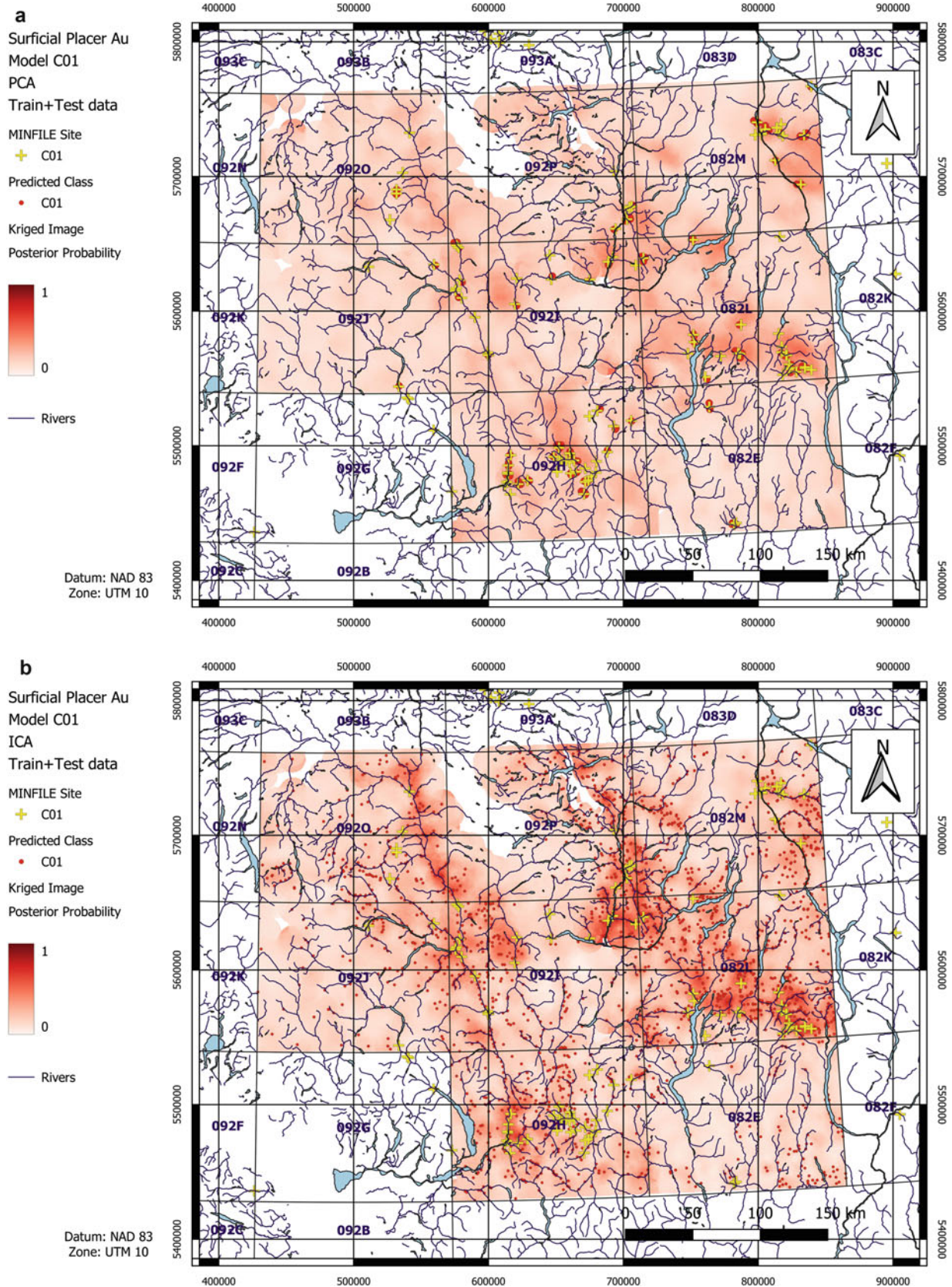
Computational Geoscience, Fig. 12 Charts that show the predictive accuracy of each mineral deposit type based on linear discriminant analysis (a), quadratic discriminant analysis (b), and random forests (c). Figure (d) shows the overall accuracy of each prediction method and metric

The location of a MINFILE site and the associated stream sediment site may not be within the same catchment area. In the example presented here, the assumption was made that the effect of catchment is not significant. If there is a requirement for the location of a MINFILE site and associated stream sediment site to be in the same catchment, the number of sites for the training set would be significantly reduced.

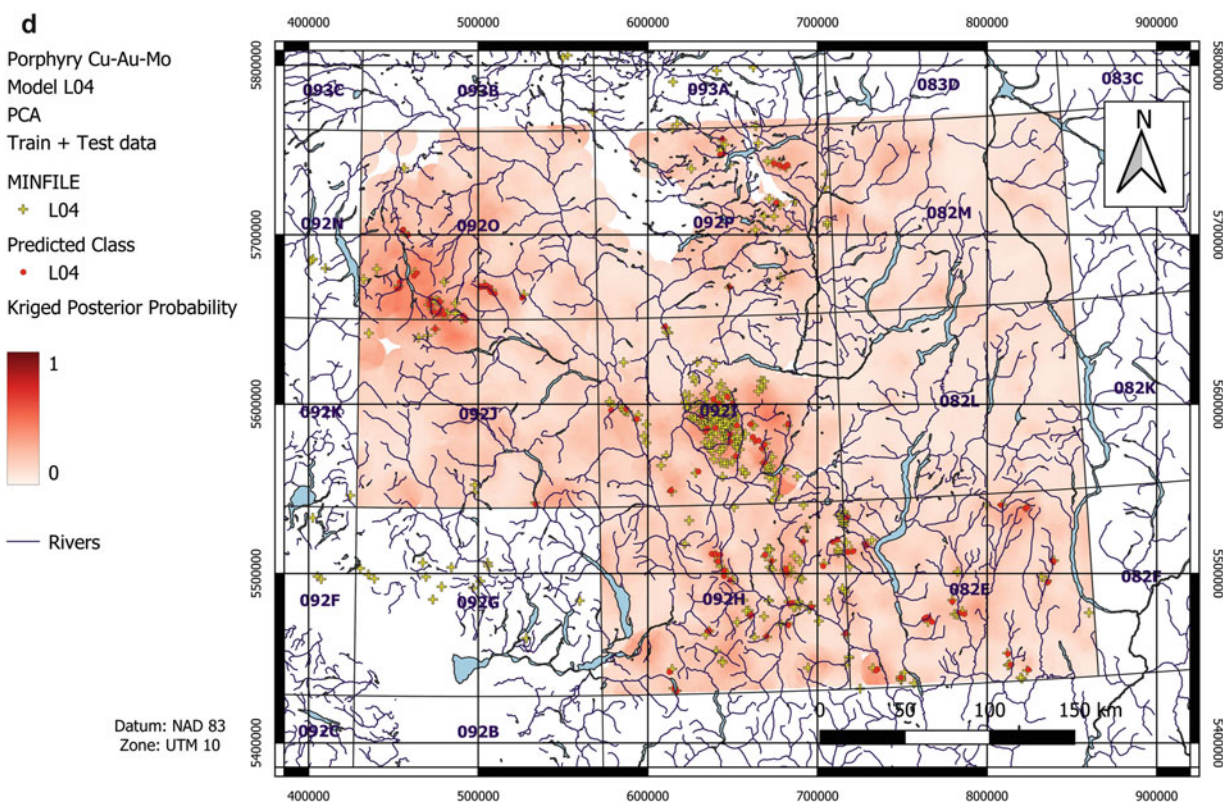
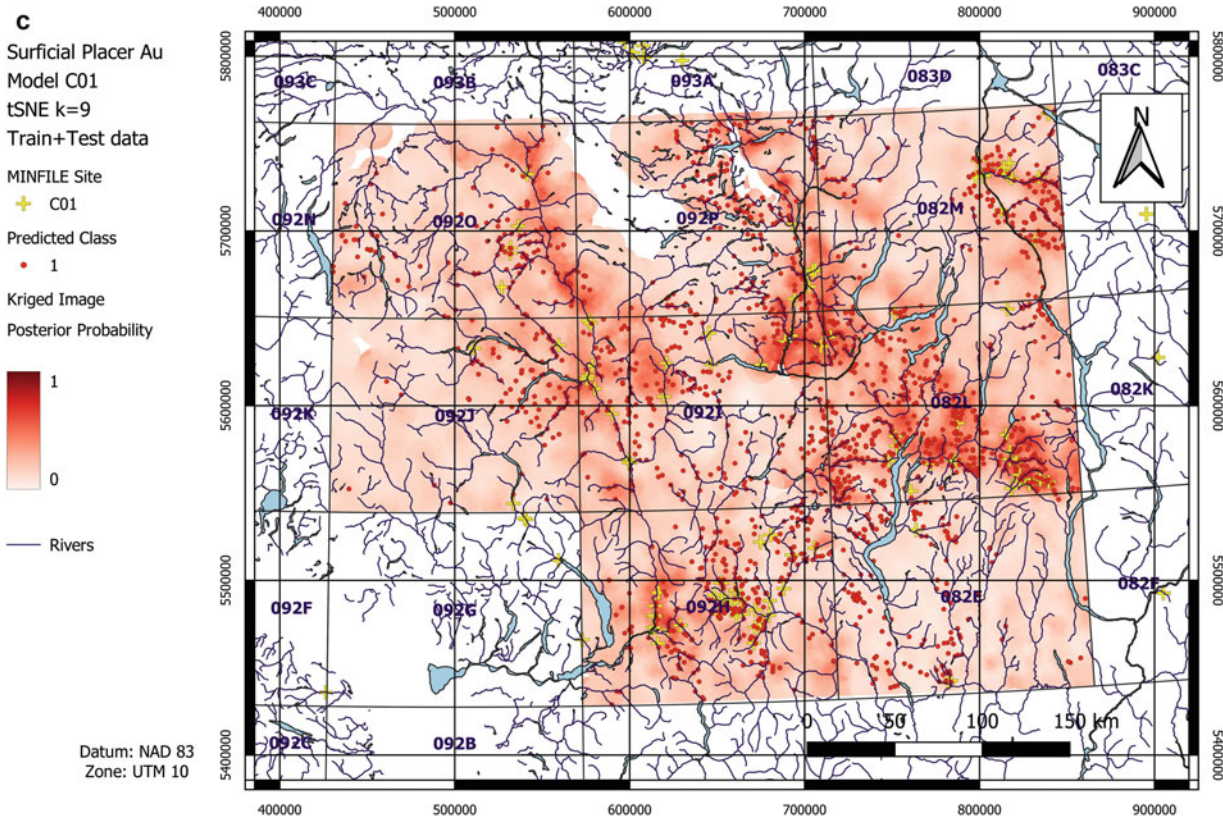
The fact that the corresponding MINFILE site and the stream sediment sample site are not colocated means that there is always the likelihood that the stream sediment composition does not reflect the observed mineralization at the MINFILE site. The distance threshold of 3.0 km appears to work for some mineral deposit types, but not necessarily for others. Changing the threshold distance for different mineral deposit models may help in a more refined estimate of mineral resource prediction.

Summary

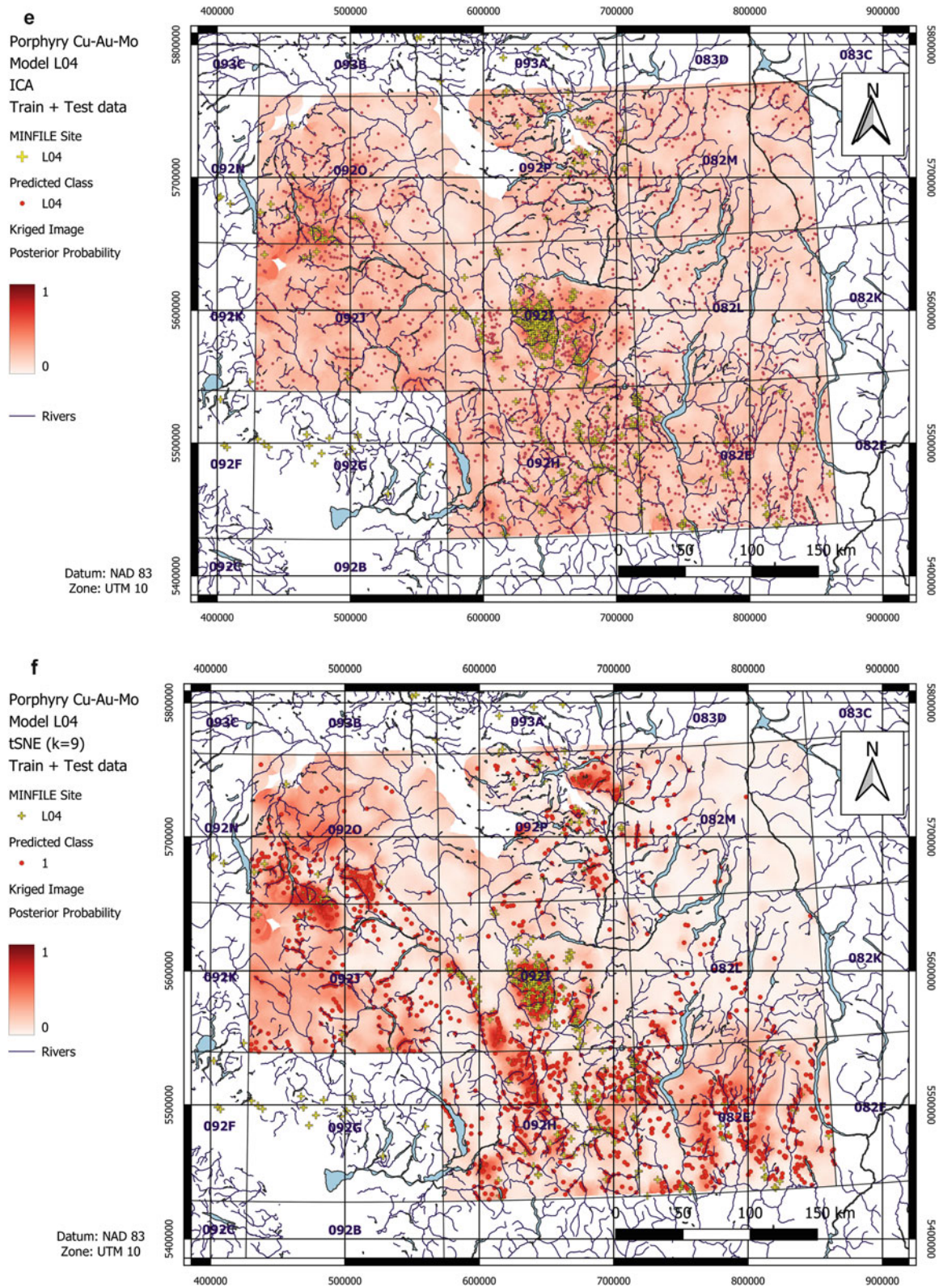
In the realm of computational geosciences, the integration and use of geoscience data as variables (e.g., geochemistry) and attributes (e.g., geology, mineral deposit occurrences) can be studied to identify/discover processes. In geochemical data, processes are recognized through the application of multivariate methods using different metrics (three in the example shown here) that demonstrate element associations related to mineralogy and patterns of geospatial coherence that are related to primary lithology and other processes including metamorphism, weathering, mass wasting, hydrothermal alteration, and base – precious-metal mineralization. These patterns, once identified as processes, can be used as training sets to validate and predict these processes using other data.



Computational Geoscience, Fig. 13 Kriged images of posterior probabilities based on random forests prediction for the three metrics (PCA, ICA, t-SNE) and two mineral deposit model types (C01-placer gold), (L04-porphyry Cu-Au). (a–c) C01 predictions. (d–f) L04 predictions



Computational Geoscience, Fig. 13 (continued)



Computational Geoscience, Fig. 13 (continued)

Cross-References

- Compositional Data
- Data Mining
- Data Visualization
- Discriminant Function Analysis
- Exploration Geochemistry
- Exploratory Data Analysis
- Fractal Geometry in Geosciences
- Geographical Information Science
- Geoscience Signal Extraction
- Geostatistics
- Imputation
- Interpolation
- Machine Learning
- Mineral Prospectivity Analysis
- Minimum Maximum Autocorrelation Factors
- Multidimensional Scaling
- Multivariate Analysis
- Multivariate Data Analysis in Geosciences, Tools
- Predictive Geologic Mapping and Mineral Exploration
- Random Forest
- Spatial Analysis
- Stationarity
- Variogram

Bibliography

- Aitchison J (1986) The statistical analysis of compositional data. Chapman and Hall, New York, p 416
- BC Geological Survey (1996) British Columbia mineral deposit profiles. BC Geological Survey. http://cmscontent.nrs.gov.bc.ca/geoscience/PublicationCatalogue/Miscellaneous/BCGS_MP-86.pdf. Accessed Nov 2019
- BC Geological Survey (2019) MINFILE BC mineral deposits database. BC Ministry of Energy, Mines and Petroleum Resources. <http://MINFILE.ca>. Accessed Nov 2019
- Cheng Q, Agterberg FP (1994) The separation of geochemical anomalies from background by fractal methods. *J Geochem Explor* 51:109–130
- Cheng Q, Zhao P (2011) Singularity theories and methods for characterizing mineralization processes and mapping geo-anomalies for mineral deposit prediction. *Geoscience Frontiers* 2(1):67–79
- Cheng Q, Xu Y, Grunsky EC (2000) Integrated spatial and spectrum analysis for geochemical anomaly separation. *Nat Resour Res* 9(1): 43–51
- Comon P (1994) Independent component analysis: a new concept? *Signal Process* 36:287–314
- Cox MAA (2001) Multidimensional scaling. Chapman and Hall, New York
- Egozcue JJ, Pawłowsky-Glahn V, Mateu-Figueras G, Barcelo-Vidal C (2003) Isometric logratio transformations for compositional data analysis. *Math Geol* 35(3):279–300
- Greenacre M, Grunsky E, Bacon-Shone J, Erb I, Quinn T (2022) Aitchison's Compositional Data Analysis 40 Years On: A Reappraisal. *arXiv:submit/4116835 [stat.ME]* 13 Jan 2022
- Grunsky EC (2010) The interpretation of geochemical survey data. *Geochem Explor Environ Anal* 10(1):27–74. <https://doi.org/10.1144/1467-7873/09-210>
- Grunsky EC (2012) Editorial, special issue on spatial multivariate methods. *Math Geosci* 44(4):379–380
- Grunsky EC, Arne D (2020) Mineral-resource prediction using advanced data analytics and machine learning of the QUEST-South stream-sediment geochemical data, southwestern British Columbia, Canada. *Geochem Explor Environ Anal*. <https://doi.org/10.1144/geochem2020-054>
- Grunsky EC, de Caritat P (2019) State-of-the-art analysis of geochemical data for mineral exploration. *Geochem Explor Environ Anal*. Special issue from Exploration 17, October, 2017, Toronto, Canada. <https://doi.org/10.1144/geochem2019-031>
- Hariharan S, Tirodkar S, Porwal A, Bhattacharya A, Joly A (2017) Random forest-based prospectivity modelling of greenfield terrains using sparse deposit data: an example from the Tanami Region, Western Australia. *Nat Resour Res* 26:489–507. <https://doi.org/10.1007/s11053-017-9335-6>
- Jolliffe IT (1986) Principal component analysis. Springer, Berlin. <https://doi.org/10.1007/b98835>
- McInnes L, Healy J, Melville J (2020) UMAP: Uniform Manifold Approximation and Projection for dimension reduction. *arXiv:1802.03426v3 [stat.ML]*, 18 Sep 2020
- McKinley JM, Hron K, Grunsky EC, Reimann C, de Caritat P, Filzmoser P, van den Boogaart KG, Tolosana-Delgado R (2016) The single component geochemical map: fact or fiction? *J Geochem Explor* 162:16–28. ISSN 0375-6742. <https://doi.org/10.1016/j.gexplo.2015.12.005>
- Mueller U, Tolosana-Delgado R, Grunsky EC, McKinley JM (2020) Biplots for compositional data derived from generalised joint diagonalization methods. *Appl Comput Geosci*. <https://doi.org/10.1016/j.acags.2020.100044>
- Palarea-Albaladejo J, Martín-Fernández JA, Buccianti A (2014) Compositional methods for estimating elemental concentrations below the limit of detection in practice using R. *J Geochem Explor* 141:71–77
- Pearce TH (1968) A contribution to the theory of variation diagrams. *Contrib Mineral Petrol* 19:142–157
- Pebesma EJ (2004) Multivariable geostatistics in S: the gstat package. *Comput Geosci* 30:683–691
- Switzer P, Green A (1984) Min/max autocorrelation factors for multivariate spatial imaging, Technical report no. 6. Department of statistics, Stanford University, Stanford
- Thompson M, Howarth RJ (1976a) Duplicate analysis in practice – part 1. Theoretical approach and estimation of analytical reproducibility. *Analyst* 101:690–698
- Thompson M, Howarth RJ (1976b) Duplicate analysis in practice – part 2. Examination of proposed methods and examples of its use. *Analyst* 101:699–709
- van der Maaten LJP, Hinton GE (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9:2579–2605

Computer-Aided or Computer-Assisted Instruction

Madhurima Panja and Uttam Kumar

Spatial Computing Laboratory, Center for Data Sciences,
International Institute of Information Technology Bangalore
(IIITB), Bangalore, Karnataka, India

Definition

“Computer-aided or computer-assisted instruction” (CAI) refers to instruction or remediation presented on a computer, i.e., it is an interactive technique whereby a computer is used to present the instructional material and monitor the learning that takes place. The computer has many purposes in the classroom, and it can be utilized to help a student in all areas of the curriculum. CAI refers to the use of the computer as a tool to facilitate and improve learning. Many educational computer programs are available online (through cloud services) and from computer stores and textbook companies; these tools enhance instructional qualities in several ways. Computer programs are interactive and enhance the learning process by utilizing a combination of text, graphics, vivid animation, sound, and demonstrations. Unlike a teacher-led program, computers offer a different range of activities and allow the students to progress at their own pace. Students are also provided the option of working either individually or in a group. A computer can train the student on the same topic rigorously until they master it. They also provide a facility for immediate feedback so that the students do not continue to practice the wrong skills. Computers capture the students’ attention and engage the students’ spirit of competitiveness to increase their scores.

Introduction

The earliest CAI, Pressey’s multiple-choice machine (Benjamin 1988), was invented in 1925. The main purpose of this machine was to present instructions, test the user, provide feedback to the user based on their responses, and record each attempt as data. Since then, several attempts have been made by researchers to improve Pressey’s multiple-choice machine and use it in various applications.

Computer-assisted or computer-aided instruction (CAI) has been a term of increasing significance during the last few decades and can also be referred to as computer-based instruction (CBI), computer-aided or computer-assisted learning (CAL). In general, CAI can be described as the set of instructions provided by a computer program that enhance the learning experience of users. However, the keyword for

understanding the CAI is *interaction*. Computers can facilitate interaction during the learning process in several ways, e.g., the user might interact with the learning material that has been programmed systematically. Alternatively, there might be interaction between user and the tutor, even the computer might host peer interaction or interaction between members of whole “virtual” learning communities. The concept of user interaction with content was initially introduced in the 1980s, and it forms the primary reason behind the rapid success of CAI. During the last few decades, the explosion of technological advancements allowed reliable and inexpensive communication that facilitated interaction between humans via computer programs, and massive open online course (MOOC) is a popular example of CAI.

Types of CAI Programs

Several types of software, viz., drill and practice, tutorial, simulations, problem-solving, gaming, and discovery are used for providing the course content in a CAI program (Blok et al. 2002). A brief description of these software is provided next:

- In drill-and-practice programs, the student rehearses different elements of teaching and develops related skills. This program is highly user-friendly and is one of the most prevalent types of computer software for many years. This type of software relies heavily on positive reinforcement, i.e., a reward follows a correct response. For example, if a drill and practice program was designed to help users to learn multiplication tables, then the drill might be presented with a car race game and the refueling of the car would depend on the correct answer provided by the student on the multiplication question. Drill-and-practice software deals primarily with lower-order thinking skills and also have a tracking facility that records the progress of the user. Moreover, this software does not utilize the full power of a computer and hence it can be used continuously to provide students with practice sessions and evaluate them.
- The use of “computers as tutors” is as widespread as that of drill-and-practice programs. The tutorial programs are specifically designed to teach certain concepts or skills and then assess the student’s understanding of such concepts by providing them the opportunity to practice. This type of program tries to act like a simulating teacher. The tutorial sessions are usually very interactive and involve explanations, questions, feedback, as well as correctives. Students are also provided with computerized tutorials that can be accessed multiple times until they become adept in those concepts.

- A simulation is a representation of real-world event, object or phenomenon. Simulation attracts the students' attention by setting up reality in classroom where the learners can see the results of their action. This feature of CAI provides great advantages for the science curriculum. With the help of simulation, experiments that are difficult, time-consuming, and costly could be realized in the classroom. Simulations also help in mathematics by producing visualizations of three-dimensional objects, which is not possible otherwise. This is a very powerful application of computers, and the educational community has the potential to capitalize on this type of software. Simulation software are now supported by artificial intelligence, and virtual and augmented reality.
- Problem-solving is a type of CAI program that allows the user to manipulate the variables, and feedback is provided based on these manipulations. The major difference between a problem-solving software and a simulation software is that the former does not necessarily utilize realistic scenarios like the latter one.
- Gaming-based CAI program is the most stimulating use of computers. These programs are a kind of simulation that offer competitive games from academic and nonacademic background to its users. The games are based upon various themes, e.g., some games might encourage their users to win individually, or some might give opportunity to the users to work as a team and explore the environment with other team members. However, irrespective of the theme based on which the game is designed, the primary objective behind these games is to produce valuable learning environment.
- Discovery is another approach used in a CAI program to engage its users for skill development. In this approach, the users are provided with a large database of information regarding a specific content area and are asked to analyze, compare, and evaluate their findings based on explorations of the database.

Features of CAI

CAIs have served the purpose of enhancing students' learning by affecting cognitive processes and increasing motivation. There are certain factors by which computer programs can facilitate this learning experience.

- Personalized information – CAI captures learner's interest in a given task by providing them with personalized information. For example, if a CAI program is allotted the task of explaining the concept of big data to the users, then it will use several animated objects to demonstrate the concepts and thereby increase learning by decreasing the

cognitive load on the learner's memory. This will allow the student to perform search and recognition processes and make more informational relationships (Fig. 1).

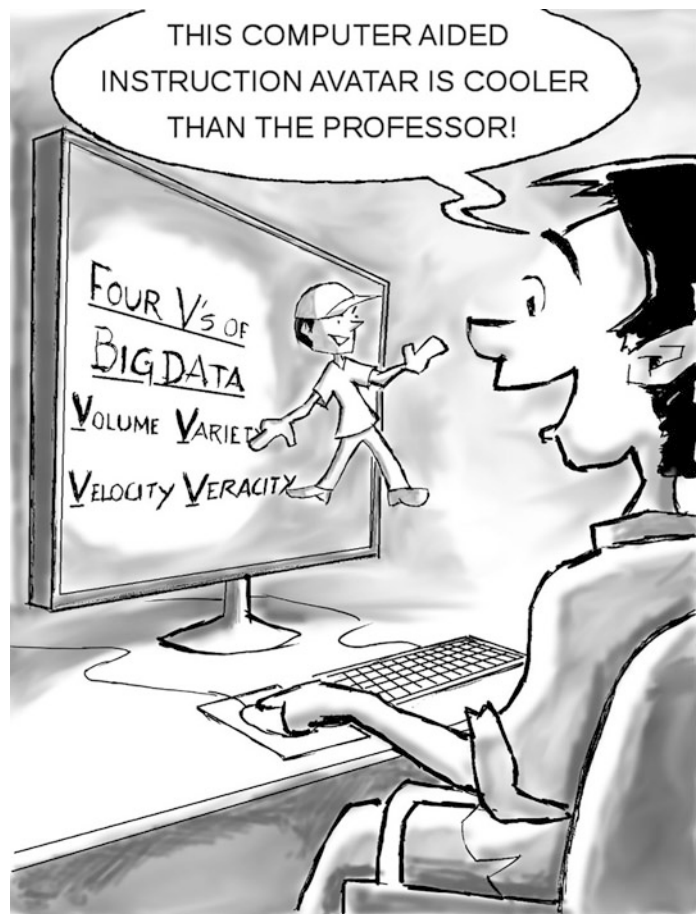
- Trial sessions – CAI encourages its learners to practice challenging problems. These practice sessions are intrinsically motivating and also carry other significant advantages such as personal satisfaction, challenge, relevance, and promotion of a positive perspective on a lifelong journey.

Types of Student Interactions in CAI

1. Recognition – In recognition type of interactions, the students are only required to identify whether or not a particular study material or question presented by a machine has been demonstrated previously.
2. Recall – This type of interaction requires the student to do more than just recognizing the information previously presented. The students are expected to reproduce the information precisely in a similar manner as was done previously by the machine.
3. Reconstruction – The reconstruction type of interaction does not engage the learners superficially like that of recognition and recall. In this type of interaction, the learners are required to reproduce concepts or principles that have been previously learnt. A basic difference between recall and reconstruction is that the latter does not show any relevant information except the question to the users when they are constructing the answers.
4. Intuitive understanding – The most difficult type of interaction between a CAI program and a student is intuitive understanding. These interactions often involve prolonged activity and are directed at “getting a feel” for an idea, developing sophisticated pattern-recognition skills, or developing a sense of strategy. Here, subject matter understanding must be demonstrated through experimental analysis and will be judged accordingly by experts. In this interaction, it is not possible to evaluate the students' performance by some explicit criteria stored in the computer. The activities that are used for assessing the intuitive understanding of learners include discovering principles behind simulations, problem-solving using classical techniques, and developing a feel for diagnostic strategies.
5. Constructive understanding – Unlike the other interactions mentioned above, constructive understanding is extremely open-ended and encourages the students to “create” knowledge. The learners are usually provided with an “open” enquiry and are expected to do genuine research instead of just providing solutions to questions on the content that is already known. From the users' point of view, she/he will be going beyond her/his knowledge and

Computer-Aided or Computer-Assisted Instruction,

Fig. 1 Demonstration of a learner using a CAI program © [Sujit Kumar Chakrabarti.]



may be testing her/his own hypothesis, developing new theory, and drawing conclusions based on her/his own work.

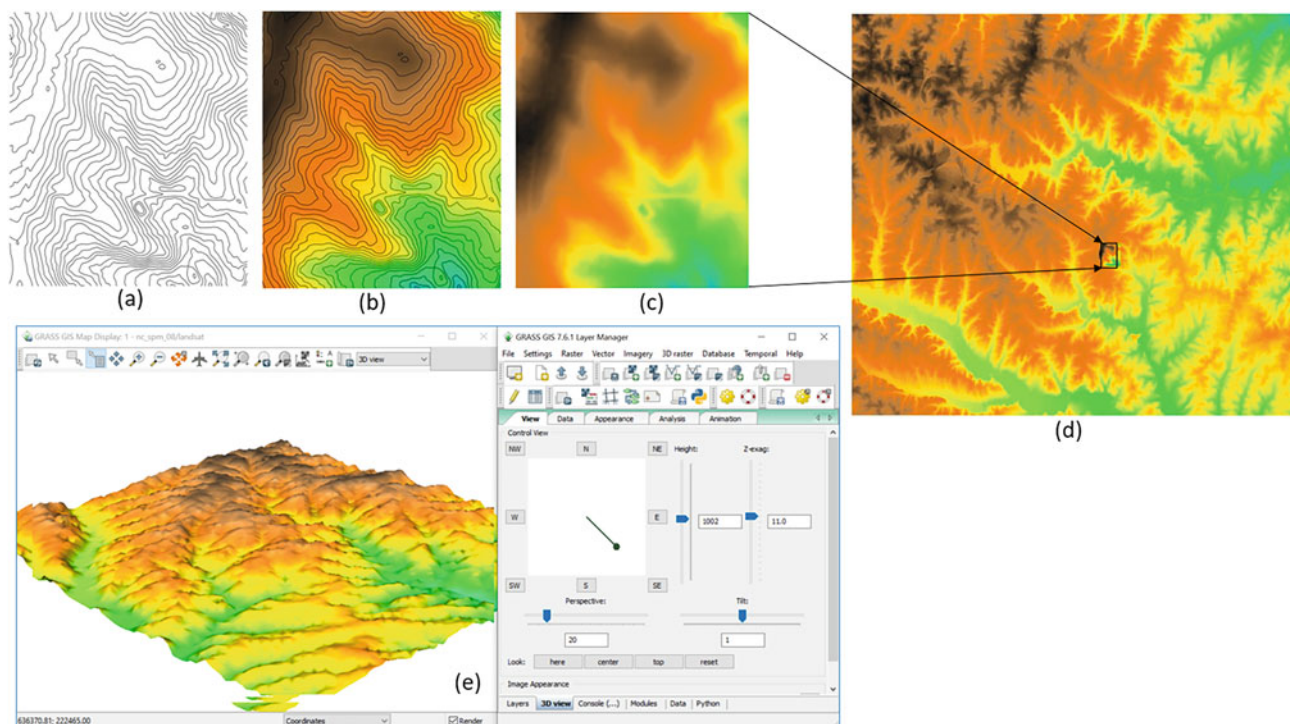
How Is CAI Implemented?

Teachers are primarily responsible to review the computer programs or games or online activities before they are introduced to the students. The mentors should understand the context of the lessons in any CAI program and then determine their usefulness. The CAI program must satisfy some basic requirements. Firstly, the program should be able to supplement the lesson and give basic skills to the learners. Secondly, the content should be presented in such a manner that the student will remain interested in the topic. However, it must be ensured that the students don't waste a substantial amount of time with too much animation in a content. Finally, the program must be at the correct level for the class or the individual students, and should neither be too simple nor too complex. Once these requirements are fulfilled, the teachers might allow the students to operate these computer programs to enhance their learning experience.

Advantages of CAI

There are several advantages of using a CAI program as listed below:

1. CAI programs are self-pacing; they allow the students to learn at their own pace, totally unaffected by the performance of others. Owing to its self-directed learning nature, the learners have the freedom to choose when, where, and what to learn. It is also very useful for slow learners.
2. Information is presented in a structured manner in CAI software, which is useful when there exists a hierarchy of facts and rules on the subject.
3. CAI requires active participation from its users, which is not necessarily true while reading a book or attending a lecture.
4. The immediate feedback mechanism of CAI software provides a clear picture to the student about their progress. Learners can identify the subject areas in which they have improved and in which they need to improve further.
5. While teaching a concept, the CAI programs enable students to experiment with different options and explore the results after any manipulations, thereby reducing the time taken to comprehend difficult concepts.



Computer-Aided or Computer-Assisted Instruction, Fig. 2 Three-dimensional surface generation in GRASS GIS with a simulation tool

6. CAI offers a wide range of experience that is otherwise not available to the student. It enables the students to understand concepts clearly using various techniques such as animation, blinking, and graphical displays. CAI software acts as multimedia that helps to understand difficult concepts through a multisensory approach.
7. CAI provides individual attention to each and every student. Additionally, a lot of practice sessions are also given to the students that can prove useful, especially for low-aptitude ones.
8. CAI programs are great motivators and encourage the students to take part in challenging activities, thus, enhancing their reasoning and decision-making abilities.

Limitations of CAI

Apart from the abovementioned advantages, there are certain restrictions of CAI software as described next:

1. Although CAI programs allow experimentation through simulation, the hands-on experience is missing. These programs cannot develop manual skills like handling apparatus in a chemical experiment, working with machines in a physical laboratory, workshop experience, etc.
2. There are real costs associated with the development of CAI systems. They require dedicated time from

programmers to design such a software as well as sophisticated infrastructure for deployment.

3. The course content of any CAI program needs to be updated frequently, otherwise, it will become obsolete.
4. Since the objective and method of teaching decided by a teacher may be different from that of a CAI program designer, the program might fail to fulfill the expectation of the teacher.
5. Teachers might be reluctant about using a computer in teaching and may be unwilling to spend extra time on the preparation, selection, and use of CAI packages. It may be perceived as a threat to their job.
6. The overuse of multimedia in CAI software may divert the attention of the student from the course.
7. CAI packages depend upon the configuration of the hardware; previously built packages might become completely useless after major hardware upgradation.

Application of CAI in Geospatial Analysis

In geospatial science, CAI programs are very useful in the simulation of real-world scenarios. There are many free and open-source software that provide such simulation and visualization capabilities. Geographic Resources Analysis Support System (GRASS) GIS (<https://grass.osgeo.org/>) is a software built for vector and raster geospatial data

management, geoprocessing, spatial modeling, and visualization released under the GNU General Public License. It has programs to render three-dimensional views from contours at a given spatial location that allow users to generate interpolated surface of the terrain and simulate the environment. Figure 2a shows contours and Fig. 2b, c shows interpolated raster (digital elevation model) of a small region from a larger terrain shown in Fig. 2d. Three-dimensional surface generation in GRASS GIS with a simulation tool is shown in Fig. 2e. The dataset used to generate the figures are from the North Carolina sample open data available for GRASS GIS users (<https://grass.osgeo.org/download/data/>).

These types of simulations can be realized in a classroom environment without actually visiting the spatial point of interest. A fly-through can be generated and stored in the form of popular video formats for multimedia display.

Conclusion

This chapter was a brief overview of the features and usage of CAI. Nowadays, with the rapid technological development, it is almost impossible to keep away from electronic devices like laptops and smartphones. However, with the help of CAI software, students can be motivated to use electronic devices for their greater good. CAI provides the learners with the freedom to experiment with different options and get instant feedback. Another crucial feature of CAI is its one-to-one interaction, which not only provides individual attention to the students but also allows them to learn at their own pace. This assistance from technology provides opportunities to the teachers to work with students who need more of their time. It is also a proven concept that the animations that a CAI package offers enable the students to learn difficult concepts quite easily and are useful in virtual laboratories. With all these great features that CAI has to offer to its users, it is always recommended to use this technology under proper supervision, otherwise, the learning becomes too mechanical. If all these factors are taken into consideration, one might appropriately declare that CAI has the potential to transform the educational system of our society (Bayraktar 2001).

Cross-References

- Artificial Intelligence in the Earth Sciences
- Big Data
- Cloud Computing and Cloud Service
- Pattern Recognition

Bibliography

- Bayraktar S (2001) A meta-analysis of the effectiveness of computer-assisted instruction in science education. *J Res Technol Educ* 34(2): 173–188
- Benjamin L (1988) A history of teaching machines. *Am Psychol* 43(9):703
- Blok H, Oostdam R, Otter M, Overmaat M (2002) Computer-assisted instruction in support of beginning reading instruction: a review. *Rev Educ Res* 72(1):101–130

Concentration-Area Plot

Behnam Sadeghi

EarthByte Group, School of Geosciences, University of Sydney, Sydney, NSW, Australia

Earth and Sustainability Science Research Centre, University of New South Wales, Sydney, NSW, Australia

Definition

Element contour maps are representative tools to discriminate the anomalous areas from background areas using optimum thresholds. Such thresholds could be defined via the target element concentration-area (C-A) log-log plot. Such an idea was developed as the C-A fractal model, by Cheng et al. (1994), which is the first fractal/multifractal model applied in geochemical anomaly classification.

Introduction

Classification models, represented as classified maps, are the main tools to discriminate anomalous areas of mineralization from background areas. To expose the subtle patterns (e.g., geochemical patterns) associated with various mineralization and lithologies, a variety of mathematical and statistical models have been developed and applied to the available data. Traditional statistical methods, advanced geostatistical models and simulations, fractal/multifractal models, machine learning, deep learning, and artificial intelligence are some of the models that have been developed (Sadeghi 2020). Fractal geometry was proposed by Mandelbrot (1983); later, in 1992, it was taken into further consideration in geosciences by Bölviken et al. – see the ► “Fractal Geometry in Geosciences” and ► “Singularity Analysis” entries for further details. Then, Cheng et al. (1994) developed the first geochemical anomaly classification fractal model, called concentration-area (C-A). The C-A model is simply based on the relation between the target element concentration and the area confined by the relevant concentration contours (Cheng et al. 1994; Zuo and

Wang 2016; Sadeghi 2020, 2021; Sadeghi and Cohen 2021). In this entry, the C-A fractal model and its log-log plot will be studied in detail. After that, several other fractal models, particularly concentration-volume (C-V) (Afzal et al. 2011) and concentration-concentration (C-C) (Sadeghi 2021) fractal models, which were developed based on the C-A model, will be discussed.

Methodology: Concentration-Area Fractal Model

The C-A fractal model, as one of the important fractal models, describes the spatial distribution of geochemical data based on the spatial relationship of geochemical concentrations to the occupied areas (Cheng et al. 1994):

$$A(\rho \leq v) \propto \rho^{-a_1}; A(\rho > v) \propto \rho^{-a_2}$$

where $A(\rho)$ denotes the area occupied by concentrations greater than a concentration contour value ρ ; v represents a threshold concentration contour; and a_1 and a_2 are characteristic exponents. The thresholds defined by the C-A fractal model are cross-points of lines fitted, by Least Squares (LS) method, through log-log plots of A versus ρ .

One question that may come to mind is about the relation between the fractal geometry, log-log plots, e.g., log-log plots of A versus ρ , fitting straight lines and calculating their slopes (fractal dimensions). Such slopes represent the exponents of the power-law relation in the C-A model equation (Zuo and Wang 2016). The answer, in summary, is that if we implement the logarithm (e.g., log 10) on both sides of the C-A fractal model equation, the result will be the equation of a straight line as below:

$$Y = mX + C$$

where X and Y are the axis coordinates of the points generating the lines, m denotes the slope of the lines, and C is a constant.

Therefore, analysts need to fit straight lines to the fractal log-log plots to find their cross-points as the desired thresholds. The main method implemented to do so is the LS fitting method (Carranza 2009), as will continue to be discussed.

The LS method, proposed by Carl Friedrich Gauss (1795) (Stigler 1981; Bretscher 1995), is a regression analysis to do linear and nonlinear regression. It minimizes the sum of the difference between the actual values and the fitted values predicted by a model, which are the squared residuals (r^2).

The data pairs are (x_i, y_i) that $i = 1, \dots, n$, where x_i is an independent variable and y_i is a dependent variable. If the model function is $f(x, \beta)$, the i th residual would be calculated as given below (cf. Stigler 1981; Bretscher 1995; Strutz 2011):

$$r_i = y_i - f(x_i, \beta) = y_i - \sum_{j=1}^n X_{ij} \beta_j$$

So the sum of the squared residuals is:

$$S = \sum_{i=1}^n r_i^2$$

If β_0 is the y intercept, and β_1 is the slope of the straight line, the model function is given by:

$$f(x, \beta) = \beta_0 + \beta_1 x$$

If the sum of the squared residuals, S , is minimum, the LS method would be optimum. So, in order to have the $S_{\min}$, its gradient should be set to 0, meaning:

$$\frac{\partial S}{\partial \beta_j} = 2 \sum_{i=1}^m r_i \frac{\partial r_i}{\partial \beta_j} = 0, (j = 1, \dots, n)$$

Then, the derivatives are:

$$\frac{\partial r_i}{\partial \beta_j} = -X_{ij}$$

so,

$$\frac{\partial S}{\partial \beta_j} = 2 \sum_{i=1}^m \left(y_i - \sum_{k=1}^n X_{ik} \beta_k \right) (-X_{ij}), (j = 1, \dots, n)$$

If $\hat{\beta}$ minimizes S , we will have:

$$2 \sum_{i=1}^m \left(y_i - \sum_{k=1}^n X_{ik} \hat{\beta}_k \right) (-X_{ij}), (j = 1, \dots, n)$$

and now, the normal equations are obtained:

$$\sum_{i=1}^m \sum_{k=1}^n X_{ij} X_{ik} \hat{\beta}_k = \sum_{i=1}^m X_{ij} y_i, (j = 1, 2, \dots, n)$$

The matrix notation of this equation is:

$$(X^T X) \hat{\beta} = X^T y$$

where X^T is the matrix transpose of X .

Another significant parameter in the straight-line fitting using the LS technique that should be taken into account is the determination coefficient (R^2), which is the square of the correlation coefficient (R) between predicted values and actual values; so, the range of R^2 is between 0 and 1. It is considered as a key output of regression analysis. In other words, it is the proportion of the variance in the dependent variable, which is predictable from the independent variable. The reason why the coefficient is important is because a linear

relation is supposed to be strong/weak if the points on a scatterplot are close to/far from the straight line. Therefore, scientifically, instead of simply saying there is a strong or weak relationship between the variables, analysts must apply a measure of how strong/weak the association is.

If $R^2 = 0$, it means that the dependent variable cannot be predicted from the independent variable; if $R^2 = 1$, it explains the dependent variable can be predicted from the independent variable, without error; and if R^2 is between 0 and 1, it denotes the extent to which the dependent variable is predictable. For instance, if $R^2 = 0.10$, it means that 10% of the variance in Y is predictable from X .

The formula of R considering R^2 for a linear regression model with one independent variable is (Caers 2011):

$$R = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \cdot \left(\frac{y_i - \bar{y}}{s_y} \right)$$

where n is the number of samples or points on the scatter plot; x_i and y_i are the x and y values of the i th sample, $\bar{x}$ and $\bar{y}$ are the

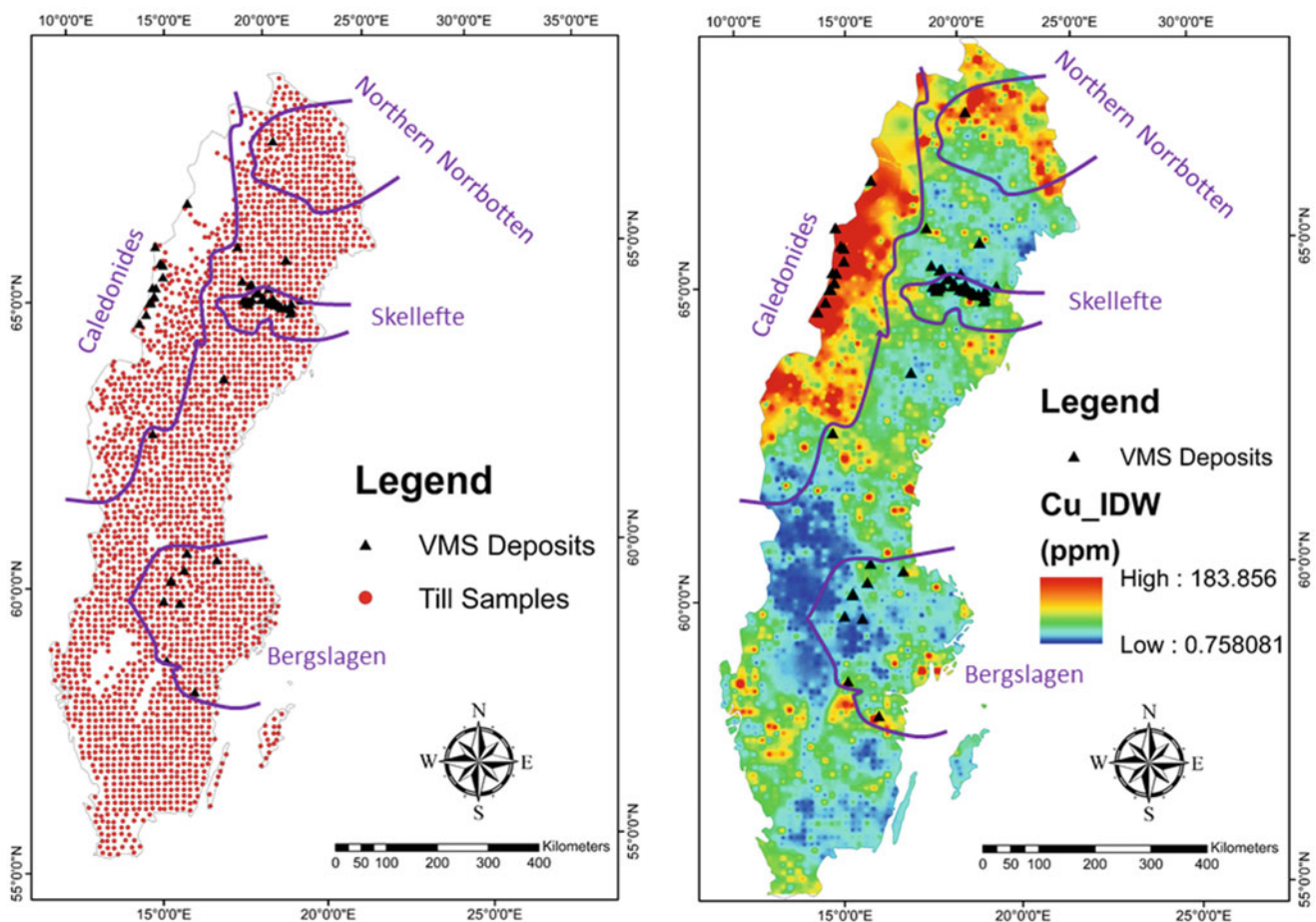
mean of the x and y values; and s_x and s_y are the standard deviations of x and y .

Case Study: Sweden

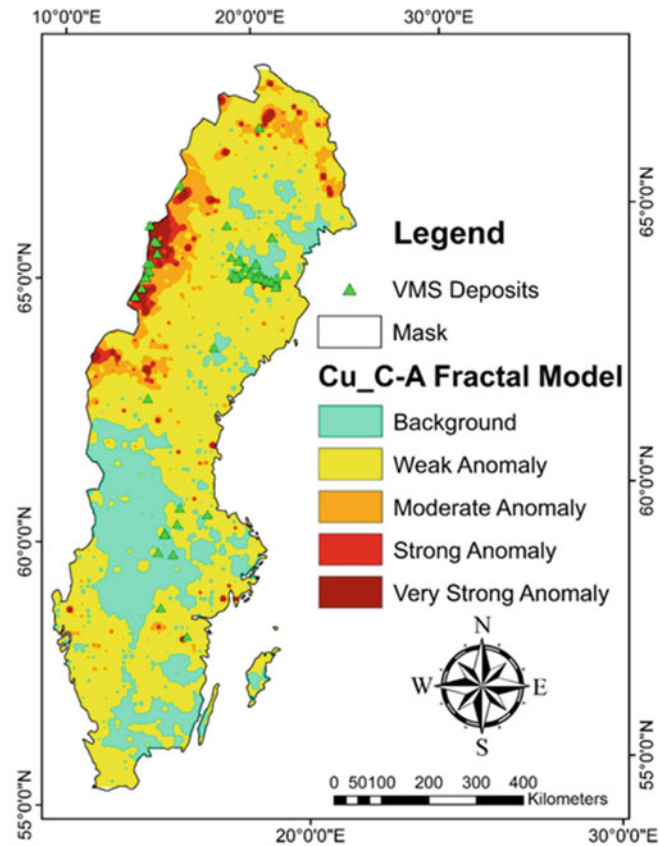
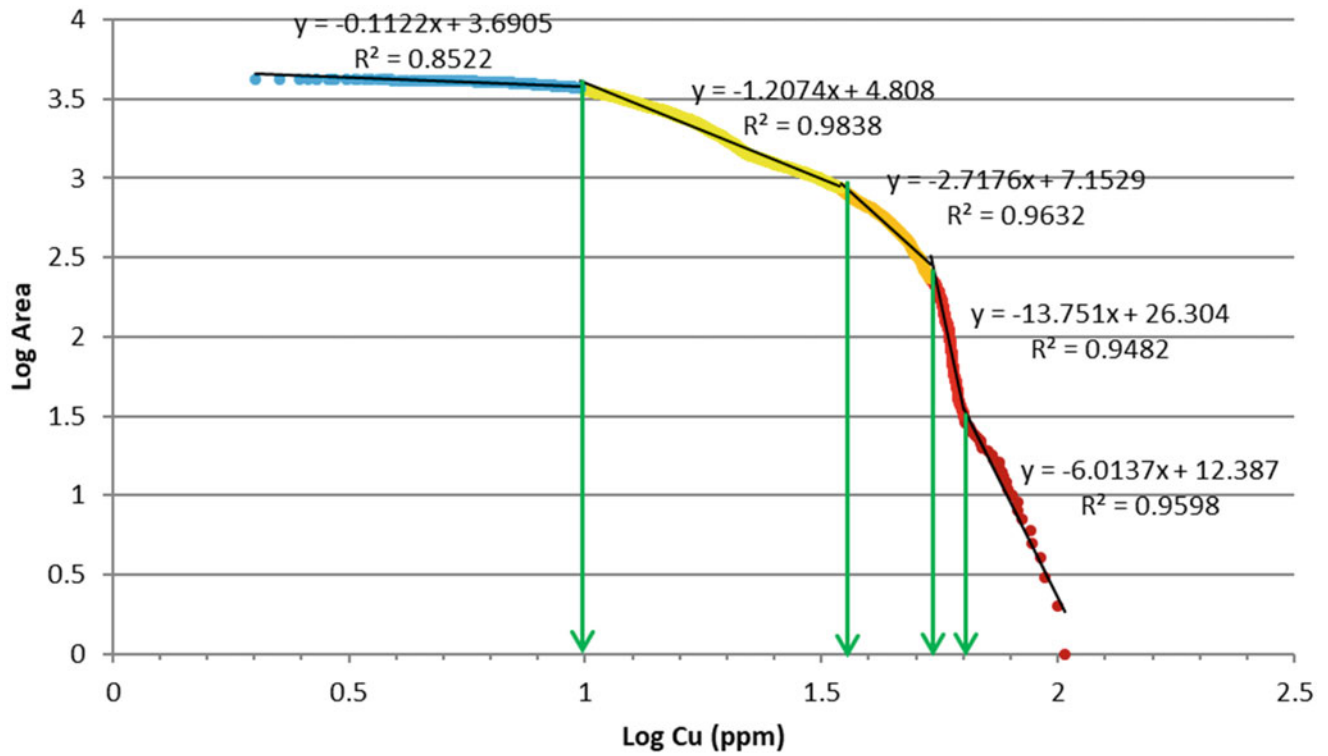
Sadeghi and Cohen (2019) applied the C-A model to Swedish till geochemical data, collected by Geological Survey of Sweden (Andersson et al. 2014) (Fig. 1). The thresholds obtained were applied to generate the classified volcanic massive sulfide (VMS) Cu geochemical anomaly map (Fig. 2). Main anomalies are concentrated in Caledonides and Northern Norrbotten metallogenic provinces.

Summary of the Other Models Developed Based on the C-A Fractal Model

Several other models have been developed based on the C-A model such as the 3D format of C-A, called concentration-



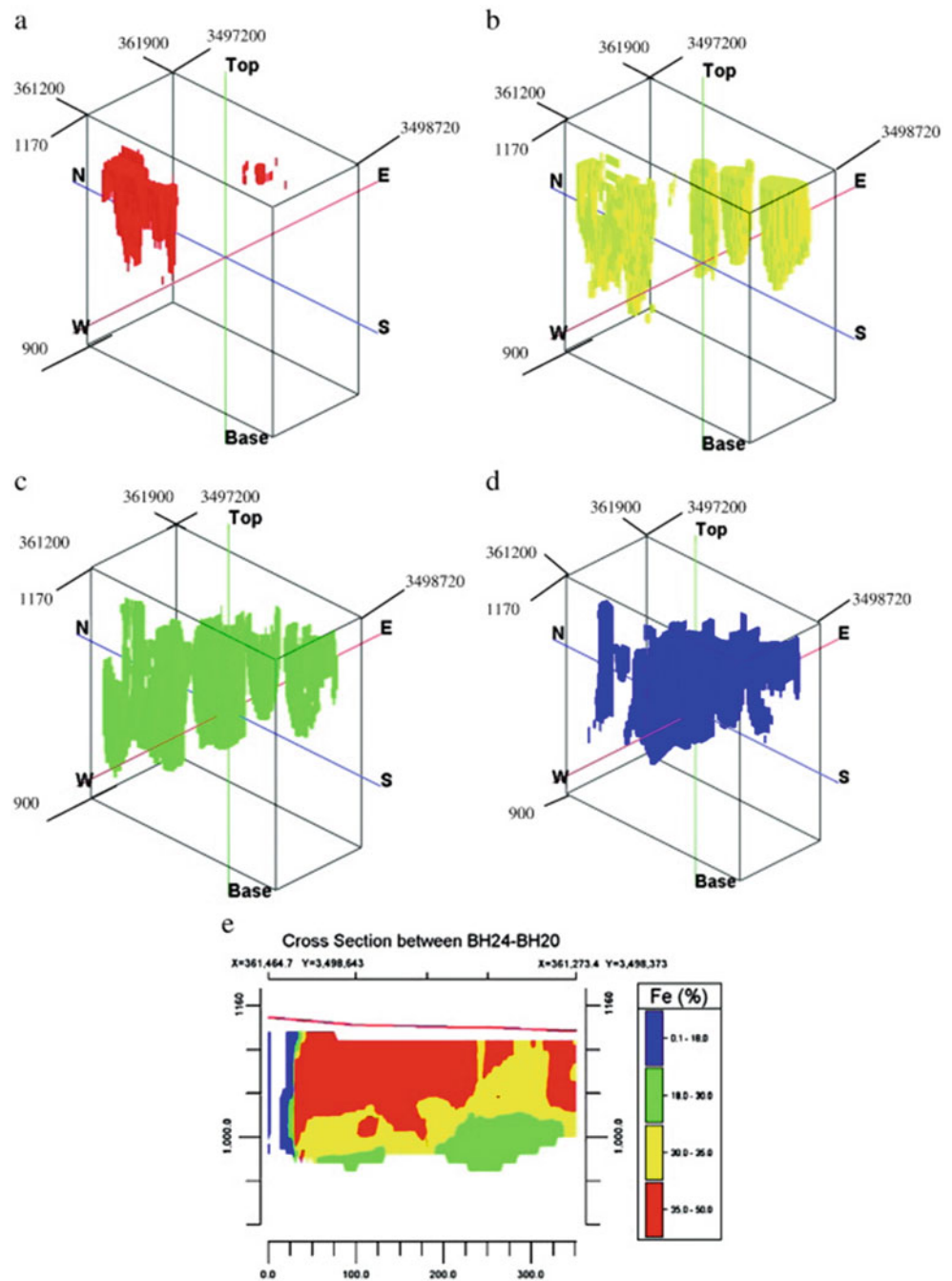
Concentration-Area Plot, Fig. 1 Swedish till sample locations, and the IDW interpolated map of Cu. (From Sadeghi 2020; Sadeghi and Cohen 2021)



Concentration-Area Plot, Fig. 2 VMS Cu geochemical anomalies recognized by C-A fractal modeling of till data in Sweden. (From Sadeghi and Cohen 2019)

Concentration-Area Plot,

Fig. 3 (a) Highly, (b) moderately, and (c) weakly mineralized zones, in addition to (d) wall rocks, and (e) the cross-section of the mineralized zones, characterized by C-V fractal model. (From Sadeghi et al. 2012)



volume (C-V) (Afzal et al. 2011), spectrum-area (S-A) (Cheng et al. 2000), and its 3D format, called power spectrum-volume (P-V) (Afzal et al. 2012), and concentration-concentration (C-C) (Sadeghi 2021) fractal models. For S-A and P-V fractal models see the ► “Spectrum-Area Method” entry. Here C-V and C-C fractal models are introduced.

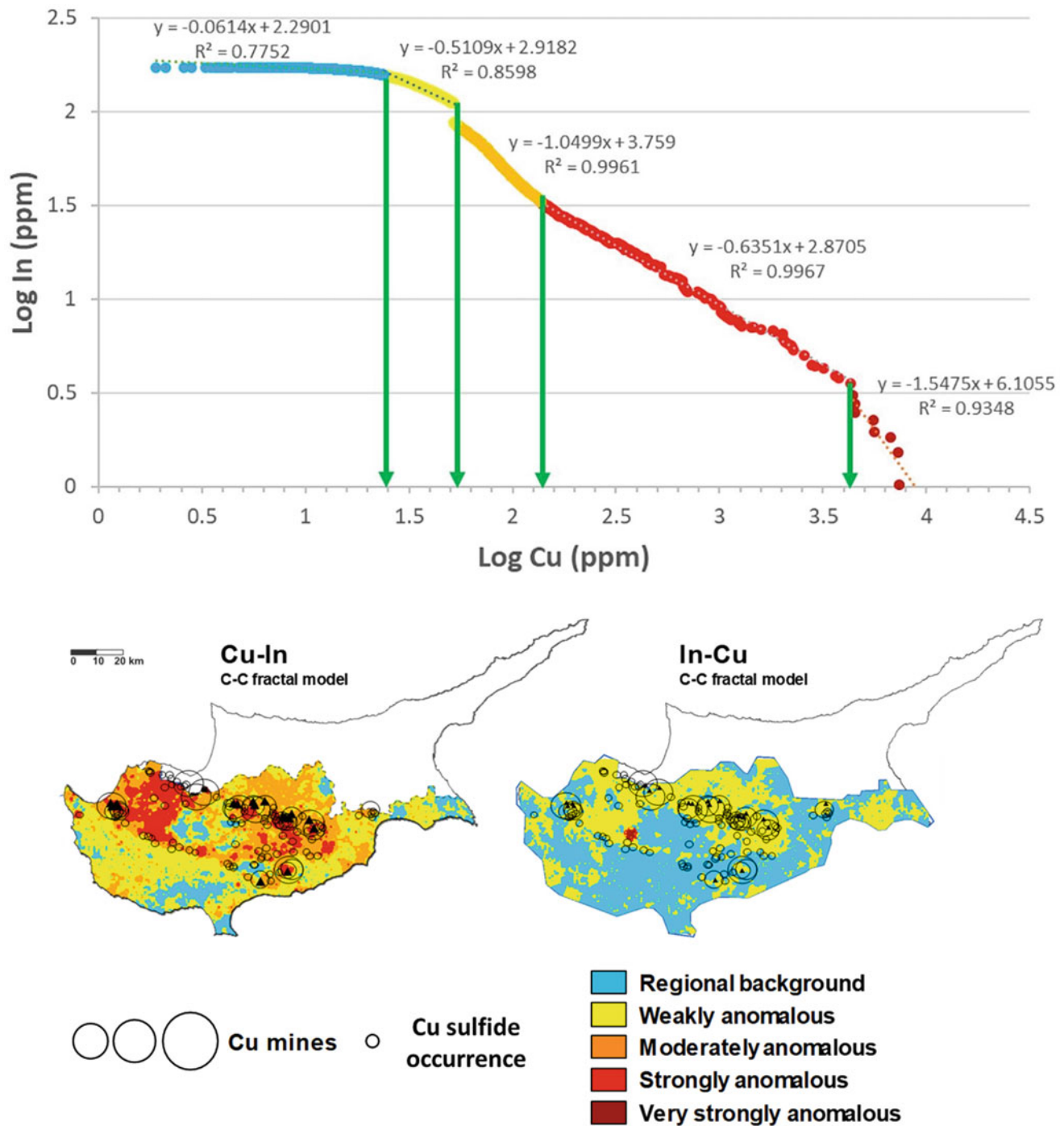
The C-V fractal model was developed by Afzal et al. (2011) to define different mineralization zones in 3D:

$$V(\rho \leq v) \propto \rho^{-a_1}; V(\rho > v) \propto \rho^{-a_2}$$

where $V(\rho)$ denotes the volume occupied by concentrations greater than ρ .

In Fig. 3, Sadeghi et al. (2012) have applied the C-V model to the Fe data of Zaghia iron ore deposit, Central Iran, to characterize different mineralized zones.

Sadeghi (2020, 2021) proposed the C-C fractal model as a bivariate modification of the C-A model (Fig. 4):



Concentration-Area Plot, Fig. 4 The C-C model applied to Cyprus soil data to classify Cu and In geochemical anomalies based on each other. (From Sadeghi 2021)

$$C_2 (\geq C_1) = FC_1^{-D}$$

where C_1 is the target element concentration, $C_2(\geq C_1)$ is the cumulative concentration of the element, which is highly correlated with the target element, with concentration values

higher than or equal to C_1 ; F is a constant and D is the fractal dimension. Unlike the C-A model, in the C-C model the element C_2 is not measured in the area, i.e., it is measured in a subspace containing subareas in which C_1 values exceed a certain value.

Summary and Conclusions

Fractal/multifractal models provide robust tools to characterize mineralization zones and discriminate geochemical anomalies from the background using the obtained thresholds. The C-A fractal model is the first fractal model, which has been efficiently applied to geochemical anomaly classifications. It works based on the relation between element concentrations and their occupied contour areas. It has been applied to various geological studies, then modified to several other models such as C-V in 3D, S-A and its 3D format, P-V, and recently C-C fractal models.

Cross-References

- [Fractal Geometry in Geosciences](#)
- [Singularity Analysis](#)
- [Spectrum-Area Method](#)

Bibliography

- Afzal P, Fadakar Alghalandis Y, Khakzad A, Moarefvand P, Rashidnejad Omran N (2011) Delineation of mineralization zones in porphyry Cu deposits by fractal concentration-volume modeling. *J Geochem Explor* 108:220–232
- Afzal P, Fadakar Alghalandis Y, Moarefvand P, Rashidnejad Omran N, Asadi Haroni H (2012) Application of power-spectrum-volume fractal method for detecting hypogene, supergene enrichment, leached and barren zones in Kahang Cu porphyry deposit, Central Iran. *J Geochem Explor* 112:131–138
- Andersson M, Carlsson M, Ladenberger A, Morris G, Sadeghi M, Uhlbäck J (2014) Geochemical atlas of Sweden. Geological Survey of Sweden (SGU), Uppsala, p 208
- Bölviken B, Stokke PR, Feder J, Jössang T (1992) The fractal nature of geochemical landscapes. *J Geochem Explor* 43:91–109
- Bretscher O (1995) Linear algebra with applications, 3rd edn. Prentice Hall, Upper Saddle River
- Caers JK (2011) Modeling uncertainty in earth sciences. Wiley, Hoboken
- Carranza EJM (2009) Geochemical anomaly and mineral prospectivity mapping in GIS. In: Handbook of exploration and environmental geochemistry, vol 11. Elsevier, Amsterdam, p 368
- Cheng Q, Agterberg FP, Ballantyne SB (1994) The separation of geochemical anomalies from background by fractal methods. *J Geochem Explor* 51:109–130
- Cheng Q, Xu Y, Grunsky E (2000) Integrated spatial and spectrum method for geochemical anomaly separation. *Nat Resour Res* 9:43–52
- Mandelbrot BB (1983) The fractal geometry of nature, 2nd edn. Freeman, San Francisco
- Sadeghi B (2020) Quantification of uncertainty in geochemical anomalies in mineral exploration. PhD thesis, University of New South Wales
- Sadeghi B (2021) Concentration–concentration fractal modelling: a novel insight for correlation between variables in response to changes in the underlying controlling geological–geochemical processes. *Ore Geol Rev*. <https://doi.org/10.1016/j.oregeorev.2020.103875>
- Sadeghi B, Cohen D (2019) Selecting the most robust geochemical classification model using the balance between the geostatistical precision and sensitivity. In: International Association for Mathematical Geology (IAMG) conference, State College
- Sadeghi B, Cohen D (2021) Category-based fractal modelling: a novel model to integrate the geology into the data for more effective processing and interpretation. *J Geochem Explor*. <https://doi.org/10.1016/j.gexplo.2021.106783>
- Sadeghi B, Moarefvand P, Afzal P, Yasrebi AB, Daneshvar Saein L (2012) Application of fractal models to outline mineralized zones in the Zaghia iron ore deposit, Central Iran. *J Geochem Explor* 122: 9–19
- Stigler SM (1981) Gauss and the invention of least squares. *Ann Stat* 9(3):465–474
- Strutz T (2011) Data fitting and uncertainty. Vieweg+Teubner Verlag/ Springer, Berlin
- Sun H (1992) A general review of volcanogenic massive sulfide deposits in China. *Ore Geol Rev* 7:43–71
- Zuo R, Wang J (2016) Fractal/multifractal modeling of geochemical data: a review. *J Geochem Explor* 164:33–41

Constrained Optimization

Deeksha Aggarwal and Uttam Kumar
Spatial Computing Laboratory, Center for Data Sciences,
International Institute of Information Technology Bangalore
(IIITB), Bangalore, Karnataka, India

Definition

Constrained optimization is defined as the process of optimizing an objective function with respect to some variables on which constraints are defined. In general, for optimization, the objective function is either minimized in case of a cost or energy function or maximized in case of a reward or utility function. The constraints defined on the variables can be hard constraints which put a hard limit on the value of the variable or it can be a soft constraint having some penalized values of the variables in the objective function if the condition on the variables is not satisfied.

Introduction

Optimization problem can be easily seen in most of the real world economic decisions which are generally subject to one or more constraint(s). Some real world examples of the economic decisions are:

1. When buying goods or services, consumers make decisions on what to buy, constrained by the fact that it must be affordable and under their budget.
2. Industries often make decisions on maximizing their profit subject to having limited production capacity, limited

costing, and also maintaining high quality of goods or services that are above standards.

Basic Nomenclature

A set $S \subset R^n$ is called:

- Affine if $x, y \in S$ implies $\theta x + (1 - \theta)y \in S$ for all $\theta \in R$.
- Convex if $x, y \in S$ implies $\theta x + (1 - \theta)y \in S$ for all $\theta \in [0, 1]$.
- Convex cone if $x, y \in S$ implies $\lambda x + \mu y \in S$ for all $\lambda, \mu \geq 0$.

A function $f: R^n \rightarrow R$ is called

- Affine if $f(\theta x + (1 - \theta)y) = \theta f(x) + (1 - \theta)f(y)$ for all $\theta \in R$ and $x, y \in R^n$.
- Convex if $f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y)$ for all $\theta \in [0, 1]$ and $x, y \in R^n$.

General form: A constrained optimization problem has the following general form:

$$\min f(x)$$

such that

$$g_i(x) = c_i; \quad \text{for } i = 1, \dots, n; c \in \mathfrak{R} \text{ Equality constraint}$$

$$h_j(x) \geq d_j; \quad \text{for } j = 1, \dots, m; d \in \mathfrak{R} \text{ Inequality constraint}$$

Here, $f(x)$ is the objective function that needs to be optimized in terms of minimization such that g_i and h_j (hard constraints) are satisfied.

Convexity: This problem is called a convex problem if the objective function f is convex, the inequality constraints $h_j(x)$ are convex, and the equality constraints $g_i(x)$ are affine. Convex analysis is the study of properties of convex functions and convex sets with applications in convex minimization.

Solution methods: To solve the constrained optimization problem, several methods have been proposed. Some of these methods are described below:

- (i) **Substitution method:** For simple objective functions having two or three variables with equality constraints, this method is favorable as the solution. Here, a composite function is made by substituting the value of the equality constraint in the objective function as shown below:

$$\begin{aligned} \text{Let } \min f(x) &= x \cdot y \\ \text{such that } y &= x - 11 \end{aligned}$$

Hence, substituting y in $f(x)$ we get:

$$\begin{aligned} f(x) &= x(x - 11) \\ f(x) &= x^2 - 11x \end{aligned}$$

According to first-order necessary condition, $\frac{\partial f}{\partial x} = 0$. Therefore,

$$\begin{aligned} \frac{\partial f}{\partial x} &= 2x - 11 \\ 2x - 11 &= 0 \\ x &= 5.5 \end{aligned}$$

- (ii) **Lagrange multiplier:** This method is used to minimize or maximize the objective function subject to one or more equality constraints only. To solve the objective function, a Lagrangian function is defined which is a relation between the gradient of the objective function and the gradients of the equality constraints, which lead to the reformulation of the original problem. The Lagrange multiplier theorem for the objective function $f(x)$ and equality constraint $g(x)$ is stated as below:

$$L(x, \lambda) = f(x) + \lambda g(x)$$

Here, $\lambda \in \mathfrak{R}$ is the unique Lagrange multipliers such that $\partial f(x) = \lambda^T \partial g(x)$.

First order condition: A differentiable function f of one variable defined on an interval $F = [ae]$. If an interior-point $\hat{x}$ is a local/global minimizer, then $f'(\hat{x}) = 0$; if the left-end-point $\hat{x} = a$ is a local minimizer, then $f(a) \geq 0$; if the right-end-point $\hat{x} = e$ is a local minimizer, then $f(e) \leq 0$. First-order necessary condition (FONC) summarizes the three cases by a unified set of optimality/complementarity slackness conditions: $a \leq x \leq e$, $f'(x) = y^a + y^e$, $y^a \geq 0$, $y^e \leq 0$, $y^a(x - a) = 0$, $y^e(x - e) = 0$. Note: If $f(x) = 0$, then it is also necessary that $f(x)$ is locally convex at $\hat{x}$ for being a local minimizer.

Second order condition: Second-order condition states that if first-order condition satisfies and if $f(\hat{x}) \geq 0$, or the function f is strictly locally convex, then $\hat{x}$ is a local minimizer.

Duality: Consider the Lagrange multiplier theorem for objective function $f(x)$ and equality constraint $g(x)$ where $L(x, \lambda) = f(x) + \lambda g(x)$ and there is a Lagrange dual function $D(\lambda) = \inf_{x \in \mathfrak{R}^n} L(x, \lambda)$. λ and v are dual feasible if $\lambda \geq 0$ and $D(\lambda)$ is finite. The dual function D is always concave. Weak duality holds when $D(\lambda) \leq f(x)$.

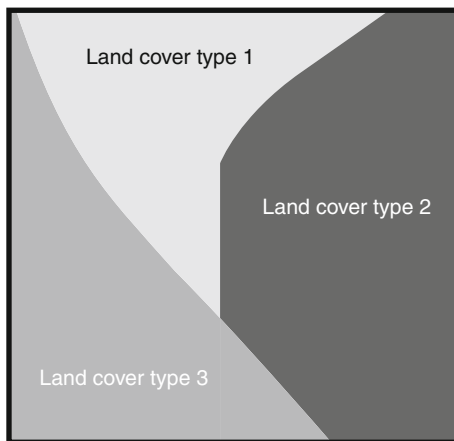
- (iii) **Karush–Kuhn–Tucker (KKT) conditions:** Constrained optimization with objective function having

both inequality and equality constraints can be solved as a more generalized method (as compared to the Lagrange method) called the Karush–Kuhn–Tucker or the KKT conditions. Similar to Lagrange method, the KKT condition method solves the constrained optimization problem using the Lagrange function $L(x, \lambda)$ whose optimal point is the saddle point.

Some of the other popular optimization algorithms like gradient descent, simulated annealing, local search algorithm are based on approximation methods in order to find a local or global minimum to optimize the objective function. As one of the most popular optimization algorithms, the gradient descent is a first-order iterative optimization algorithm for finding a local minimum of a differentiable objective function. The idea is to take repeated steps in the opposite direction of the gradient of the function at the current point, because this is the direction of steepest descent. In order to use gradient descent algorithm in the presence of constraints, projected gradient descent algorithm was proposed where the constraint is imposed on the feasible set of the parameters.

Constrained Optimization Application in Geospatial Science

Now, let us see one geospatial application of constrained optimization in land cover (LC) classification. Most land cover features occur at finer spatial scales compared to the resolution of primary remote sensing satellites. Therefore, observed data are a mixture of spectral signatures of two or more LC features resulting in mixed pixels (see Fig. 1); the mixed pixel problem has been commonly observed in hyperspectral remote sensing data. In other words, the observed data are mixture of spectral signatures of the individual objects present in mixed pixels, which could be due to two



Constrained Optimization, Fig. 1 A mixed pixel comprised of three land cover types

reasons: (1) the spatial resolution of the sensor is not high enough to separate different objects, these can jointly occupy a single pixel, and the resulting spectral measurement is a composite of the individual spectra, (2) mixed pixels may also result when distinct objects are combined into homogeneous (intimate) mixtures (Plaza et al. 2004).

One of the solutions to the mixed pixel problem is the use of spectral unmixing techniques such as linear mixture model (LMM) or linear unmixing to disintegrate pixel spectrum into its constituent spectra. Here, the observation vector in the low spatial resolution imagery is modeled as a linear combination of the endmembers, and given the endmembers present in the scene, the fractional abundances are extracted through the ordinary least squares error approaches. In most cases, the fractional abundances are constrained to be positive and sum to one (Heinz and Chang 2001). LMM infers a set of pure object's spectral signatures (called endmembers) and fractions of these endmembers (abundances) in each pixel of the image, that is, the objects are present with relative concentrations weighted by their corresponding abundances. The endmembers can be either derived using endmember extraction algorithms, from the image pixels, or obtained from an endmember spectral library available a priori. In LMM, the observation vector $\mathbf{y}$ for each pixel is related to endmember signature $\mathbf{E}$ by a linear model as

$$\mathbf{y} = \mathbf{E}\boldsymbol{\alpha} + \boldsymbol{\eta} \quad (1)$$

where each pixel is a M -dimensional vector $\mathbf{y}$ whose components are the digital numbers corresponding to the M spectral bands. $\mathbf{E} = [e_1, \dots, e_{n-1}, e_n, e_{n+1}, \dots, e_N]$ is a $M \times N$ matrix, where N is the number of classes, $\{e_n\}$ is a column vector representing the spectral signature of the n th target material, and $\boldsymbol{\eta}$ accounts for the measurement noise. For a given pixel, the abundance of the n th target material present in the pixel is denoted by α_n , and these values are the components of the N -dimensional abundance vector $\boldsymbol{\alpha}$. Further, components of the noise vector $\boldsymbol{\eta}$ are zero-mean random variables that are independent and identically distributed (i.i.d.). Therefore, covariance matrix of the noise vector is $\sigma^2 \mathbf{I}$, where σ^2 is the variance and $\mathbf{I}$ is $M \times M$ identity matrix.

LMM renders an optimal solution that is in unconstrained or partially constrained or fully constrained form. A partially constrained model imposes either the abundance non-negativity constraint (ANC) or abundance sum-to-one constraint (ASC) while a fully constrained model imposes both. ANC restricts the abundance values from being negative and ASC confines the sum of abundances of all the classes to unity. The conventional approach to extract the abundance values is to minimize $\|\mathbf{y} - \mathbf{E}\boldsymbol{\alpha}\|$ as in (2):

$$\boldsymbol{\alpha}_{\text{UCLS}} = (\mathbf{E}^T \mathbf{E})^{-1} \mathbf{E}^T \mathbf{y} \quad (2)$$

which is termed as unconstrained least squares (UCLS) estimate of the abundance. UCLS with full additivity is a nonstatistical, nonparametric algorithm that optimizes a squared-error criterion but does not enforce the non-negativity and unity conditions.

To avoid deviation of the estimated abundance fractions, ANC given in (3) and ASC given in (4) are imposed on the model in fully constrained least squares (FCLS):

$$\alpha_n \geq 0 \forall n : 1 \leq n \leq N \quad (3)$$

$$\sum_{n=1}^N \alpha_n = 1 \quad (4)$$

ANC and ASC constrain the value of abundance in any given pixel between 0 and 1. When only ASC is imposed on the solution, the sum-to-one constrained least squares (SCLS) estimate of the abundance is

$$\alpha_{\text{SCLS}} = \mathbf{E}^T \mathbf{E}^{-1} \left(\mathbf{E}^T \mathbf{y} - \frac{\lambda}{2} \mathbf{1} \right), \quad (5)$$

where

$$\lambda = \frac{2 \left(\mathbf{1}^T (\mathbf{E}^T \mathbf{E})^{-1} \mathbf{E}^T \mathbf{y} - 1 \right)}{\mathbf{1}^T (\mathbf{E}^T \mathbf{E})^{-1} \mathbf{1}}. \quad (6)$$

The SCLS solution may have negative abundance values, but they add to unity. FCLS (Chang 2003; Heinz and Chang 2001) extends non-negative least squares (NNLS) algorithm (Lawson and Hanson 1995) to minimize $\|\mathbf{E}\alpha - \mathbf{y}\|$ subject to $\alpha \geq 0$ by including ASC in the signature matrix $\mathbf{E}$ by a new signature matrix (**SME**) defined by

$$\mathbf{SME} = \begin{bmatrix} \theta \mathbf{E} \\ \mathbf{1}^T \end{bmatrix} \quad (7)$$

with

$$\mathbf{1} = (11111 \dots N \text{ times})^T \text{ and } \mathbf{s} = \begin{bmatrix} \theta \mathbf{y} \\ \mathbf{1} \end{bmatrix} \quad (8)$$

θ in (7) and (8) regulates ASC. Using these two equations, the FCLS algorithm is directly obtained from NNLS by replacing signature matrix $\mathbf{E}$ with **SME** and pixel vector $\mathbf{y}$ with $\mathbf{s}$. ANC is a major difficulty in solving constrained linear unmixing problems as it forbids the use of Lagrange multiplier. One solution is the replacement of $\alpha_n \geq 0 \forall n : 1 \leq n \leq N$ with absolute ASC (AASC), $\sum_{n=1}^N |\alpha_n| = 1$ (Chang 2003).

AASC allows usage of Lagrange multiplier along with exclusion of negative abundance values. This leads to optimal constrained least squares solution satisfying both ASC and AASC, known as Modified FCLS (MFCLS). MFCLS utilizes SCLS solution and the algorithm terminates with all nonnegative components (Kumar et al. 2017).

Case Study

Problem: Solve the below constraint optimization problem by applying KKT conditions for the objective function $f(x, y)$ subject to the constraint function $h_1(y)$ and $h_2(y)$.

$$\min f(x, y) = (x - 2)^2 + 2(y - 1)^2$$

subject to:

$$h_1(x, y) \Rightarrow x + 4y \leq 3 \quad (\text{A})$$

$$h_2(x, y) \Rightarrow -x + y \leq 0 \quad (\text{B})$$

Solution: Applying generalized Lagrangian function:

$$L(x, y, \lambda_1, \lambda_2) = f(x, y) + \lambda_1 h_1(x, y) + \lambda_2 h_2(x, y)$$

$$L(x, y, \lambda_1, \lambda_2) = (x - 2)^2 + 2(y - 1)^2 + \lambda_1 (x + 4y - 3) + \lambda_2 (-x + y)$$

with $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$. Calculating the partial derivatives to find the optimal point as below:

$$\frac{\partial L}{\partial x} = 2(x - 2) + \lambda_1 - \lambda_2 = 0 \quad (9)$$

$$\frac{\partial L}{\partial y} = 4(y - 1) + 4\lambda_1 + \lambda_2 = 0 \quad (10)$$

$$\lambda_1 (x + 4y - 3) = 0 \quad (11)$$

$$\lambda_2 (-x + y) = 0 \quad (12)$$

$$x + 4y \leq 3 \quad (13)$$

$$-x + y \leq 0 \quad (14)$$

$$\lambda_1 \geq 0 \quad (15)$$

$$\lambda_2 \geq 0 \quad (16)$$

We need to check four possible cases as shown below:

Case I: When $\lambda_1 = 0$ and $\lambda_2 = 0$.

Using Eqs. 1 and 2, we have $x = 2$ and $y = 1$; these values of x and y do not satisfy Eqs. 5 and 6. So this is not a possible or optimal solution.

Case II: When primal constraint A is inactive and primal constraint B is active, that is, $\lambda_1 = 0$, and $-x + y = 0$.

then substituting the above values in Eqs. 1 and 2 gives:

$$2(x - 2) - \lambda_2 = 0$$

and

$$4(y - 1) + \lambda_2 = 0.$$

By solving above two equations, we have:

$$x = \frac{4}{3}, y = \frac{4}{3}, \lambda_2 = -\frac{4}{3}$$

The values of λ_2 does not satisfy Eqs. 7 and 8. So, this is not a possible or optimal solution.

Case III: When primal constraint A is active and primal constraint B is inactive, that is, $x + 4y = 3$ and $\lambda_2 = 0$ then substituting the above values in Eqs. 1 and 2 gives:

$$2(x - 2) + \lambda_1 = 0$$

and

$$4(y - 1) + 4\lambda_1 = 0.$$

By solving the above two equations, we have:

$$x = \frac{5}{3}, y = \frac{1}{3}, \lambda_1 = \frac{2}{3}$$

The values of x, y, λ_1 , and λ_2 satisfy all Eqs. 1–8. So, this case is one of the possible or optimal solution.

Case IV: When both primal constraint A and B are active, that is, $x + 4y = 3$ and $-x + y = 0$.

then by solving above two equations we have:

$$x = \frac{3}{5}, y = \frac{3}{5}.$$

Substituting the above values in Eqs. 1 and 2:

$$\lambda_1 = \frac{22}{25}, \lambda_2 = -\frac{48}{25}$$

The values of λ_2 do not satisfy Eq. 8. So this is not a

possible or optimal solution. Hence, the only possible solution to the objective function f is case III. Therefore,

$$\text{Optimal solution : } x = \frac{5}{3}, y = -\frac{1}{3}$$

$$\text{Optimal value : } f^*(x, y) = \left(\frac{5}{3} - 2\right)^2 + 2\left(\frac{1}{3} - 1\right)^2$$

$$\text{Optimal value : } f^*(x, y) = 1$$

Summary and Conclusions

In this chapter, we saw some basic method of solutions for solving constrained optimization problem. Solution methods such as substitution method, Lagrange multipliers, and KKT conditions were discussed in detail with respective case studies at the end of the chapter. Furthermore, the global optimality conditions can be checked by KKT conditions and the dual problem provides a lower-bound and an optimality gap.

Cross-References

- [Convex Analysis](#)
- [Hyperspectral Remote Sensing](#)
- [Least Squares](#)
- [Linear Unmixing](#)
- [Ordinary Least Squares](#)

Bibliography

- Chang C-I (2003) Hyperspectral imaging techniques for spectral detection and classification. Kluwer Academic/Plenum Publishers, New York
- David Lay, Steven Lay Judi McDonald, Linear Algebra and its Applications
- Fletcher R (2013) Practical methods of optimization. Wiley
- Gilbert Strang, Linear Algebra and its Applications, 4th Edition
- Heinz DC, Chang C-I (2001) Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. IEEE Trans Geosci Remote Sens 39(3):529–545
- Kenneth Hoffman Ray Kunze, Linear Algebra, 2nd Edition
- Kumar U, Ganguly S, Nemani RR, Raja KS, Milesi C, Sinha R, Michaelis A, Votava P, Hashimoto H, Li S, Wang W, Kalia S, Gayaka S (2017) Exploring subpixel learning algorithms for estimating global land cover fractions from satellite data using high performance computing. Remote Sens 9(11):1105
- Lawson L, Hanson RJ (1995) Solving least squares problems. SIAM, Philadelphia, PA
- Plaza J, Martinez P, Pérez R, Plaza A (2004) Nonlinear neural network mixture models for fractional abundance estimation in AVIRIS Hyperspectral Images. Proceedings of the NASA Jet Propulsion Laboratory AVIRIS Airborne Earth Science Workshop, Pasadena, California, March 31–April 2, 2004

Convex Analysis

Indu Solomon and Uttam Kumar

Spatial Computing Laboratory, Center for Data Sciences,
International Institute of Information Technology Bangalore
(IIITB), Bangalore, Karnataka, India

Definition

Convex analysis is a subfield of mathematical optimization, the word meaning of “optimal” is “best” and optimization techniques strive to bring out the best solution. Convex analysis studies the properties of convex functions and convex sets and aims to minimize convex functions over convex sets. It has application in various fields such as automatic control systems, signal processing, deep learning, data analysis and model building, finance, statistics, economics, and so on.

Introduction

In the beginning of 1960s, the works of mathematicians Ralph Tyrrell Rockafellar and Jean Jacques Moreau brought about great advancement in the field of convex analysis. An example for optimization in the field of economics could be profit maximization or cost minimization. As the system complexity increases, analytical solutions become infeasible, and we can reach the optimal solution only through convex analysis and optimization. The goal of convex analysis is to determine x^* , such that $f(x^*) \leq f(x)$ (Au 2007) for all other points of x . Loss functions used in machine learning techniques are convex in nature, and hence, best model parameters are estimated through optimization techniques like gradient decent. Linear programming is a special case of convex optimization where the objective function is linear and the constraints consist of linear equalities and inequalities.

An optimization problem can be defined as finding x that satisfies the following objective function

$$\begin{aligned} & \text{minimize} && f_0(x) \\ & \text{subject to} && f_i(x) \leq 0, \quad i = 1, \dots, m \\ & && g_j(x) = 0, \quad j = 1, \dots, p \end{aligned} \quad (1)$$

where $f_0: \mathcal{R}^n \rightarrow \mathcal{R}$ is the objective function to be optimized and $x \in \mathcal{R}^n$ is the n -dimensional variable, $f_i(x)$ are inequality constraints, and $g_j(x)$ are equality constraints of the objective function. A valid solution exists if the optimization variable x is a feasible point and x has to satisfy m inequality constraints, $f_i: \mathcal{R}^n \rightarrow \mathcal{R}$ and p equality constraints $g_i: \mathcal{R}^n \rightarrow \mathcal{R}$. A convex optimization problem is the one which satisfies an

extra requirement, that is, all functions $\{f_0, \dots, f_m\}$ are convex functions and $\{g_1, \dots, g_p\}$ are affine functions.

Affine Functions

A function $f: \mathcal{R}^n \rightarrow \mathcal{R}^m$ is affine, if it is a sum of linear function and a constant. It has a form $f(x) = Ax + b$, where $A \in \mathcal{R}^{m \times n}$ and $b \in \mathcal{R}^m$.

Affine Sets

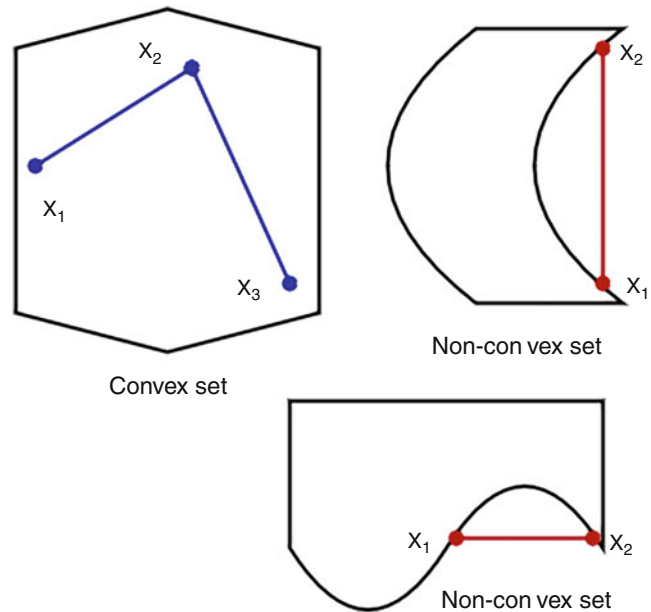
A set $C \subseteq \mathcal{R}^n$ is an affine set if a line connecting any two points in C lies in C . If any $x_1, x_2 \in C$ and $\theta \in \mathcal{R}$, $\theta x_1 + (1 - \theta)x_2 \in C$ (Boyd et al. 2004). Therefore, the set C must contain all linear combination of points in C , provided the coefficients θ_i in the linear combination sum to one. Examples are line, plane, and hyperplane.

Convex Sets

Any set C is convex if the line segment between any two points in C lies in C . For any $x_1, x_2 \in C$ and any θ in $0 \leq \theta \leq 1$, the combination is given by

$$\theta x_1 + (1 - \theta)x_2 \in C. \quad (2)$$

Every affine set is also convex. A convex combination of points in C is the weighted average of points with θ_i as weight for the point x_i . Examples of convex and non-convex sets are shown in Fig. 1.



Convex Analysis, Fig. 1 Convex sets and nonconvex sets

Convex Functions

A function $f: \mathcal{R}^n \rightarrow \mathcal{R}$ is convex, if $\text{dom}f$ is a convex set, and if for all $x, y \in \text{dom}f$ and $\theta, 0 \leq \theta \leq 1$ given by

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) \quad (3)$$

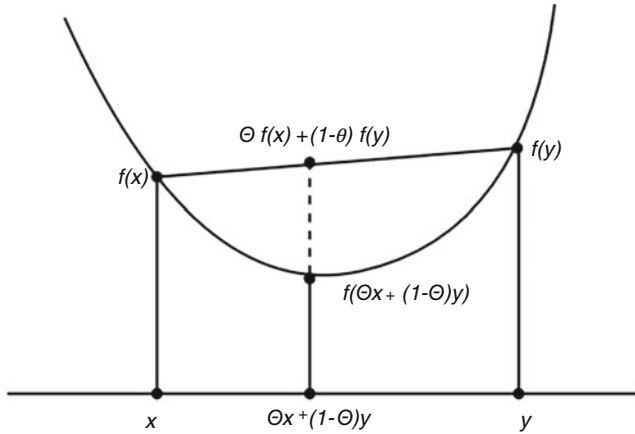
where $\text{dom}f$ is the domain of function f and is the subset of points of $x \in \mathcal{R}^n$ for which function f is defined (Boyd et al. 2004). The criteria in Eq. (3) are geometrically explained in Fig. 2. It means that the function f lies below the line segment connecting any two points on the function.

First-Order Conditions

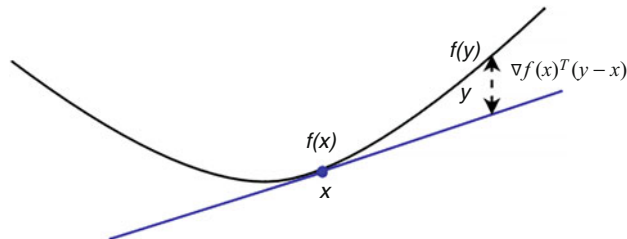
If the function f is differentiable and ∇f exists at each point in $\text{dom}f$, then f is convex if and only if $\text{dom}f$ is convex and

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) \quad (4)$$

for all $x, y \in \text{dom}f$. Figure 3 is the geometrical representation of the inequality described in Eq. (4) and it states that a convex function always lies above its tangent. Hence, the tangent to a convex function is its global underestimator and if we have the local information about a function, we can derive the global information.



Convex Analysis, Fig. 2 Convex function criteria



Convex Analysis, Fig. 3 First-order condition

Second-Order Conditions

Suppose if f is twice differentiable and its Hessian or second derivative exists at each point in $\text{dom}f$, then the function f is convex iff its $\text{dom}f$ is convex and its second derivative or hessian is positive semi-definite for all $x \in \text{dom}f$.

$$\nabla^2 f(x) \geq 0 \quad (5)$$

The condition $\nabla^2 f(x) \geq 0$ can be geometrically interpreted as the function $f(x)$ must have positive curvature. The gradient ∇ of $f(x)$ with $x \in \mathcal{R}^n$ is given by

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}. \quad (6)$$

The hessian ∇^2 for $f(x)$ with $x \in \mathcal{R}^n$ is given by,

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \quad (7)$$

Example-1: $f = x^2 + xy + y^2$

$$\nabla f = \begin{bmatrix} 2x + y \\ x + 2y \end{bmatrix}$$

$$\nabla^2 f = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

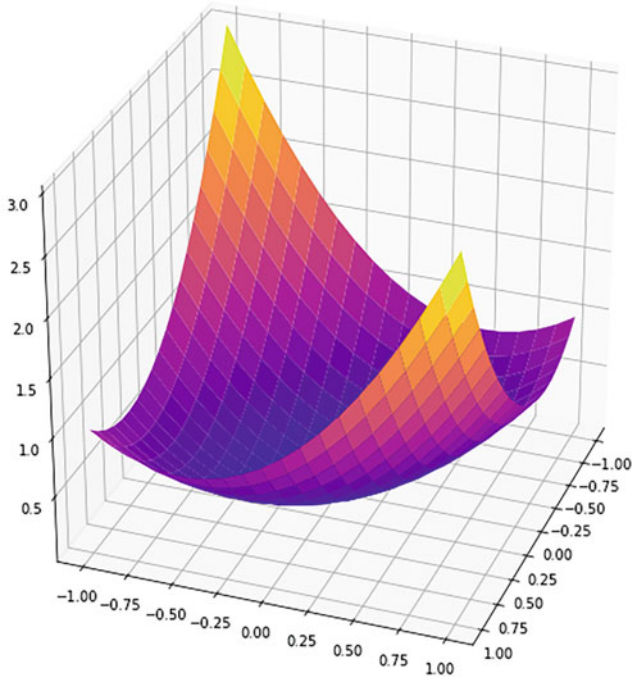
Eigen values of $\nabla^2 f$ is $\lambda = 1, 3$ and the eigen values are > 0 . Hence, $\nabla^2 f$ is positive semi-definite and the second-order conditions of convexity are satisfied (Fletcher 2013). Figure 4 shows the 3D plot of $f = x^2 + xy + y^2$.

Example-2: $f = x^2 + 3xy + y^2$

$$\nabla f = \begin{bmatrix} 2x + 3y \\ 3x + 2y \end{bmatrix}$$

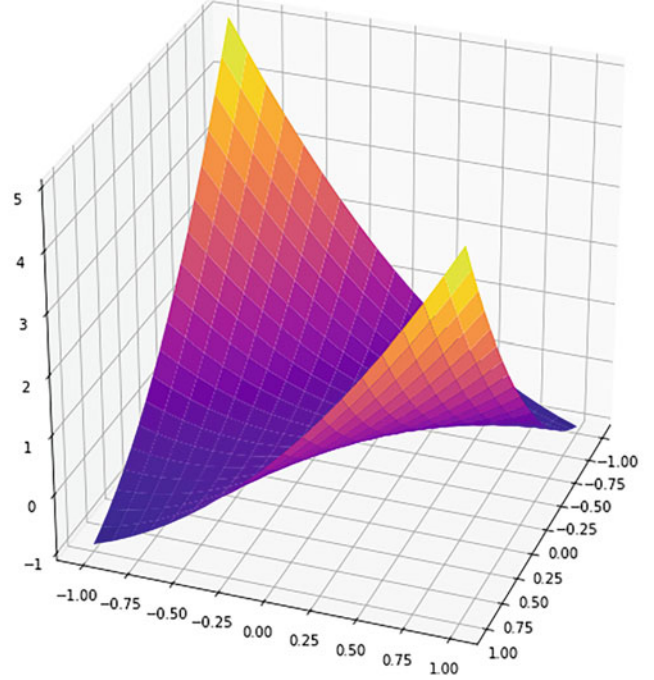
$$\nabla^2 f = \begin{bmatrix} 2 & 3 \\ 3 & 2 \end{bmatrix}$$

Eigen values of $\nabla^2 f$ are $\lambda = 5, -1$, and all the eigen values are not ≥ 0 . Hence, $\nabla^2 f$ is not positive semi-definite and the



Convex Analysis, Fig. 4 3D plot of the Convex function $f = x^2 + xy + y^2$

second-order conditions of convexity are not satisfied. Figure 5 shows the 3D plot of $f = x^2 + 3xy + y^2$.



Convex Analysis, Fig. 5 3D plot of the concave function $f = x^2 + 3xy + y^2$

Operations that preserve Convexity

Non-negative Weighted Sums

If f is a convex function and $\alpha \geq 0$, then αf is a convex function.

Sum

If f_1 and f_2 are convex functions, then their sum $f_1 + f_2$ is also convex. A non-negative weighted sum of convex functions is convex.

Point-wise maximum

If $\{f_1, f_2, \dots, f_n\}$ are convex functions, then $f = \max \{f_1, f_2, \dots, f_n\}$ is convex.

Duality

The Lagrangian

Consider the optimization problem defined in Eq. (1). Lagrangian duality is used to modify the objective function by including the constraints. The updated objective function

is a weighted sum of the constraints with the existing objective function. Lagrangian L is defined as $L : \mathcal{R}^n \times \mathcal{R}^m \times \mathcal{R}^p \rightarrow \mathcal{R}$.

$$L(x, \lambda, v) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p v_i h_i(x) \quad (8)$$

In Eq. (8), λ_i is the Lagrange multiplier associated with the i^{th} inequality constraint $f_i(x) \leq 0$ and v_i is the Lagrange multiplier associated with the i^{th} equality constraint $h_i(x) = 0$.

The Lagrange Dual Function

Lagrange dual function $g : \mathcal{R}^m \times \mathcal{R}^p \rightarrow \mathcal{R}$ is the minimum value of the Lagrangian over x : for $\lambda \in \mathcal{R}^m, v \in \mathcal{R}^p$ given by

$$\begin{aligned} g(\lambda, v) &= \inf_{x \in \mathcal{D}} L(x, \lambda, v) \\ &= \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p v_i h_i(x) \right) \end{aligned} \quad (9)$$

The dual function is a pointwise infimum of a family of affine functions, and hence, it is concave. The dual function gives the lower bound of the optimal value p^* for any $\lambda \geq 0$ and any v , $g(\lambda, v) \leq p^*$. Let $\hat{x}$ be a feasible point, then $f_i(\hat{x}) \leq 0$ and $h_i(\hat{x}) = 0$ and $\lambda_i \geq 0$. Mathematically,

$$\sum_{i=1}^m \lambda_i f_i(\hat{x}) + \sum_{i=1}^p v_i h_i(\hat{x}) \leq 0 \quad (10)$$

Then Eq. (9) becomes

$$\begin{aligned} L(x, \lambda, v) &= \left(f_0(\hat{x}) + \sum_{i=1}^m \lambda_i f_i(\hat{x}) + \sum_{i=1}^p v_i h_i(\hat{x}) \right) \\ &\leq f_0(\hat{x}) \end{aligned} \quad (11)$$

Therefore,

$$g(\lambda, v) = \inf_{x \in \mathcal{D}} L(x, \lambda, v) \leq L(x, \lambda, v) \leq f_0(\hat{x}) \quad (12)$$

Every feasible point satisfies $g(\lambda, v) \leq f_0(\hat{x})$ and hence for every feasible point, $g(\lambda, v) \leq p^*$. For a nontrivial lower bound p^* , $\lambda \geq 0$ and $\lambda, v \in \text{dom } g$.

Example – Least Squares Solution

Minimize $x^T x$ subject to $Ax = b$. The problem has equality constraints and the Lagrangian is given by, $L(x, v) = x^T x + v(Ax - b)$ in the domain $R^n \times R^p$. The dual function is given by $g(v) = \inf_x L(x, v)$. The optimality condition is given by

$$\nabla_x L(x, v) = 2x + A^T v = 0 \quad (13)$$

which gives $x = -\frac{1}{2}A^T v$. Hence, the dual function can be written as $g(v) = -\frac{1}{4}v^T A A^T v - b^T v$ and the dual function is a concave quadratic function with domain R^p . The lower bound property of dual function for any $v \in R^p$ gives

$$-\frac{1}{4}v^T A A^T v - b^T v \leq \inf \{x^T x | Ax = b\} \quad (14)$$

Examples of Optimization Techniques in Geospatial Applications

Spatial data analysis and linear programming can be combined to obtain an optimal solution to optimal location problems. An optimal geographical location for a desired building of a restaurant/mall/park/anything of interest can be found by setting the mathematical model with the right optimization criteria (Guerra and Lewis 2002). Convex optimization technique was used for hyperspectral image visualization (Cui et al. 2009). Maintaining the spectral distances, spectral discrimination and interactive visualization were three objectives of hyperspectral image visualization techniques. A combination of principal component analysis and linear programming techniques was used for mapping the spectral

signature of pixels to RGB colors. Cat swarm optimization technique was used for feature extraction from satellite images (Prabhakaran et al. 2018). Cat optimized algorithm distinguishes the inner, outer, and extended boundary along with the land cover. Examples of convex analysis in geospatial applications are terrain analysis (Prabhakaran et al. 2018), landslide susceptibility mapping (Pham et al. 2019), and in interactive hyperspectral visualization (Cui et al. 2009).

Summary and Conclusions

In this chapter, we have explored mathematical optimization and especially convex optimization. Optimization concepts like affine sets, convex sets, and convex functions were explained. The first-order and second-order optimality conditions for a convex function put a constraint on the function that the hessian of the function must be positive semi-definite for the function to be convex. Duality concept combines the function and the constraints into a single equation with the help of the Lagrange multiplier. Optimization concepts are useful in geospatial applications and a few examples were mentioned in this chapter.

Cross-References

- [Constrained Optimization](#)
- [Hyperspectral Remote Sensing](#)
- [Optimization in Geosciences](#)

Bibliography

- Au JKT (2007) An ab initio approach to the inverse problem-based design of photonic bandgap devices (Doctoral dissertation, California Institute of Technology)
- Boyd S, Boyd SP, Vandenberghe L (2004) Convex optimization. Cambridge university press
- Cui M, Razdan A, Hu J, Wonka P (2009) Interactive hyperspectral image visualization using convex optimization. IEEE Trans Geosci Remote Sens 47(6):1673–1684
- Fletcher R (2013) Practical methods of optimization
- Guerra G, Lewis J (2002) Spatial optimization and gis. Locating and optimal habitat for wildlife reintroduction. McGill University
- Prabhakaran N, Ramakrishnan SS, Shanker NR (2018) Geospatial analysis of terrain through optimized feature extraction and regression model with preserved convex region. Multimed Tools Appl 77(24): 31855–31873
- Pham BT, Prakash I, Chen W, Ly HB, Ho LS, Omidvar E et al (2019) A novel intelligence approach of a sequential minimal optimization-based support vector machine for landslide susceptibility mapping. Sustainability 11(22):6323

Copula in Earth Sciences

Sandra De Iaco and Donato Posa

Department of Economics, Section of Mathematics and Statistics, University of Salento, Lecce, Italy

Synonyms

Multivariate distribution; Spatial dependence

Definition

A **copula** is a multivariate distribution, defined on the unit hypercube, which is characterized by uniform marginals. In particular, it describes the dependence of **multivariate distributions** by marginal distributions. By recalling the result of Sklar (1959), any multivariate distribution $F(z_1, z_2, \dots, z_n)$ can be written as a copula of its one-dimensional marginal distribution $F_i(z)$, $i = 1, 2, \dots, n$, that is

$$F(z_1, \dots, z_n) = C(F_1(z_1), F_2(z_2), \dots, F_n(z_n)). \quad (1)$$

Moreover, the function C in (1) is unique, if the **marginal distributions** are assumed to be continuous.

Introduction

Quality parameters in Earth Sciences are characterized by significant spatial variability. In Geostatistics the evaluation of this variability is based on the use of the variogram. In addition, copulas provide a valid chance to give information on the dependence structures for multivariate distributions and in particular bi-dimensional copulas can be an alternative to the use of the traditional measures of spatial correlation (i.e., variograms and covariance functions) in describing the spatial variability.

Thus, the copula-based models are widely utilized in stochastic interpolation (alternatively to kriging) or for simulation.

Measures of Spatial Variability

Traditionally, the **spatial variability** of various variables encountered in Earth Sciences are analyzed by using geostatistical methods (Matheron 1971). For geostatistical studies, spatial observations, for a given variable Z of interest, are assumed to be a finite realization of a random function (or random field), which is usually supposed to satisfy the

hypothesis of second-order stationarity or at least the hypothesis of intrinsic stationarity. Under these conditions, the covariance function or the variogram of Z at two different locations $\mathbf{x}_1$ and $\mathbf{x}_2$ of the spatial domain, depends on the spatial separating vector $\mathbf{h} = \mathbf{x}_1 - \mathbf{x}_2$ and can be estimated from the available sample data. Thus, the spatial variability or spatial correlation is well described by these moments, however they cannot highlight potential differences in the spatial dependence related to extreme values or to central values, since they represent the spatial correlation obtained as an integral over the whole distribution of the variable values. Regarding this aspect, it is well known that different quantiles can be characterized by a different spatial dependence structure (Journel and Alabert 1989). For example, extreme values can have a spatial dependence structure which is different from the one related to central values.

Indicator variograms represent a way to assess the difference in dependence with respect to the values of the variable under study. For this reason, the indicator approach used in simulation and interpolation studies requires that variogram models are fitted for each selected threshold. However, an indicator variogram model is fitted for each threshold separately and it is not applied any stochastic model. Therefore the results might be not consistent; the monotonicity of the estimated indicator variogram (corresponding to the estimated distribution function) for the various cutoff values might not be guaranteed. With the support of the indicator variograms, Journel and Alabert (1989) pointed out that the dependence between variables often deviates significantly from the **Gaussian**. Another advantage in using indicator variogram emerges when data have skewed distributions, as usually happened in Earth Sciences applications. In these cases, the estimation of the variogram (which is based on the squared differences between the values measured at different locations) might be severely influenced by few anomalous squares. These problems can be mitigated by adopting the indicator approach, with the effort of computing several indicator variograms. In alternative, another technique to be used to face the same problem is to consider the ranks for interpolation, as illustrated in Journel and Deutsch (1996).

In the following, the concept of copulas and its help in the study of the spatial variability of earth quality variables is underlined. In particular, it is clarified that the dependence structure between random variables can be represented through copulas with no information on the marginal distributions. In other terms, copulas is interpreted as the key measure of the dependence over the range of quantiles.

Some Fields of Application

Nowadays, an emerging interest in applying copulas in the spatial context is noticeable. There are numerous applications

of copulas in finance, where it is important to study the dependence between extremes (Embrechts et al. 2002) and to adopt appropriate non-Gaussian type of dependence for the estimation of financial risks. The non-Gaussian structure of dependence can also be found in other contexts, such as in hydrology, for groundwater modeling (Gomez and Wen 1998), where the consequences of the departure from normality on some aspects regarding ground water flow and transport were analyzed. Copulas were also applied on stochastic rainfall simulation or on extreme value statistics, among other by De Michele and Salvadori (2003); Favre et al. (2004).

Basic Concepts

As already specified, copulas are multivariate distributions defined on the unit hypercube with uniform marginals. In particular, for $n = 2$ the definition of bivariate copula (or bi-dimensional copula) C can be given as follows:

$$C : [0, 1]^2 \rightarrow [0, 1] \quad (2)$$

and has to fulfill the following properties:

- I $C(u, 0) = C(0, u) = 0$ and $C(1, u) = C(u, 1) = u$,
- II given $u_1 \geq v_1$ and $u_2 \geq v_2$, then $C(u_1, u_2) + C(v_1, v_2) - C(u_1, v_2) - C(v_1, u_2) \geq 0$.

These last property guarantees the non-negativity of the probability associated to any rectangle in the unit square. For a n -dimensional space, the multivariate copula is a function

$$C : [0, 1]^n \rightarrow [0, 1] \quad (3)$$

which is characterized by uniform marginals,

$$C(\mathbf{u}^{(i)}) = u_i, \quad \text{if } \mathbf{u}^{(i)} = (1, \dots, 1, u_i, 1, \dots, 1),$$

and is zero if at least one of the arguments is zero, that is

$$C(\mathbf{u}) = 0, \quad \text{if } \mathbf{u} = (u_1, \dots, u_i, \dots, u_n), \\ \exists i : u_i = 0.$$

For every n -dimensional hypercube in the domain $[0, 1]^n$ (called unit hypercube), that is for $[\mathbf{u}', \mathbf{v}'] \subseteq [0, 1]^n$, the associated probability has to be nonnegative:

$$V_C([\mathbf{u}', \mathbf{v}']) \geq 0,$$

where $V_C([\mathbf{u}', \mathbf{v}']) = \Delta_{u'_n}^{v'_n} \Delta_{u'_{n-1}}^{v'_{n-1}} \dots \Delta_{u'_1}^{v'_1} C(\mathbf{u})$ is the n -difference of C on $[\mathbf{u}', \mathbf{v}']$ and a first-order difference is defined as $\Delta_{u'_k}^{v'_k} C(\mathbf{u}) = C(u_1, \dots, u_{k-1}, v'_k, u_{k+1}, \dots, u_n) - C(u_1, \dots, u_{k-1}, u'_k, u_{k+1}, \dots, u_n)$.

The readers may refer to Joe (1997) or Nelsen (1999) for more information.

Some Features of Copulas

If C is assumed to be an absolutely continuous function, the corresponding copula density, denoted with $c(\cdot)$, is expressed as

$$c(u_1, \dots, u_n) = \frac{\partial^n C(u_1, \dots, u_n)}{\partial u_1 \dots \partial u_n}. \quad (4)$$

In addition, it is worth introducing the conditional copula can which be written as

$$C(u_1 | U_2 = u_2, \dots, U_n = u_n) = \frac{1}{c(u_2, \dots, u_n)} \cdot \frac{\partial^{n-1} C(u_1, \dots, u_n)}{\partial u_2 \dots \partial u_n} \quad (5)$$

Thus, a copula C can be viewed as a multivariate cumulative distribution function, while a copula density c is interpreted as the strength of dependence. Given two independent variables, then the bivariate copula of their joint distribution function is obtained through the product of the marginals $C[F(z_1), F(z_2)] = F(z_1)F(z_2)$; for this reason, it is known as the product copula and the density copula is equal to 1. In the case of full dependence, the copula becomes $C[F(z_1), F(z_2)] = \min[F(z_1), F(z_2)]$.

The invariance of the copula of a multivariate distribution with respect to monotonic transformations of the marginal variables represents one of the main advantages. Thus, some common data transformations (a normal score transformation and Box-Cox transformations, included logarithms) do not affect the copula. On the basis of this characteristic, the use of copula is preferred to the approaches based on covariance functions or variograms, as these last are strongly influenced by the marginal distributions. Another relevant property of a copula is associated to the possibility of pointing out whether the corresponding dependence is related to the intensity of the variable realizations, in other term, whether high (low) values present a strong (weak) spatial dependence.

A bi-dimensional copula highlights a symmetrical dependence if

$$C(u_1, u_2) = C(1 - u_1, 1 - u_2) - 1 + u_1 + u_2, \quad (6)$$

or in terms of its density

$$c(u_1, u_2) = c(1 - u_1, 1 - u_2). \quad (7)$$

Note that the symmetric property is given with respect to the axis $u_2 = 1 - u_1$ of the unit square.

Copula Families

Building new families of copulas is not an easy task. Traditional copula families include the Archimedean class (Genest and MacKay 1986) as well as the Farlie-Gumbel-Morgenstern class (Nelsen 1999). Moreover, the Gaussian copulas is another well-known family of copulas which is used to explain the dependence form of the multivariate normal distribution. They are frequently used in hydrology, although the term copula is not popular; they are applied for regressions, or in case of normal score transformation. In Table 1, some well-known families of bivariate copula are given together with their parameter space and the dependence structure.

Note that the normal copula class presents a symmetric density, which can be properly expressed as follows:

$$c(u_1, \dots, u_n) = k_n \exp\left(0.5 \mathbf{z}^T (\boldsymbol{\Sigma}^{-1} - \mathbf{I}) \mathbf{z}\right) \quad (8)$$

where $u_i = \Phi(z_i)$, Φ is the standardized normal distribution function, k_n is a normalizing constant, $\boldsymbol{\Sigma}$ is the correlation matrix, and $\mathbf{I}$ is called identity matrix.

For $n = 2$, $\boldsymbol{\Sigma}$ depends on the only parameter ρ , which is the correlation coefficient between the two variables. A graphical representation of the bivariate Gaussian copula together with the associated copula density given $\rho = 0.85$ is shown in Fig. 1. Note that the symmetry of the density, as specified in (7). It is also worth highlighting that, for the user, the dependence can be better visualized through the density rather than the copula itself.

Similarly to the Gaussian family, several other bivariate copula classes have multivariate extensions, which are desirable in spatial modeling. However, these families are considered not flexible enough, since they depend only on one or two parameters. Moreover, it might happen that the multivariate distribution has different pairwise dependence structures across its margins, as a consequence a single family might not be able to capture this behavior. These restrictions can be get through with vine copulas, as will be clarified in one of the next sections. The introduction of a bivariate spatial copula into a vine copula for interpolation has been described by Gräler and Pebesma (2011) and is extended in Gräler (2014), where convex combinations of bivariate copulas parametrized by distance are combined in a vine copula (also known as pair-copula construction).

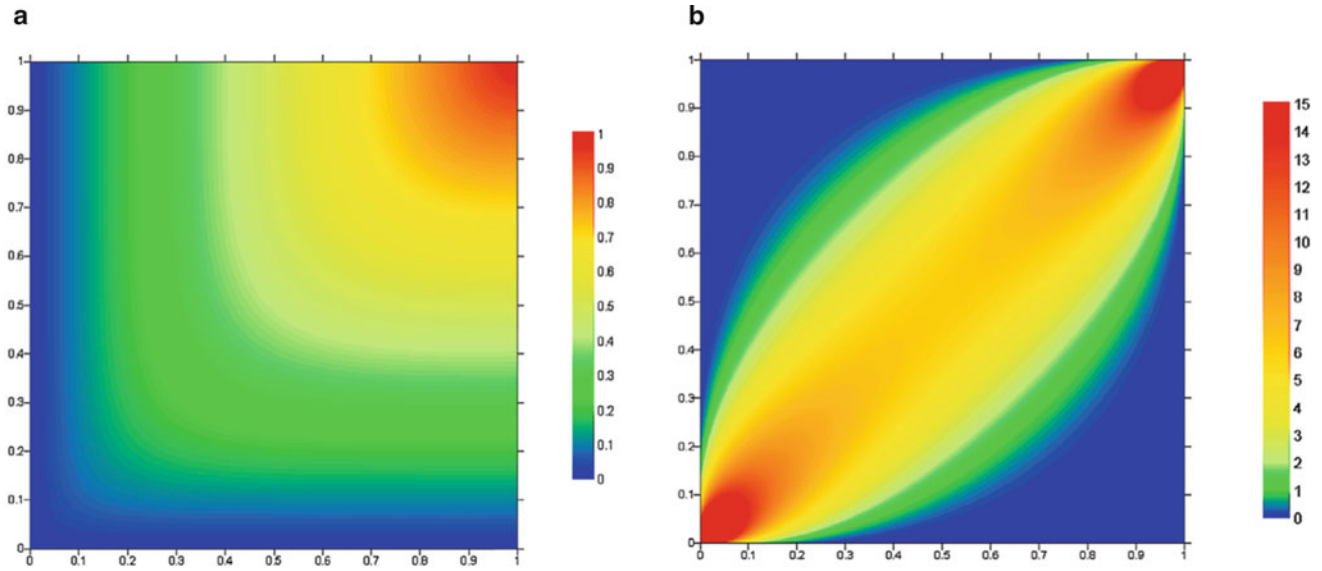
Copulas for Describing Spatial Variability

In spatial statistics, copulas characterize the joint multivariate distribution associated to the regionalized variables at different locations of the domain of interest. In this context, the variable at each location of the domain is assumed to have the same probability distribution.

Let $Z = \{Z(\mathbf{x}), \mathbf{x} \in D\}$ be a random function, where D represents the spatial domain (one-, two-, or three-

Copula in Earth Sciences, Table 1 Some well-known families of copulas

Name	Copula	Parameter	Dependence structure
Gaussian	$C_N(u_1, u_2; \rho) = \Phi(\Phi^{-1}(u_1), \Phi^{-1}(u_2))$	ρ	Tail independence: $\lambda_U = \lambda_L = 0$
Student- t	$C_{ST}(u_1, u_2; \rho, \nu) = T(t_v^{-1}(u_1), t_v^{-1}(u_2))$	ρ, ν	Symmetric tail dependence: $\lambda_U = \lambda_L =$ $2t_{\nu+1}\left(-\sqrt{\nu+1}\sqrt{1-\rho}/\sqrt{1+\rho}\right)$
Gumbel	$C_G(u_1, u_2; \delta) = \exp\{-(\log u_1)^\delta + (\log u_2)^\delta\}^{1/\delta}$	$\delta \geq 1$	Asymmetric tail dependence: $\lambda_U = 2 - 2^{1/\delta}, \lambda_L = 0$
Rotated Gumbel	$C_{RG}(u_1, u_2; \delta) = u_1 + u_2 - 1 + C_G(1 - u_1, 1 - u_2; \delta)$	$\delta \geq 1$	Asymmetric tail dependence: $\lambda_U = 0, \lambda_L = 2 - 2^{1/\delta}$
Plackett	$C_P(u_1, u_2; \theta) = \frac{1}{2(\theta-1)}[1 + (\theta-1)(u_1 + u_2)] +$ $-\sqrt{[1 + (\theta-1)(u_1 + u_2)]^2 - 4\theta(\theta-1)u_1u_2}$	$\theta \geq 0,$ $\theta \neq 1$	Tail independence: $\lambda_U = \lambda_L = 0$
Frank	$C_F(u_1, u_2; \theta) = -\frac{1}{\theta} \log\{1 + [(exp^{-\theta u_1} - 1)(exp^{-\theta u_2} - 1)/exp^{-\theta} - 1]\}$	$0 < \theta < \infty$	Tail independence: $\lambda_U = \lambda_L = 0$
Clayton	$C_{CL}(u_1, u_2; \theta) = \max\{(u_1^{-\theta} + u_2^{-\theta} - 1), 0\}$	$\theta > 0$	Asymmetric tail dependence: $\lambda_U = 2 - 2^{1/\theta}, \lambda_L = 0$



Copula in Earth Sciences, Fig. 1 (a) Gaussian copula characterized by correlation $\rho = 0.85$ and (b) the associated copula density

dimensional space) and $(\mathbf{x})$ a point in the same domain. The available spatial data represent a single finite realization of the random function Z , thus it is not possible to assess the multivariate copulas from them. However, in order to describe the spatial variability, it is interesting to introduce the bivariate marginal copulas, similarly to what is proposed with variograms or covariance functions. For this aim, according to a stationarity hypothesis, the spatial copula C_S of two random variables associated to two locations $\mathbf{x}$ and $\mathbf{x}'$ of the domain is assumed to be dependent on the separating vector $\mathbf{h} = \mathbf{x} - \mathbf{x}'$, rather than the specific locations $\mathbf{x}$ and $\mathbf{x}'$.

For any two selected quantiles u_1, u_2 ,

$$C_S(\mathbf{h}, u_1, u_2) = P\{F_Z[Z(\mathbf{x})] < u_1, F_Z[Z(\mathbf{x} + \mathbf{h})] < u_2\} \\ = C(F_Z[Z(\mathbf{x})], F_Z[Z(\mathbf{x} + \mathbf{h})]), \quad (9)$$

where F_Z denotes the marginal distribution of the variable Z and it is assumed to be independent from the location $\mathbf{x}$. Note that the behavior of any couple of variables at two locations, characterized by a separation vector $\mathbf{h}$, is described by the bivariate distribution function whose dependence is described by the bivariate copula.

Similarly to variogram, copula is not a function of the locations $\mathbf{x}$ and $\mathbf{x}'$, since it depends only on the separation vector $\mathbf{h}$. It is worth to underline that the hypothesis that the random function Z has bivariate marginal copulas which are translation invariant is stronger than the **second-order stationarity (or the intrinsic hypothesis)**.

Differently from the approach of Journel and Deutsch (1996) who used the variogram to describe the dependence, on average, in the rank space, this approach considers copulas to assess the dependence structure with respect to the entire range of quantiles.

It is worth noting that building a multivariate copula, given two-dimensional marginals, is not an easy task. For example, the Gaussian copula represents one of the multivariate copula obtained from its bivariate marginals, given a positive definite matrix (covariance matrix). Since this is not always possible, the construction of an alternative theoretical copula model representing the dependence of regionalized variables is presented just by way of explanation; even in this case, it is shown that the construction of the copula can be obtained from its bivariate marginals. The copulas can be constructed by transforming the variables and using their marginal distribution, similarly to the case of indicator variables.

Indeed, an evident link between copulas and indicator variograms can be provided. In particular, for any threshold z_α , the indicator variogram γ_α can be expressed as follows:

$$\gamma_\alpha(\mathbf{h}) = F_Z^{-1}(\alpha) - C_S[\mathbf{h}, F_Z^{-1}(\alpha), F_Z^{-1}(\alpha)]. \quad (10)$$

The indicator cross-variogram $\gamma_{\alpha_1, \alpha_2}$, for the thresholds α_1, α_2

$$\gamma_{\alpha_1, \alpha_2}(\mathbf{h}) = 0.5 \min[F_Z^{-1}(\alpha_1), F_Z^{-1}(\alpha_2)] \\ - C_S[\mathbf{h}, F_Z^{-1}(\alpha_1), F_Z^{-1}(\alpha_2)]. \quad (11)$$

In other terms, copulas provide the same information given by the direct and cross-indicator variograms; indeed copulas have the advantage to offer a representation with more parsimonious parameterization.

In conclusion, the hypothesis that the bivariate marginals is dependent on the lag vector $\mathbf{h}$ is analogous to the second-order stationarity hypothesis which is considered in

geostatistics, however the former is more general than the latter.

Estimation Aspects

The bivariate copulas estimation from the observations $z(\mathbf{x}_1), \dots, z(\mathbf{x}_n)$ at the samples locations $\mathbf{x}_1, \dots, \mathbf{x}_n$ can be obtained by first computing the sample distribution function $F_n(z)$. On the basis of this sample distribution function and for any given vector $\mathbf{h}$, the set $S(\mathbf{h})$ of those pairs of distribution function values associated to the observations at points with separation vector $\mathbf{h}$, is computed, that is

$$S(\mathbf{h}) = \{ [F_n(z(\mathbf{x}_i)), F_n(z(\mathbf{x}_j))] \mid (\mathbf{x}_i - \mathbf{x}_j) \approx \mathbf{h} \text{ or } (\mathbf{x}_j - \mathbf{x}_i) \approx \mathbf{h} \} \quad (12)$$

Note that $S(\mathbf{h})$ is a set of points in the unit square and is, by definition, symmetric with respect to the major axis $u_1 = u_2$ of the unit square, therefore, it is clear that if $(u_1, u_2) \in S(\mathbf{h})$, then $(u_2, u_1) \in S(\mathbf{h})$. In practice, the set $S(\mathbf{h})$ is defined for various lags $\mathbf{h}$. The sample copula density are easily computed through the sample frequencies $\hat{c}$, for a fixed lag $\mathbf{h}$, as follows:

$$\hat{c}\left(\mathbf{h}, \frac{2i-1}{2k}, \frac{2j-1}{2k}\right) = \frac{k^2}{|S(\mathbf{h})|} |S_{ij}(\mathbf{h})|, \quad i, j = 1, \dots, k, \quad (13)$$

where $|\cdot|$ denotes the cardinality and

$$S_{ij}(\mathbf{h}) = \left\{ (u_1, u_2) \in S(\mathbf{h}) : \frac{i-1}{k} < u_1 \leq \frac{i}{k} \text{ and } \frac{j-1}{k} < u_2 \leq \frac{j}{k} \right\}.$$

Note that if the cardinality set $S(\mathbf{h})$ is sufficiently high, a kernel smoothing can be used to approximate the bivariate density.

Copula Modeling

As done for structural analysis based on variogram, where the sample variogram is fitted to a model which satisfies the conditional negative definiteness condition, even the empirical copulas have to be described through a model. Thus, after the estimation of copulas, which reflect the spatial dependence structure of the observations, the modeling step is necessary.

It is important to highlight that in some cases, the complexity of phenomenon under study requires (for interpolation or simulation) a very high dimension of the dependence structure. Then, the multivariate copula (associated to the

random variables at the locations in the studied domain) should be characterized by bivariate marginals which recall the empirical bivariate copulas.

In modeling the multivariate copula, the properties listed below should be fulfilled:

1. The multivariate marginals should be stable, in the sense that any multivariate marginal copula for a fixed set of locations should not depend on the set of other fixed locations considered to define the multivariate copula.
2. The dependence should be characterized by wide range, with strong dependence for geographically close set of points and weaker and weaker dependence or independence for points at increasing distance.
3. The multivariate copula should be flexible enough that the dependence structure takes into account the geometry of the related set of points.

In literature, there exist various copula models, however not all satisfy the features which are desirable for analyzing the spatial dependence of random functions. The main problem derives from the difficulty to find a joint distribution (or copula) for n ($n > 2$) random variables that optimally honors the bivariate marginals. For this aim, although there are different conditions that ensure the existence of such a copula (Drouet and Kotz 2001), building a multivariate copula (with $n > 2$), on the basis of given bivariate marginals, is not an easy task and sometimes it is not possible at all on for the incompatibility of the bivariate marginals. Rueschendorf (1985) proposed an interesting method for constructing multivariate distributions and it can be used to define copulas (Drouet and Kotz 2001). Bardossy (2006) started from the selection of a multivariate distribution, then a suitable multivariate copula was found. Indeed, a multivariate copula can be defined through a multivariate distribution, then the multivariate model is obtained through a combination with the univariate marginal of the regionalized variable.

On the basis of **Sklar's theorem**, given a multi-dimensional distribution function F , whose margins are absolutely continuous, the Eq. (1) is valid, where C (uniquely determined) is a copula function. A copula function C is determined from the function F by considering

$$C(u_1, u_2, \dots, u_n) = F(F_1^{-1}(u_1), F_2^{-1}(u_2), \dots, F_n^{-1}(u_n)), \quad (14)$$

where F_i , $i = 1, 2, \dots, n$ are the corresponding univariate distribution functions of F . On the other hand, any copula C can be applied to combine a set of univariate distribution functions $F_1, F_2, \dots, F_n$ in order to build a multivariate distribution F . This technique is often adopted, for example, in case of multivariate normal or t copula.

Various well-known multivariate copula functions only depend on a restricted number of parameters which can cause that the dependence would not change with respect to the spatial distance between two points and that the third property above-mentioned would be violated. The multivariate normal copula and the Farlie-Gumbel-Morgenstern copula present $\binom{n}{2}$ parameters. However, the former is symmetric, that is satisfies Eq. (7), the latter has the drawback of showing a decaying correlation between pairs of variables with the increase in the number of points (Drouet and Kotz 2001). This last feature represents a critical aspect especially when the two variables corresponds to very close points $\mathbf{x}_1$ and $\mathbf{x}_2$ and it is expected to have a high correlation. Thus, this multivariate copula does not respect the first two properties of the above list. Copula models can be built on the basis of Eq. (14) starting from a given multivariate distribution. Moreover, interesting constructions of multivariate copulas can be obtained through non-monotonic transformations from the Gaussian copula; indeed, as underlined in Bardossy (2006), they are asymmetric. For example, given a nonmonotonic function $g(t)$ and an n -dimensional normal random variable $Y \sim N(\mathbf{m}, \mathbf{\Gamma})$, whose expected value is $\mathbf{m}^T = (m, \dots, m)$ and covariance matrix $\mathbf{\Gamma}$, then $\mathbf{V}$ is defined for each coordinate $j = 1, \dots, n$ as

$$V_j = g(Y_j). \quad (15)$$

For simplicity, all marginals are assumed to have unit variance. By using (15), the copula of the transformed multivariate distribution is not necessarily Gaussian. For example, the copula of the non-centered multivariate chi-square distribution can be easily obtained by considering the transformation function $g(t) = t^2$ and $V_j = Y_j^2$. In this way, the corresponding n -dimensional distribution has identical one-dimensional marginals. In particular, the marginal variables follows a χ^2 distribution with a degree of freedom equal to 1, under the hypothesis that the expected value is equal to zero ($\mathbf{m} = \mathbf{0}$). Thus, after getting the multivariate density function f_n , the multivariate copula density can be determined through the following equality:

$$f_n(z_1, z_2, \dots, z_n) = c(F_1(z_1), F_2(z_2), \dots, F_n(z_n)) \cdot f_1(z_1) \cdot f_2(z_2) \cdots f_n(z_n), \quad (16)$$

where f_i , $i = 1, 2, \dots, n$ are the marginal density.

The presence of an **asymmetric dependence** can be assessed by comparing the strength of the dependence of high values and the one related to low values. In particular, the following ratio of the joint probability between high values (exceeding the quantile $1 - u$) and the joint probability between low values (not exceeding a quantile u):

$$A(u) = \frac{2u - 1 + C(1 - u, 1 - u)}{C(u, u)} \quad (17)$$

is such that $A(u) > 1$ if high values present a stronger dependence than the low values, while $A(u) < 1$ if high values present a weaker dependence than the low values; the case $A(u) \approx 1$ occurs when the dependence is the same. Moreover, as highlighted in Joe (1997), the tail dependence for a given copula offers information about multivariate extreme value distributions, that is about the dependence in the upper and lower quadrants. However, the tail dependence is rarely applied in spatial statistics, since the main focus does not often regard the extremes, but the location where the observations that are greater than certain thresholds are recorded.

For the spatial context, the elements of the correlation matrix $\mathbf{\Gamma}$ are the correlation function $\rho(\mathbf{h})$, where $\mathbf{h} = \mathbf{x}_i - \mathbf{x}_j$, $i = 1, 2, \dots, n$. Thus, for any set of points $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, the generic element of the covariance matrix associated to the Gaussian variable $\mathbf{Y}$, from which the copula is generated, is $\mathbf{\Gamma}_{[ij]} = [\rho(\mathbf{x}_i - \mathbf{x}_j)]$. At this point, the estimation of the correlation function $\rho(\mathbf{h})$ of $\mathbf{Y}$, on the basis of the observations, has to be faced. However, this is not an easy task since on one hand $\mathbf{Y}$ is not an observed variable and on the other hand it cannot be obtained from the measured variable $\mathbf{Z}$ because the transformation function is non-monotonic.

Regarding modeling, note that $\mathbf{\Gamma}$ has to be a positive semi-definite matrix, then the function ρ has to be a positive semi-definite function. Nowadays, the family of positive semi-definite models includes a wide range of parametric functions, among which users can choose. The maximum likelihood method as well as the rank correlations and the asymmetry of the copulas can be used for the parameters estimation of the selected correlation model. Maximum likelihood is often applied only for the bivariate marginals rather than for the n -dimensional copula (whose density requires the sum of 2^n elements). Indeed, in the former case, the complexity is greatly reduced to the order of n^2 , since observations are paired for computation; however, the existence of independence among different pairs implicitly assumed is not supported by the model. Fitting the rank correlation is often done by graphical inspection or through least squares, analogously to the variogram fitting. If the least squares method is used, then it is necessary to calculate first the sample rank correlations corresponding to the empirical copulas and the theoretical rank correlation can be computed by using the Spearman's rank correlation (Joe 1997)

$$\rho' = 12 \int_0^1 \int_0^1 uv \cdot c(u, v) dv du - 3. \quad (18)$$

Thus, it is minimized, with respect to the parameters of the correlation function, the sum of the squared differences

between the rank correlations of the theoretical and the empirical copulas plus the squared difference of the asymmetries.

Validation of the Copula Models

The goodness of fit of a selected copula model can be assessed by comparing the sample bivariate copulas with respect to the fitted theoretical ones. For this aim, some statistical tests are available in the literature, such as the Kolmogorov-Smirnov distance D_1 or a Cramér von Mises type distance D_2 :

$$D_1 = \sup \left\{ |\hat{C}(u_1, u_2) - C(u_1, u_2)|; (u_1, u_2) \in [0, 1]^2 \right\}, \quad (19)$$

$$D_2 = \int_0^1 \int_0^1 [\hat{C}(u_1, u_2) - C(u_1, u_2)]^2 du_1 du_2. \quad (20)$$

However, performing the test statistics is not straightforward since the independence of samples cannot be assumed. Alternatively, different realizations of random functions having the same multivariate copula can be simulated, so that the assessment of the significance can be obtained by using a bootstrap framework (Efron and Tibshirani 1986). Thus, an algorithm of statistical testing based on **bootstrap**, useful to analyze the suitability of the spatial dependence models, is described hereafter. The validation procedure consists of the following steps:

1. Computation of the sample copulas from the observations on the basis of (12).
2. Parameter estimation of the vector-dependent bivariate copula by using the available sample.
3. Comparison between the sample copula and the theoretical copula, for each distance lag, through the distances D_1 and D_2 . These are indicated with the notation D_1^* and D_2^* .
4. Fixing the counter i equal to zero.
5. Simulating one realization for the multivariate copula related to the sample points. In other terms, the simulation aims to produce a realization for the defining multivariate distribution, then the corresponding multivariate copula is constructed through the formulation given in (14).
6. Calculation of the sample copulas of the simulated field.
7. Computation of the indexes $D_1(i)$ and $D_2(i)$, for these copulas.
8. Increasing the counter i by one.
9. Replication of the previous steps 5–8 until the number H of simulations is reached.
10. Ranking of the distances $D_1(i)$ and $D_2(i)$, $i = 1, 2, \dots, H$, such that $D_j(1) \leq \dots \leq D_j(H)$ with $j = 1, 2$.
11. Rejection of the copula at the level α of significance for the index D_j , if $D_j(i_0) \leq D_j^* \leq D_j(i_0 + 1)$ and $i_0 > H(1 - \alpha)$.

Note that simulating the χ^2 copula is an easy task, since after simulation of the normal $\mathbf{Y}$, then the transformation according to the equation $V_j = Y_j^2$ can be performed.

Vine Copula

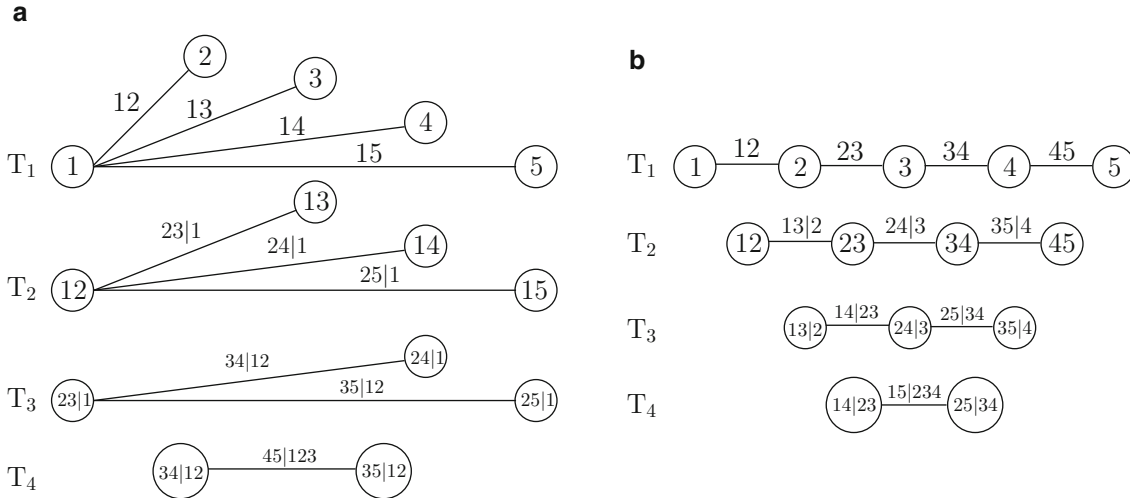
In spatial statistics, the interpolation approaches are based on different variants of kriging. However, the choice of assuming that the random field is Gaussian might be too limiting with respect to other more flexible probabilistic models. In this case, the theory of copula can be of some help. Gaussian multivariate distributions have two restrictions: elliptical symmetry and tail independence. This last means that joint extreme events are independent and this is not influenced by the covariance model, while the first limitation implies that pair of high values and pair of low values, at two spatial points, have the same level of dependence. In turn, this characteristic generates smooth interpolation when kriging is used. Thus, using the concept of copula and copula families, which are different from the Gaussian, can overcome these restrictions. The idea, known as the **pair-copula construction**, is to connect bivariate copulas to higher-dimensional vine copulas, through vines (Bedford and Cooke 2002). Since the statistical applicability of vine copulas with non-Gaussian building bivariate copulas was recognized by Aas et al. (2009), vine copulas became a standard tool to describe the dependence structure of multivariate data (Aas 2016). Thus, it is worth introducing first the concept of regular vines from Kurowicka and Cooke (2006).

Definition 1

(Regular vine) A collection of trees $V = (T_1, \dots, T_{d-1})$ is a regular vine on d elements if.

1. T_1 is a tree with nodes $N_1 = \{1, \dots, d\}$ connected by a set of non-looping edges E_1 .
2. For $i = 2, \dots, d$, T_i is a connected tree with edge set E_i and node set $N_i = E_i$, where $|N_i| = d - (i - 1)$ and $|E_i| = d - i$ are the number of edges and nodes, respectively.
3. For $i = 2, \dots, d - 1$, $\forall e = \{a, b\} \in E_i$: $|a \cap b| = 1$ two nodes $a, b \in N_i$ are connected by an edge e in T_i if the corresponding edges a and b in T_i share one node (proximity condition).

A tree $T = (N, E)$ is an acyclic graph, where N is its set of nodes and E is its set of edges. Acyclic means that there exists no path such that it cycles. In a connected tree we can reach each node from all other nodes on this tree. *R-vine* is simply a sequence of connected trees such that the edges of T_i are the nodes of T_{i+1} . A traditional example of these structures are **canonical vines** (C-vines) and **drawable vines** (D-vines) (Aas et al. 2009) in Fig. 2. Every tree of a C-vine is defined by a root node, which has d_i incoming edges, in each tree T_i ,



Copula in Earth Sciences, Fig. 2 (a) *C*-vine and (b) *D*-vine representation for $d = 5$, as given in Aas et al. (2009)

$i \in \{1, \dots, d-1\}$, whereas a *D*-vine is solely defined through its first tree, where each node has at most two incoming edges.

After this brief review of the basic notion of vine, it become easy to introduce the construction of vine copulas to be used in modeling a random field in space or spacetime. In this context, the n -dimensional multivariate distribution to be modeled is referred to the given random field evaluated only at a set of discrete points $\mathbf{x}_1, \dots, \mathbf{x}_n$ (which can be either spatial locations, temporal points, or spatio-temporal points). By following the approach of Gräler and Pebesma (2012); Gräler and Pebesma (2011), bivariate spatial copulas lead to only spatial, only temporal, or spatio-temporal multivariate copulas, differently from the other approaches which have built spatial copulas on the basis of single-family multivariate copulas (Bárdossy 2006; Kazianka and Pilz 2011). Note that a software package is also available on the R environment and is called **spcopula**.

The **pair-copula construction** is based on the idea that multivariate copulas can be approximated with nested bivariate copulas blocks. It is worth to underline that the use of the

term “approximate” is due to the consideration that not all copulas can be rebuild as a vine copula, thus although vine copulas are multivariate copulas, in some cases they can only approximate the target copula.

As well-known the decomposition of the multivariate copula into nested bivariate blocks is not unique, then different decompositions can be obtained from different regular vines, and the choice of the actual vine will depend on the goodness of fit to the multivariate target copula. For the specific spatial and spatio-temporal context, where there is a neighbourhood that can be used to define the decomposition and a central location with respect to which all initial dependencies can be fixed, the canonical vine copula is a reasonable choice. In Fig. 3, a five-dimensional canonical vine is illustrated, although the conditional cumulative distribution functions have been reported without their arguments. The density of the respective pair-copula construction is the product of all bivariate copula densities following the decomposition structure:

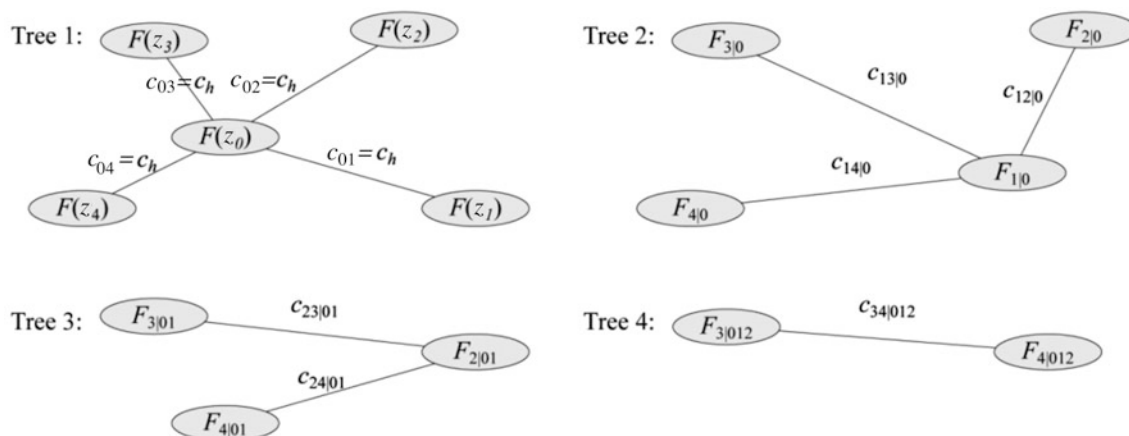
$$c(u_0, \dots, u_4) = c_{01}(u_0, u_1) \cdot c_{02}(u_0, u_2) \cdot c_{03}(u_0, u_3) \cdot c_{04}(u_0, u_4) \cdot c_{12|0}(u_1|0, u_2|0) \cdot c_{13|0}(u_1|0, u_3|0) \cdot c_{14|0}(u_1|0, u_4|0) \cdot c_{23|01}(u_2|01, u_3|01) \cdot c_{24|01}(u_2|01, u_4|01) \cdot c_{34|012}(u_3|012, u_4|012), \quad (21)$$

where the recalled conditional cumulative distribution functions $u_{k|V} = F_{k|V}(z_k)$ correspond to partial derivatives of the copulas involved, as specified below

$$u_{k|v} = \frac{\partial C_{jk|v-j}(u_{j|v-j}, u_{k|0..j-1})}{\partial u_{j|v-j}},$$

with v the set of indexes of the conditioning variables and $v-j$ the set of indexes excluding j .

It is worth recalling that the use of vine copulas is advisable not only to model **non-Gaussian random fields** Z (where the non-Gaussianity is referred to marginal distributions), but also to adopt non-Gaussian dependence structure between locations. Then, since every n -dimensional neighbourhood of a random field Z can be interpreted as n -variate distribution, the vine copulas describing these



Copula in Earth Sciences, Fig. 3 Canonical vine structure for the construction of a five-dimensional spatial random field, as given in Gräler and Pebesma (2011)

neighbourhoods need to capture changing correlations across the domain (assuming often isotropy and stationarity), in order to flexibly reflect different neighbourhoods arrangements. To this aim, Gräler (2014) proposed to use blocks of bivariate copulas, which are convex combinations of well-known bivariate copula families whose parameters depend on distance (in space or in spacetime) in the vine copula. These leads to spatial or spatio-temporal vine copulas which are distance-aware. Indeed, an earlier contribution Gräler and Pebesma (2011) highlighted the potentiality of spatial vine copulas for heavily skewed spatial random fields and a first attempt to extend the spatial approach to the spatio-temporal one was given in Gräler and Pebesma (2012). In this way, modeling the dependence structure, in terms of strength and shape, between points in space or in space-time is very flexible. Note that the bivariate space or space-time copulas on the first tree essentially recall a **Gumbel copula**, since it allows for a strong dependence in the tails, differently from the Gaussian case. Moreover, considering marginals together to this vine copula makes a local multivariate probabilistic model of the random field available for prediction at unsampled points. Unlike kriging where each predictive distribution is always Gaussian, the conditional distributions of the covariate vine copula can assume any functional form and then can provide more realistic uncertainty estimates. However, modeling separately marginals and dependence structure offers a wide range of alternatives in the definition of a random field's distribution. Nevertheless, it is crucial to recover reliable models for both components in order to guarantee good results. Simulation from the modeled random function is also practicable and can be performed by using the same package **spcopula**.

For further details on bivariate spatial copulas, where distance is incorporated as parameter, together with the use of many bivariate copula families, the readers can refer to Gräler (2014).

Conclusions

Nowadays, there are wide areas of research that involves relevant theoretical and computational challenges to face interpolation or prediction problems as well as applications in various scientific fields. In particular, some significant lines of research can be found in the construction of new class of models for spatial and spatio-temporal data, as well as in the analysis of big spatial and spatio-temporal data sets or in machine learning methods developed by the computational science researchers.

Cross-References

- [Markov Random Fields](#)
- [Random Function](#)
- [Variogram](#)

Bibliography

- Aas K (2016) Pair-copula constructions for financial applications: a review. *Econometrics* 4:43
- Aas K, Czado C, Frigessi A, Bakken H (2009) Pair-copula constructions of multiple dependence. *Insurance* 44:182–198
- Bárdossy A (2006) Copula-based geostatistical models for groundwater quality parameters. *Water Resour Res* 42:W11416. (1–12)
- Bedford T, Cooke RM (2002) Vines – a new graphical model for dependent random variables. *Ann Stat* 30(4):1031–1068
- De Michele C, Salvadori G (2003) A generalized Pareto intensity-duration model of storm rainfall exploiting 2-Copulas. *J Geophys Res* 108(D2):4067
- Drouot-Mari D, Kotz S (2001) Correlation and dependence. Imperial College Press, London
- Efron B, Tibshirani R (1986) Bootstrap method for standard errors, confidence intervals and other measures of statistical accuracy. *Stat Sci* 1:54–77

- Embrechts PME, McNeil AJ, Straumann D (2002) Correlation and dependency in risk management: properties and pitfalls. In: Dempster M (ed) Risk management: value at risk and beyond. Cambridge University Press, Cambridge, UK, pp 176–223
- Favre A-C, El Adlouni S, Perreault L, Thiérmonge N, Bobée B (2004) Multivariate hydrological frequency analysis using copulas. *Water Resour Res* 40:W01101
- Genest C, MacKay R (1986) Archimedean copulas and families of bidimensional laws for which the marginals are given. *Can J Stat* 14:145–159
- Gomez-Hernandez J, Wen X (1998) To be or not to be multi-Gaussian? A reflection on stochastic hydrogeology. *Adv Water Resour* 21:47–61
- Gräler B (2014) Modelling skewed spatial random fields through the spatial vine copula. *Spat Stat* 10:87–102
- Gräler B, Pebesma EJ (2011) The pair-copula construction for spatial data: a new approach to model spatial dependency. *Procedia Environ Sci* 7:206–211
- Gräler B, Pebesma EJ (2012) Modelling dependence in space and time with Vine Copulas, expanded abstract collection from ninth international geostatistics congress, Oslo, Norway, June 11–15, 2012. International Geostatistics Congress. <http://geostats2012.nr.no/1742830.html>
- Joe H (1997) Multivariate models and dependence concepts. CRC Press, Boca Raton, Fla
- Journel AG, Alabert F (1989) Non-Gaussian data expansion in the earth sciences. *Terra Nova* 1:123–134
- Journel AG, Deutsch CV (1996) Rank order geostatistics: a proposal for a unique coding and common processing of diverse data. In: Baafi EY, Schofield NA (eds) Geostatistics Wollongong 96. Springer, New York, pp 174–187
- Kazianka H, Pilz J (2011) Copula-based geostatistical modeling of continuous and discrete data including covariates. *Stoch Environ Res Risk Assess* 24(5):661–673
- Kurowicka D, Cooke R (2006) Uncertainty analysis with high dimensional dependence modelling. John Wiley & Sons, New York
- Matheron G (1971) The theory of regionalized variables and its applications. École des Mines de Paris, Paris
- Nelsen R (1999) An introduction to Copulas. Springer, New York
- Rueschendorf L (1985) Construction of multivariate distributions with given marginals. *Ann Inst Stat Math* 37:225–233
- Sklar A (1959) Fonctions de répartition à n dimensions et leur marges. *Publ Inst Stat Paris* 8:131–229

Correlation and Scaling

Frits Agterberg
Central Canada Division, Geomathematics Section,
Geological Survey of Canada, Ottawa, ON, Canada

Definition

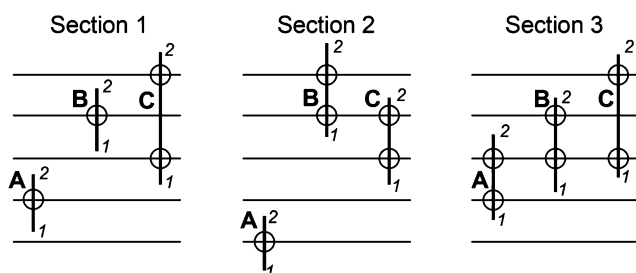
Correlation and scaling (or ranking and scaling, in some publications abbreviated to rascing) is a technique to order stratigraphic events, observed at different locations in a large study region, along a linear scale with variable distances between them. The stratigraphic events considered in applications are mostly biostratigraphic, primarily first or last occurrences of fossil species, but some lithostratigraphic

events can be included as well. These other events may be ash layers, gamma-ray signatures, or other lithostratigraphic events that can be found in different stratigraphic sections across the entire study region considered, which is usually a sedimentary basin.

Correlation Between Sections in Stratigraphy

Correlation of strata on the basis of their fossil content is one of the earliest scientific methods applied in geology. As documented by Winchester (2001), the first geological map was published in 1815 by William Smith under the title: “A Delineation of the Strata of England and Wales, with Parts of Scotland.” It was based on his discovery that strata could be correlated on the basis of their fossil content that changed with time. Shortly afterward, Charles Lyell (1833) in the first edition of his textbook *Principles of Geology* subdivided the Tertiary into different stages by counting how many fossil species they contain are still living today.

Large uncertainties are commonly associated with the positioning of events in biostratigraphic sections. This subject is treated in more detail in the chapter on quantitative biostratigraphy. Here the emphasis is on how scaled optimum sequences can be used for correlation between sections. An artificial example to illustrate uncertainty associated with observed data is shown in Fig. 1. An early statistical method in quantitative stratigraphy was developed by Shaw (1964) for use in hydrocarbon exploration. In this book, Shaw illustrates his technique of “graphic correlation” on first and last occurrences (FOs and LOs) of trilobites in the Cambrian Riley Formation of Texas for which range charts had been published by Palmer (1954). Shaw’s method consists of constructing a “line of correlation” on a scatter plot showing the FOs and LOs of taxa in two sections. If quality of information is better in one section, its distance scale is made horizontal. FOs stratigraphically above and LOs below the line of correlation are moved horizontally or vertically toward the line of correlation in such a way that the ranges of the taxa become longer. The reason for this procedure is that it can be assumed that observed lowest occurrences of fossil taxa generally occur above truly highest occurrences and the opposite rule applies to highest occurrences. Consequently, if the range of a taxon is observed to be longer in one section than in the other, the longer observed range is accepted as the better approximation. The objective is to find approximate locations of what are truly the first appearance datum (or FAD) and last appearance datum (LAD) for each of the taxa considered. True FADs and LADs probably remain unknown, but, by combining FOs and LOs from many stratigraphic sections, approximate FADs and LADs are obtained and the intervals between them can be plotted on a range chart.



Correlation and Scaling, Fig. 1 This artificial example shows three stratigraphic sections in which three fossil taxa (a, b, and c) were observed to occur. Each taxon has first and last occurrence (FO and LO) labeled 1 and 2, respectively. Consequently, there are six stratigraphic events in total. The equidistant horizontal lines represent five regularly spaced sampling levels. The discrete sampling procedure changes the positions of the FOs and LOs. Any observed range for a taxon generally is much shorter than its true range of occurrence. In this example, the observed ranges become even shorter because of the discrete sampling. FOs and LOs coincide in three places. In reality, the FO of a taxon always must occur stratigraphically below its LO. It is because of the sampling scheme that these two events can occur at exactly the same level in this example (see circles in Fig. 1). For comparison, coincident FO and LO would occur in land-based studies when only a single fossil for a taxon would be observed in a stratigraphic section. In this artificial example, it is assumed that all three taxa occur in each section. In practical applications, many taxa generally are missing from many sections

Two sections usually produce a crude approximation, but a third section can be plotted against the combination of the first two, and new inconsistencies can be eliminated as before. The process is repeated until all sections have been used. The final result or “composite standard” contains extended ranges for all taxa. Software packages in which graphic correlation has been implemented include GraphCor (Hood 1995) and STRATCOR (Gradstein 1996). The method can be adapted for constructing lines of correlation between sections (Shaw 1964). Various modifications of Shaw’s graphic correlation technique with applications in hydrocarbon exploration and to land-based sections can be found in Mann and Lane (1995). An example of Shaw’s original method will be given later in this chapter for comparison with two other methods of biostratigraphic correlation.

In another type of approach, Hay (1972) used stratigraphic information on calcareous nannofossils from sections in the California Coast Ranges as an example of application of his method of ordering biostratigraphic events. Hay’s objective was to construct a so-called optimum sequence of FADs and LADs. His example will be discussed in more detail in the next section, because it inspired the “rascing” approach for ranking and scaling later advocated by Gradstein and Agterberg (1982, also see Gradstein et al. 1985) to solve the same problem. In rascing, Hay’s optimum sequence is followed by scaling that consists of estimating intervals of

variable length between the successive events. Scaling can be regarded as a refinement of ranking. The computer program RASC (for latest version, see Agterberg et al. 2013) performs rascing calculations followed by construction of lines of correlation between sections with graphical displays of results.

In their chapter on quantitative biostratigraphy, Hammer and Harper (2005) present separate sections on five methods of quantitative biostratigraphy: (1) graphic correlation, (2) constrained optimization, (3) ranking and scaling (RASC), (4) unitary associations, and (5) biostratigraphy by ordination. Theory underlying each method is summarized by these authors, and worked-out examples are provided. Their book on paleontological data analysis is accompanied by the free software package PAST (available through <http://www.blackwellpublishing.com/hammer>) that has been under continuous development since 1998. It contains simplified versions of CONOP for constrained optimization (Sadler 2004) and RASC, as well as a comprehensive version for unitary associations (Guex 1980), a method that puts much weight on observed co-occurrences of fossil taxa in time. For comparison of RASC and unitary associations output for a practical example, see Agterberg (1990).

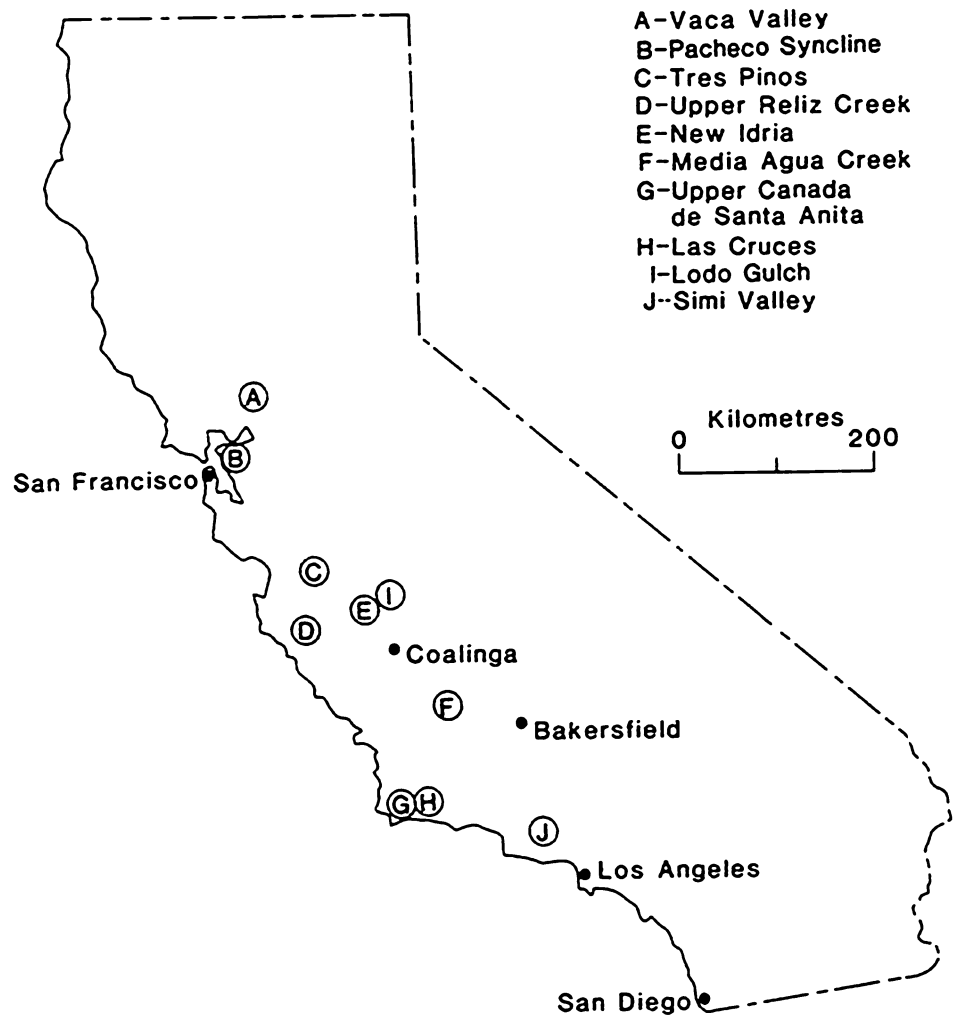
Examples of Ranking, Scaling, and Correlation

The first example to be discussed is concerned with sequencing of stratigraphic events. Nannoplankton faunozones were sampled in nine sections across California of which three are shown in Fig. 2 (Hay 1972). Stratigraphic information consisting of nine highest occurrences (HI) and one lowest occurrence (LO) of nannofossils was extracted from these sections labeled A-I in Fig. 3. The columns on the right represent a subjective ordering of the ten events and Hay’s original optimum sequence that is an objective ordering based on frequencies of how many times every event occurs above (or below) all other events considered.

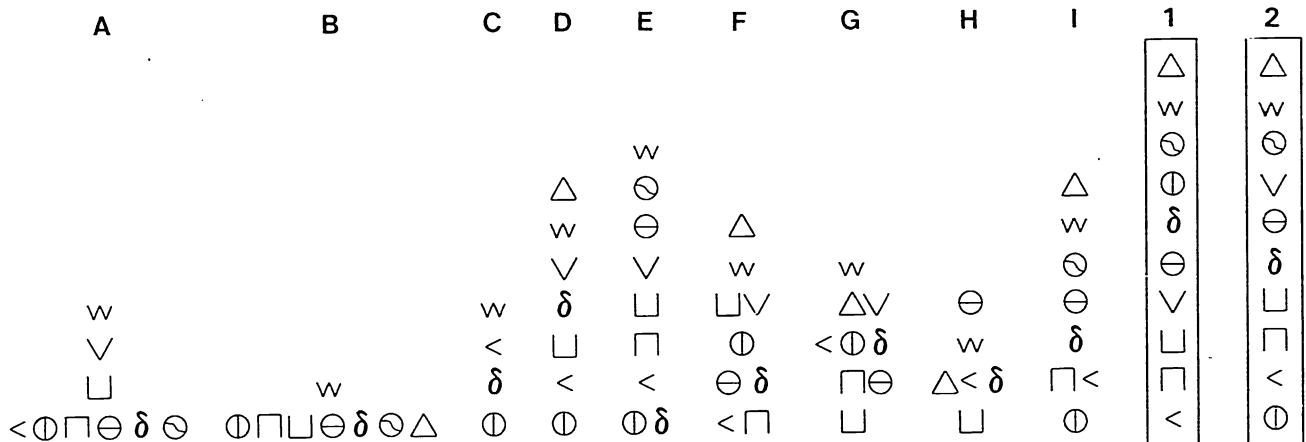
Agterberg and Gradstein (1988) took Hay’s approach one step further by estimating intervals between successive events in the optimum sequence. Each frequency f_{ij} was converted into a relative frequency $p_{ij} = f_{ij}/n_{ij}$ where n_{ij} is sample size. This relative frequency was changed into a “distance” D_{ij} by means of the probit transformation: $D_{ij} = \Phi^{-1}(p_{ij})$ (cf. Agterberg 2014). An example of this transformation is as follows. If two events turn out to be coeval in the optimum sequence, their interevent distance in the scaled optimum sequence becomes 0 as it should be. In Fig. 4, the optimum sequence for Hay’s original example is compared with the scaled optimum sequence obtained after subjecting all

Correlation and Scaling,

Fig. 2 Locations of sections in the Sullivan (1965, Table 6) database for the Eocene used by Hay (1972) for example. (Source: Agterberg 1990, Fig. 4.1)

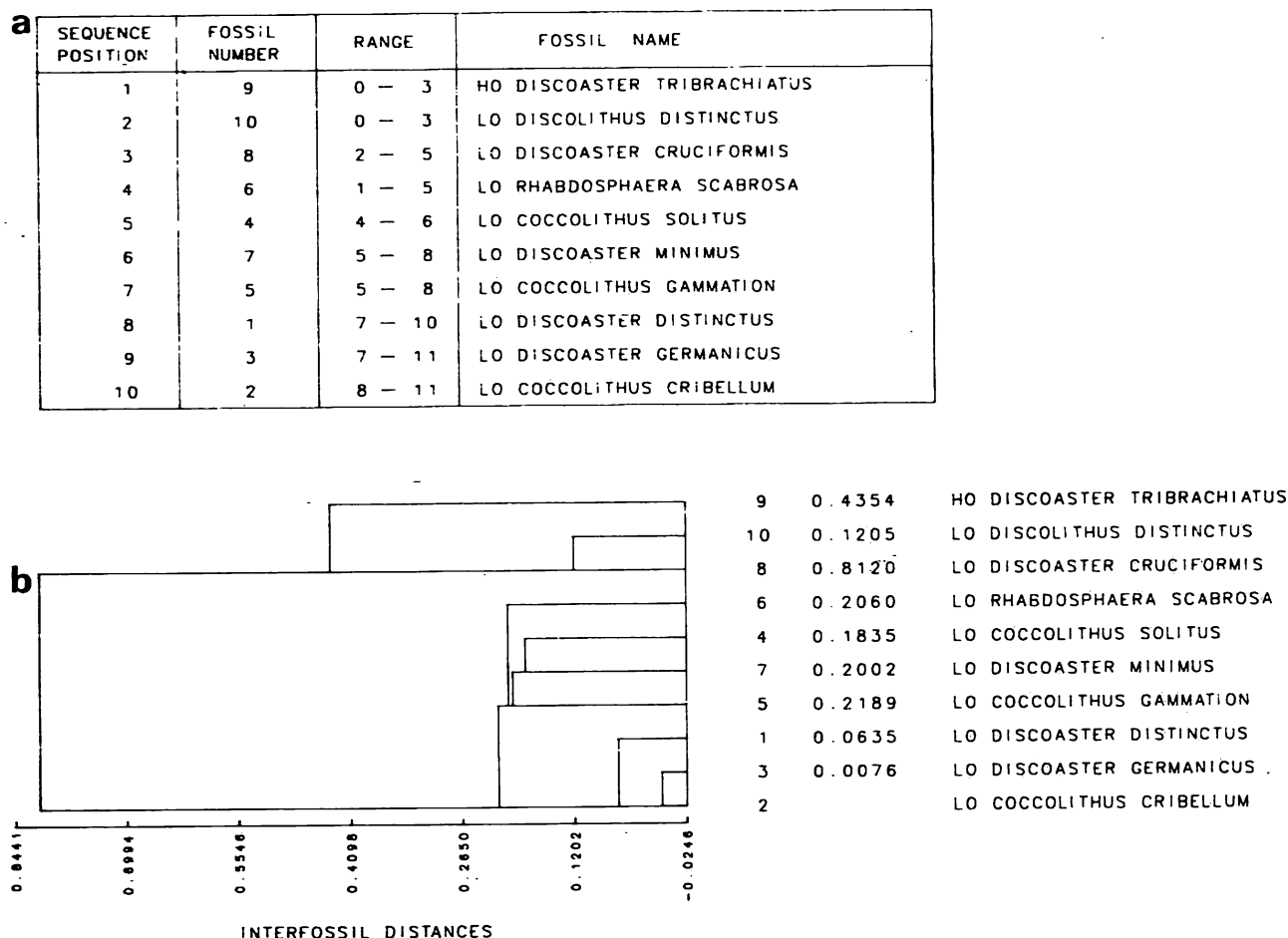


STRATIGRAPHIC INFORMATION



Correlation and Scaling, Fig. 3 Hay's (1972) example. One last occurrence and nine "first" occurrences of Eocene nannofossils selected by Hay (1972) from the Sullivan Eocene database. Explanation of symbols: δ = LO, *Coccolithus gammatum*; Φ = LO, *Coccolithus eribellum*; Θ = LO, *Coccolithus solutus*; V = LO, *Discoaster cruciformis*; $<$ = LO, *Discoaster distinctus*; Π = LO, *Discoaster*

germanicus; U = LO, *Discoaster minimus*; W = HI, *Discoaster tribrachiatus*; Δ = LO, *Discolithus distinctus*; □ = LO, *Rhabdosphaera scabrosa*. See Fig. 2 for locations of the nine sections (A-I). Some LOs are for nannofossils also found in Paleocene (Sullivan 1964, Table 3). The columns on the right represent subjective ordering of the events and Hay's original optimum sequence. (Source: Agterberg 1990, Fig. 4.2)



Correlation and Scaling, Fig. 4 RASC results for Hay example of Fig. 3 (after Agterberg and Gradstein 1988); (a) ranked optimum sequence; (b) scaled optimum sequence. Clustering of events 1 to 7 in

the dendrogram (b) reflects the relatively large number of two-event inconsistencies and many coincident events near the base of most sections used. (Source: Agterberg 1990, Fig. 6.3)

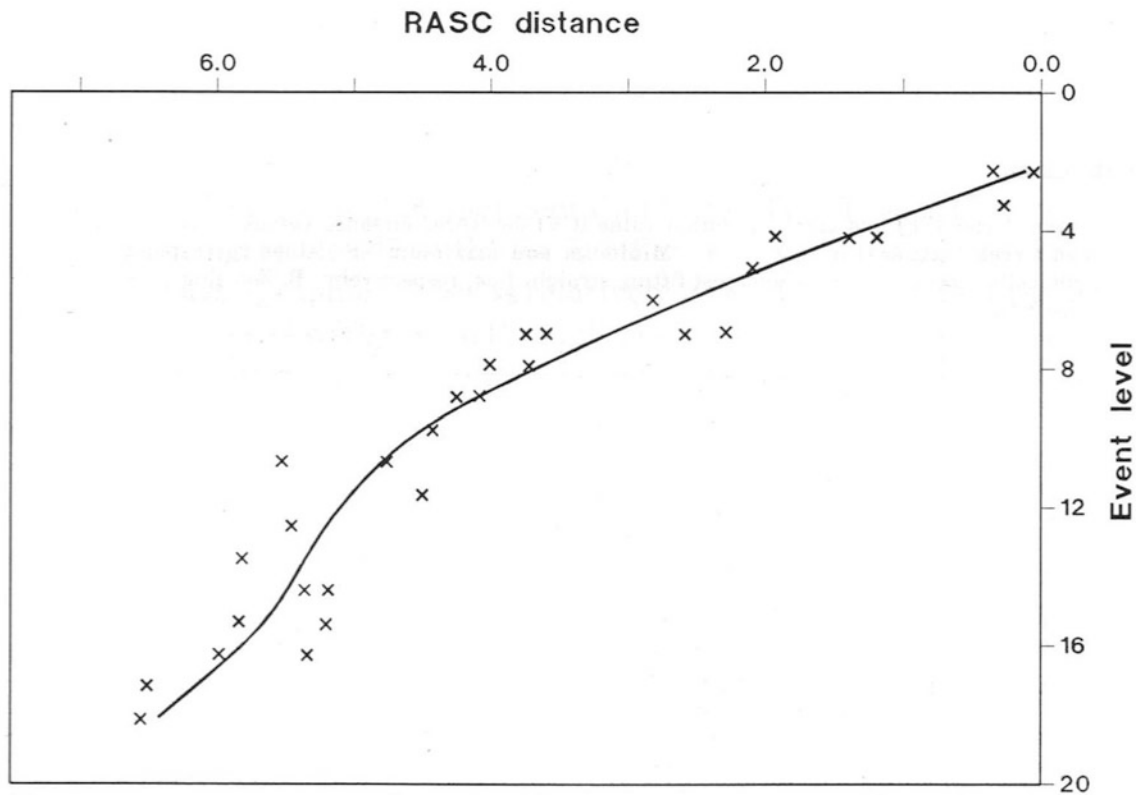
superpositional frequencies to the probit transformation and representing the result in the form of a dendrogram.

The approach can be taken a further step forward. In Fig. 5, which is for a single section in a larger database, scaled optimum sequence points are plotted against their depths, and a cubic spline curve was fitted to the data points. The points on the curve can be regarded as estimates of expected positions of the transformed relative frequencies. These expected positions of events in different sections can be used for correlation between sections.

Figure 5 is for one of seven sections used by Shaw (1964) who illustrated his original technique of “graphic correlation” on first and last occurrences of trilobites in the Cambrian Riley Formation of Texas for which range charts had been published by Palmer (1954). Shaw (1964) extended his composite standard method to construct correlation lines between sections. Figure 6 shows correlation results for three sections

in the Riley Formation obtained by three different methods: (1) Palmer’s original biozones that were obtained by conventional subjective paleontological reasoning, (2) Shaw’s so-called Riley Composite Standard (SRC) correlation lines, and (3) Agterberg’s (1990) correlation and scaling (CASC) results in which the expected positions of events were connected by lines of correlation. Error bars in Fig. 5 are projections of single standard deviations on either side of probable positions of events obtained by the CASC method. Clearly, uncertainty in positioning of the correlation lines increases rapidly in the stratigraphically downward direction (for further explanation, see Agterberg 1990). Similar results were obtained by the three different methods used.

In another application to a large database (see Fig. 7), the estimated RASC optimum sequence locations are compared with maximum deviations between observed and best-fitting spline curve for 44 offshore wells drilled along the

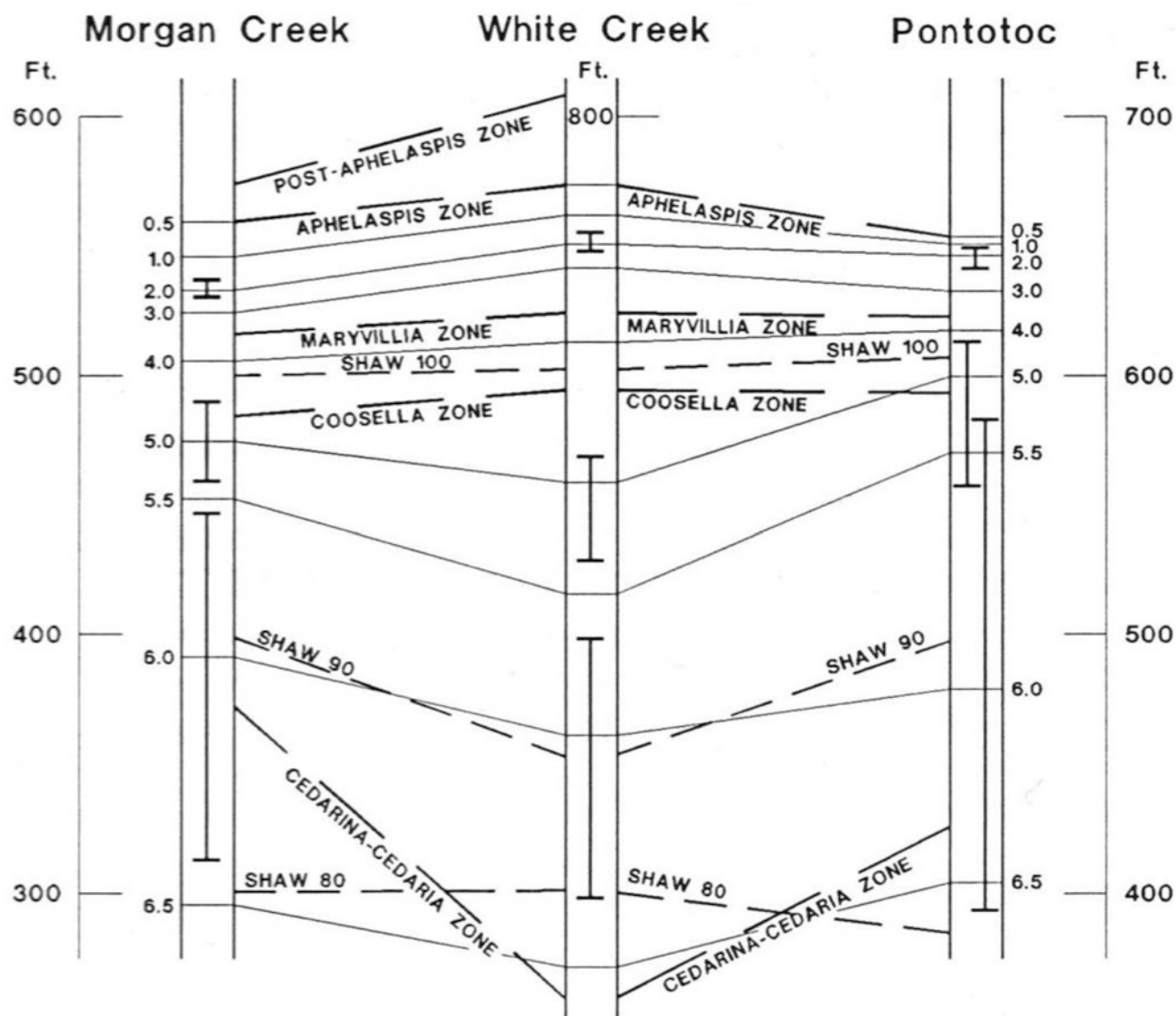


Correlation and Scaling, Fig. 5 RASC distance-event level plot for Morgan Creek section. Spline curve is for optimum (cross-validation) smoothing factor $SF = 0.382$. (Source: Agterberg 1990, Fig. 9.23)

northwestern Atlantic margin and on the Grand Banks. This example illustrates that the optimum sequence value for a stratigraphic event is the average value of a frequency distribution of observed occurrences (in this example, LOs only). Using the terminology of Shaw, the corresponding LAD occurs at a stratigraphically later position equivalent to about 9 million years in this application. In a later hydrocarbon application, Agterberg and Liu (2008) showed that the LOs for 27 Labrador and northern Grand Banks wells satisfy a bilateral gamma distribution that plots as a straight line on a normal $Q-Q$ plot when the depths of the LOs in any well are subjected to a square root transformation (see Fig. 8). The straight line fitted in Fig. 8 was fitted to the data points shown excluding those on the anomalous bulge near the center of the plot that reflects the discrete sampling method used when the wells were being drilled. For a more detailed explanation, see Agterberg et al. (2007) and Agterberg and Liu (2008).

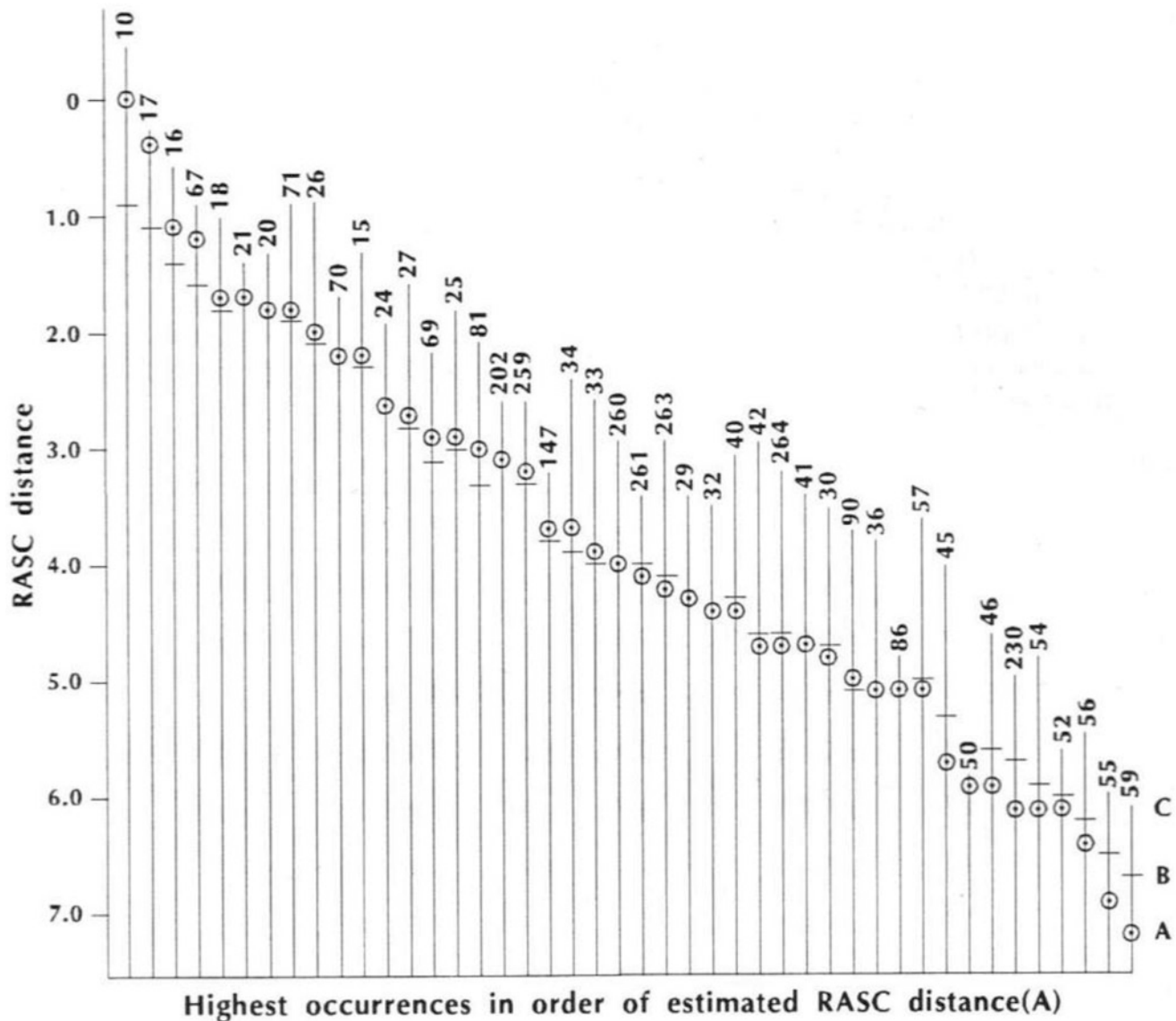
Summary and Conclusions

First and last occurrences (FOs and LOs) of fossil species in a region are subject to relatively large uncertainties. In general, FOs are observed above their truly first occurrences, and LOs are below their truly last occurrences because the period of time that any fossil species was in existence generally remains unknown. Various methods of quantitative stratigraphy have been developed to reduce uncertainties related to true period of existence, in order to allow regional correlations between estimated positions of stratigraphic events in sections at different locations in a region. The main technique illustrated in this chapter is ranking and scaling (rascing) of observed stratigraphic events in different sections in a region or in deep boreholes in a basin drilled for hydrocarbon exploration. Estimated average FOs and LOs can be used for correlation between different stratigraphic sections. Uniquely identifiable lithostratigraphic events such as ash layers resulting from the same volcanic eruption can be used along with the biostratigraphic information.



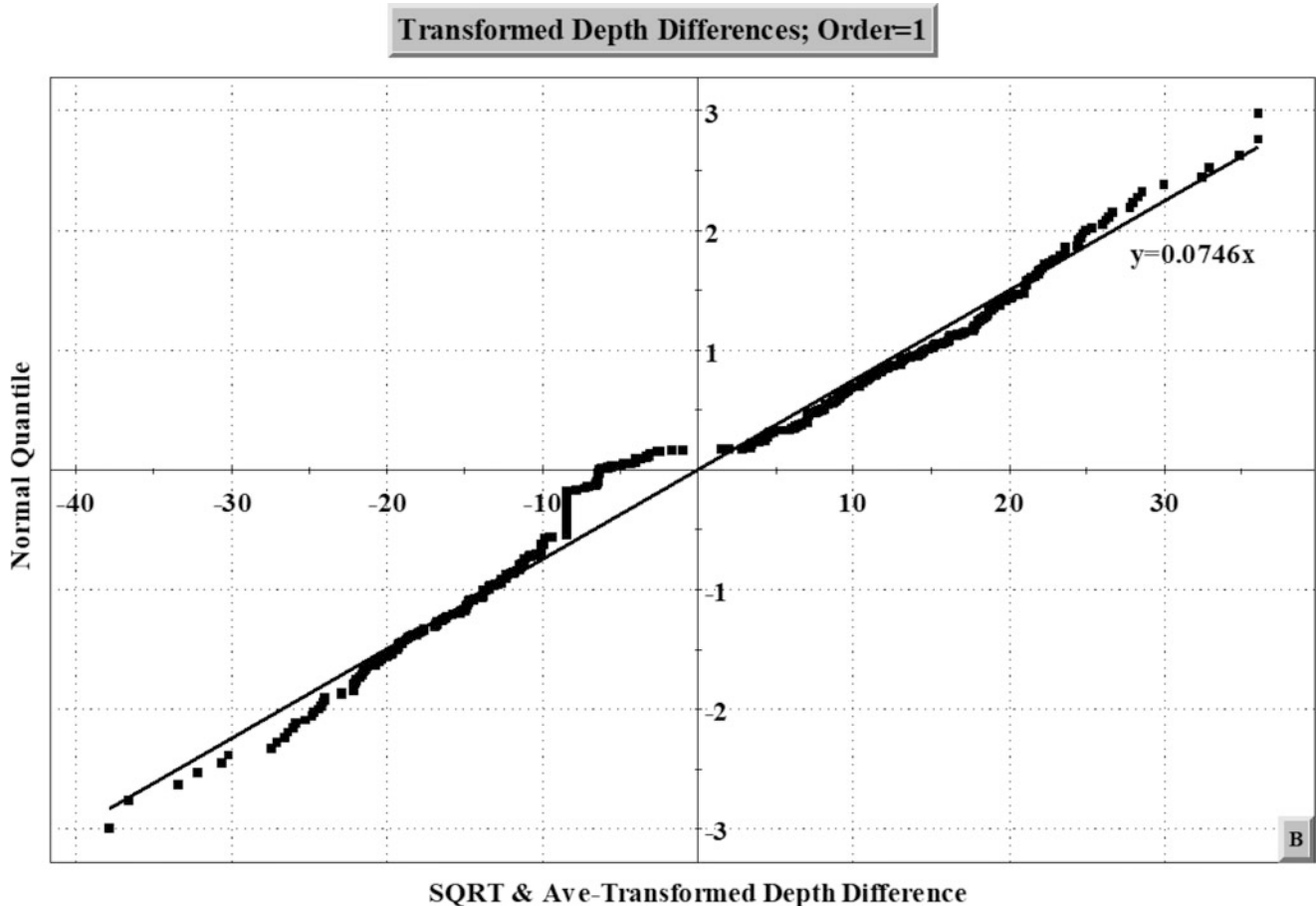
Correlation and Scaling, Fig. 6 Biostratigraphic correlation of three sections for the Riley Formation in central Texas by means of three methods. Palmer's (1955) original biozones and Shaw's (1964) R.T.S. value correlation lines were superimposed on CASC (Agterberg 1990) results. Error bars are projections of single standard deviation on either side of probable positions for cumulative RASC distance values

equal to 2.0, 5.0, and 6.0, respectively. A cumulative RASC distance value is the sum of interevent distances between events at stratigraphically higher levels. Uncertainty in positioning the correlation lines increases rapidly in the stratigraphically downward direction. The example shows that the three different methods used produce similar correlations. (Source: Agterberg 1990, Fig. 9.25)



Correlation and Scaling, Fig. 7 Extended RASC ranges for Cenozoic Foraminifera in so-called Gradstein-Thomas database for offshore wells, northwestern Atlantic margin. Letters for taxon 59 on the right represent (a) estimated RASC distance, (b) mean deviation from spline curve, and (c) highest occurrence of species (i.e., maximum for deviation from spline curve). B is shown only if it differs from A. Good “markers”

such as highest occurrence of taxon 50 (*Subbotina patagonica*) have approximately coinciding positions for A, B, and C. Note that as a first approximation it could be assumed that the highest occurrences (c) have RASC distances which are about 1.16 units less than the average position. Such systematic difference in RASC distance is equivalent to approximately 10 million years. (Source: Agterberg 1990, Fig. 8.9)



Correlation and Scaling, Fig. 8 Normal Q - Q plot of first-order square root transformed depth differences for dataset (b). Approximate normality is demonstrated except for anomalous upward bulge in the center of

the plot, which is caused by the use of discrete sampling interval. (Source: Agterberg et al. 2013, Fig. 8b)

Cross-References

- [Frequency Distribution](#)
- [Quantitative Stratigraphy](#)
- [Smoothing Filter](#)

Bibliography

- Agterberg FP (1990) Automated stratigraphic correlation. Elsevier, Amsterdam. 424 pp
- Agterberg FP (2014) Geomathematics: theoretical foundations, applications and future developments. Springer, Heidelberg. 553 pp
- Agterberg FP, Gradstein FM (1988) Recent developments in stratigraphic correlation. *Earth-Sci Rev* 25:1–73
- Agterberg FP, Liu G (2008) Use of quantitative stratigraphic correlation in hydrocarbon exploration. *Int J Oil Gas Coal Expl* 1(4):357–381
- Agterberg FP, Gradstein FM, Liu G (2007) Modeling the frequency distribution of biostratigraphic interval zone thicknesses in sedimentary basins. *Nat Resour Res* 16(3):219–233
- Agterberg FP, Gradstein FM, Cheng Q, Liu G (2013) The RASC and CASC programs for ranking, scaling and correlation of biostratigraphic events. *Comput Geosci* 54:279–292
- Gradstein FM (1996) STRATCOR – graphic zonation and correlation software – user's guide, Version 4
- Gradstein FM, Agterberg FP (1982) Models of Cenozoic foraminiferal stratigraphy – northwestern Atlantic margin. In: Cubitt JM, Reymont RA (eds) Quantitative stratigraphic correlation. Wiley, Chichester, pp 119–173
- Gradstein FM, Agterberg FP, Brower JC, Schwarzacher WS (1985) Quantitative stratigraphy. UNESCO/Reidel, Paris/Dordrecht. 598 pp
- Gradstein FM, Kaminski MA, Agterberg FP (1999) Biostratigraphy and paleoceanography of the cretaceous seaway between Norway and Greenland. *Earth-Sci Rev* 46:27–98. (erratum 50:135–136)
- Guex J (1980) Calcul, caractérisation et identification des associations unitaires en biochronologie. *Bull Soc Vaud Sci Nat* 75:111–126
- Hammer O, Harper D (2005) Paleontological data analysis. Blackwell, Ames. 351 pp
- Hay WW (1972) Probabilistic stratigraphy. *Ecl Helv* 75:255–266
- Hood KC (1995) GraphCor – interactive graphic correlation software, Version 2.2
- Lyell C (1833) Principles of geology. Murray, London
- Mann KO, Lane HR (1995) Graphic correlation, vol 53. SEPM (Soc Sed Geol), Tulsa
- Palmer AR (1954) The faunas of the Riley formation in Central Texas. *J Paleolimnol* 28:709–788
- Sadler PM (2004) Quantitative biostratigraphy achieving finer resolution in global correlation. *Annu Rev Earth Planet Sci* 32:187–231

- Shaw AG (1964) *Time in stratigraphy*. McGraw-Hill, New York. 365 pp
- Sullivan FR (1964) Lower tertiary nannoplankton from the California Coast ranges, I. Paleocene. University of California Press, Berkeley
- Sullivan FR (1965) Lower tertiary nannoplankton from the California Coast ranges, II. Eocene. University of California Press, Berkeley
- Winchester S (2001) *The map that changed the world*. Viking/Penguin Books, London. 329 pp

Correlation Coefficient

Guocheng Pan

Hanking Industrial Group Limited, Shenyang, Liaoning, China

Definition

Correlation is a term that describes the strength of dependency between two or more variables, which can be continuous, categorical, or discrete. Correlation coefficient is a measure of the degree to which the values of these variables vary in a relevant manner. The measure is generally defined as linear relationships between variables, but it can be readily generalized to take into account other relationships of more complex forms. Correlation analysis provides a means of drawing inferences about the strength of the relationship between variables. It can be more powerful when used jointly with other statistical tools for data analysis and pattern recognitions.

Introduction

The correlation coefficient is a statistical measure of the dependence or association of a pair of variables. When two sets of values move in the same direction, they are said to have a positive correlation. When they move in the opposite directions, they are said to have a negative correlation. The correlation coefficient is used as a measure of the goodness of fit between the prediction equation and the data sample used to derive the prediction equation.

Correlation analysis answers if a linear association between a pair of variables exists. A correlation relationship between variables can be judged either by quantitative methods or through graphical observations. Correlation analysis is meaningful only when there exists a relationship between variables. The method can be parametric or nonparametric depending on the form or style of data that are used in the statistical analysis (Pan and Harris 1990).

In a broad sense, correlation analysis includes the description of the mathematical form of correlation, such as fitting a regression equation, testing the rationality of regression equation, and applying regression model for statistical prediction. Correlation analysis also includes the significance test of

statistical correlation between variables. Correlation analyses are most often performed after an equation relating one variable to another has been developed (Ayyub and McCuen 2011).

Correlation analysis is an extensively used technique that identifies intrinsic relationships in geoscientific, social-economic, and biomedical fields among others. It is also used broadly in machine learning and networking in recent years to identify the relevance of attributes with respect to the target class to be predicted in environmental study (Kumar and Chong 2018).

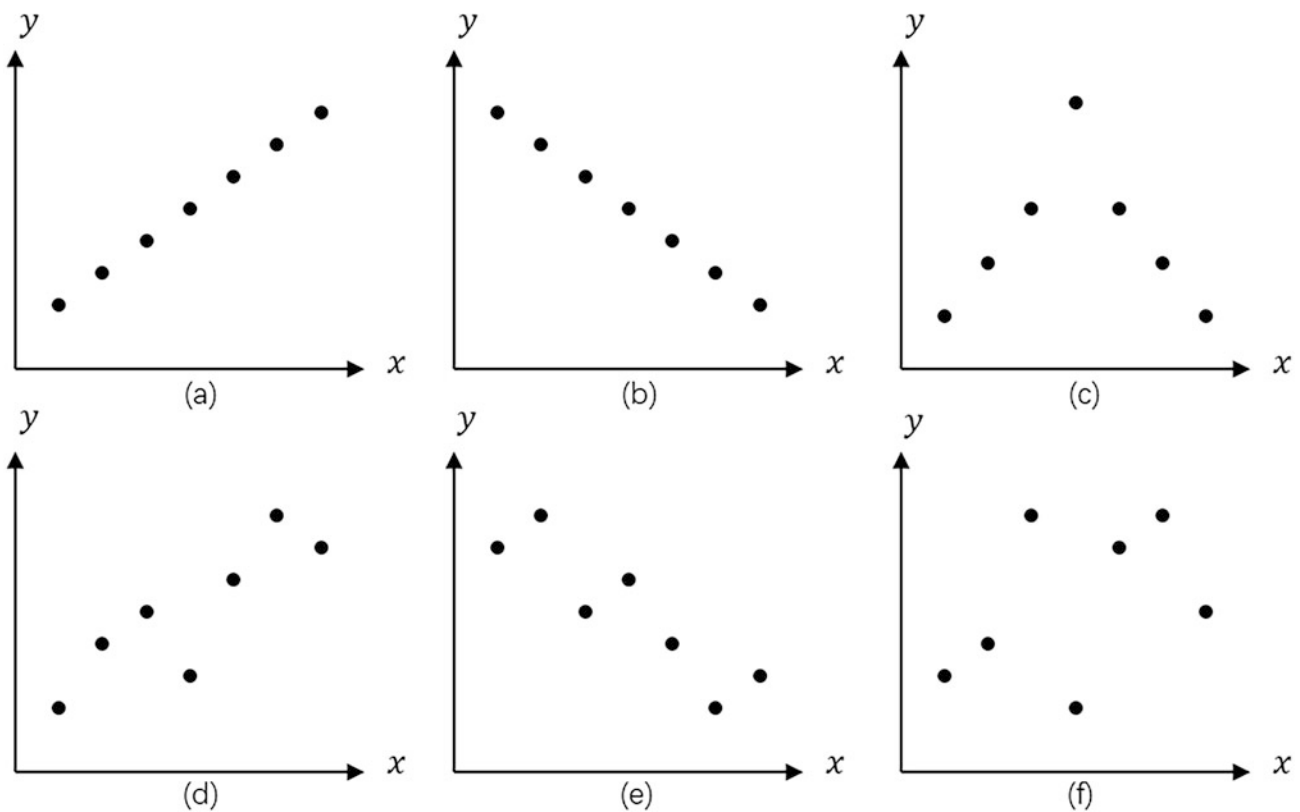
A formal correlation analysis can be carried out as follows: inspect scatter graphs of variables, select an appropriate mathematical method, calculate correlation coefficient, determine the degree of correlation between the variables, test the statistical significance of correlation, and judge the nature of correlation between variables.

It is also important to understand what correlation analysis cannot do. Correlation analysis does not provide an equation for predicting the value of a variable. It does not indicate whether a relationship is causal or physical. Correlation analysis only indicates whether the degree of common variation is statistically significant.

Scatter Graph

Graphical analysis is the first effective step in examining the relationship between variables. The relationship between a pair of variables can be visualized through plotting of scatter graphs from a set of sample values. A useful intuitive indication of correlation is the extent of the point cloud around the best-fit line when the data are plotted on a scatter graph. Positive correlation is indicated by a clear pattern of points running from down left to up right, while negative correlation is shown by an opposite pattern.

As an illustration, various degrees of variation between two variables (X and Y) are shown in Fig. 1 using a set of artificial sample data (observations) for the variables. Figure 1a represents an extreme case of perfect positive correlation, while Fig. 1b sketches the extreme case of perfect negative correlation. In Fig. 1c there is no correlation at all between the two variables. In Figs. 1d and e, the degree of correlation is moderate to strong, since in these figures, Y generally increases (decreases) as X increases (decreases) though the corresponding changes are not exactly proportional for all the sample values involved. Figure 1f provides an example of weak correlation between the two variables, implying that Y cannot be predicted by X in a satisfactory accuracy.



Correlation Coefficient, Fig. 1 Various degrees of correlation between the variables x and y : (a) perfect positive correlation; (b) perfect negative correlation; (c) zero correlation; (d) strong positive correlation; (e) strong negative correlation; (f) weak correlation

Random Variable

A random variable, either continuous or discrete, is a variable whose value is unknown or a function that assigns values to each of an experiment's outcomes (observations). A random variable has a probability distribution that represents the likelihood that any of the possible values would occur. In probability and statistics, random variables are used to quantify outcomes of a random occurrence or event and define probability distributions of the event. Random variables are required to be measurable and are typically real numbers.

Continuous random variables can represent any value within a specified range or interval and can take on an infinite number of possible values. An example of a continuous random variable is an experiment that involves measuring the grade of gold in a mineral deposit. The random variable can be represented by a capital letter, such as Z (in unit of gram per ton, g/t). Any gold grade value (for example, from drilling in the deposit), usually denoted by lower case letter z_i , is considered as an outcome of this random variable. The gold deposit is called the population of random variable Z , while the collection of gold grades from drillholes and other sampling methods is called a sample of variable Z .

Discrete random variables take on a countable number and categorical assignment of distinct values, which can be ranks of data values, type of a physical object, etc. In mineral exploration, geologists often consider many quantitative geological measurements as discrete geological phenomena. For instance, the size of ore deposits may be expressed in terms of qualitative categories, such as large, medium, and small deposits. Another example is conversion of a continuous variable to a discrete variable through optimum discretization techniques (Pan and Harris 1990).

Population and Sample

It is necessary to introduce some basic statistical concepts including population and sample, prior to discussion of the correlation and other statistical analysis for random variables. Population is the parent of a sample.

Population refers to the collection of all elements possessing common characteristics and sharing certain properties, while the sample is a finite subset of the population chosen by a sampling process. Population can be finite if its total number of elements (N) is finite; otherwise it is considered infinite. A sample is part of population chosen at random

capable of representing the population in all its characteristics or properties. The major objective of a sample is to make statistical inferences about the population.

As an example, a gold deposit can be viewed as a population of gold grade (random variable), which includes all possible sample values within the deposit. In reality, it is almost impossible to analyze gold grade values (outcomes) of all sample points within the deposit. Instead, we often obtain a limited set of sample values that are designed to reasonably represent the distribution of gold grade within the deposit, for example, through a grid drilling. Then the statistical nature of gold grade in the deposit can be estimated by analyzing a sample of the population.

Statistical characteristic of a population, such as a mean and standard deviation, are called parameters; while a measurable characteristic of a sample is called a statistic. The mean and standard deviation of a population are usually denoted by the symbols μ and σ , while the mean and standard deviation of a sample are represented by the symbols $\bar{x}$ and s .

Accuracy of a sample estimation on the statistical characteristics of a population depends on sample size and sampling method. The greater the size of a sample, the higher is the level of accuracy in representation of the population. Sampling is a procedure for selecting sample elements from a population. A common method is called simple random sampling, which satisfies the following conditions: (1) the population consists of a finite number N elements, (2) each element in population cannot be sampled in duplication, and (3) all possible sample observations are equally likely to occur. Given a simple random sample, conventional statistical methods can be employed to define various statistical characteristics of a random variable, such as a confidence interval around a sample mean of gold grade in a mineral deposit.

Basic Statistics

The most commonly used statistic of a random variable x is the sample mean, which is a central tendency descriptor. For n observations of a given sample, the average value is estimated by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (1)$$

where x_i is a sample point and $i = 1, 2, \dots, n$.

Variance or standard deviation is another important statistic that measures dispersion around the mean. The population variance of x is defined in terms of the population mean μ and population size N :

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \quad (2)$$

where σ is called standard deviation of the population.

In practice, a statistical parameter of population is estimated by its corresponding sample statistic. Given n observations in a sample, the sample variance is estimated by:

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (3)$$

where s is called the sample standard deviation.

The unit of standard deviation is always same as the unit of the variable; for example, if the variable is measured in gram per ton (gpt), the standard deviation also has the unit of gpt. For convenience, the variance of a sample can also be calculated using the following alternative equation:

$$s_x^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right] \quad (4)$$

In the equations of sample variance, note that $n-1$ (not n) is used to compute the average deviation to obtain an unbiased estimate of the variance. This statistic is always nonnegative.

As a generalization of variance for a single random variable, covariance is defined as a measure of linear associations between two random variables x and y . The population covariance between random variables X and Y is defined with respect to the population means μ_x and μ_y as:

$$\sigma_{xy} = \frac{1}{N} \sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y) \quad (5)$$

where N is the size of population.

Sample covariance, which evaluates how a pair of variables are associated relative to their means in a sample data set, is defined as follows:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (6)$$

A positive covariance would indicate a positive linear relationship between the variables, and a negative covariance would suggest the opposite.

Another useful statistic of a random variable is called the median value x_m , which is defined as the middle value when the numbers are arranged in order of magnitude. Usually, the median is used to measure the midpoint of a large set of

numbers. The median value can be determined by ranking the n values in the sample in descending order, 1 to n . If n is an odd number, the median is the value with a rank of $(n + 1)/2$. If n is an even number, the median equals the average of the two middle values with ranks $n/2$ and $(n/2) + 1$ (Illowsky and Dean 2018).

Correlation Coefficient

In geosciences, we often need to examine relationships between different variables, such as gold and silver in a mineral deposit. It is always insightful to plot out the scatter graph of the sample observations as the first step, showing a visual relationship between the two variables. Figure 2 is the scatter plot with a sample of 17 observations on gold (x) and silver (y) from a mineral deposit (population) (see Table 1), which clearly reveals that gold and silver are linearly associated or positively correlated except for a couple of outliers.

This correlation can be readily fitted by simple linear regression analysis with exclusion of the two outliers, which is not a subject of this Chapter. Quantification of the linear association between the two variables is the focus in this section.

While the covariance in Eq. (5) or (6) does measure the linear relationship between two random variables, it cannot be conveniently applied in practice since the two variables of interest could have vastly different standard deviations. It is intuitive that the covariance can be better interpreted if its value is standardized to a neat range, such as $[-1, 1]$ by removing the influence of different variances associated with variables. Thus, the Pearson coefficient (also called the product-moment coefficient of correlation) is designed in this manner so that it can be readily utilized to quantify the strength of linear correlation between variables. The Pearson coefficient in the population form is expressed as the following standardized form (Pearson 1920; Chen and Yang 2018):

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad (7)$$

The Pearson coefficient is a type of correlation coefficient that represents the relationship between two variables that are measured on the same interval or ratio scale. The Pearson coefficient is suitable to measure the strength of linear associations between two continuous variables.

Numerically, the Pearson coefficient is represented the same way as a correlation coefficient that is used in linear regression, ranging from $[-1, 1]$. A value of $+1$ is the result of a perfect positive relationship between variables. Positive correlations indicate that observations of both variable move in the same direction. Conversely, a value of -1 represents a perfect negative relationship, indicating that the two variables move in the opposite directions. A zero value means no correlation. Note that the Pearson coefficient shows correlation, not necessarily causation.

In practice, the correlation coefficient ρ is estimated by using a sample consisting of a set of observations for the pair of variables (Rumsey 2019). The sample form of the Pearson coefficient is defined as follows:

$$r = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (8)$$

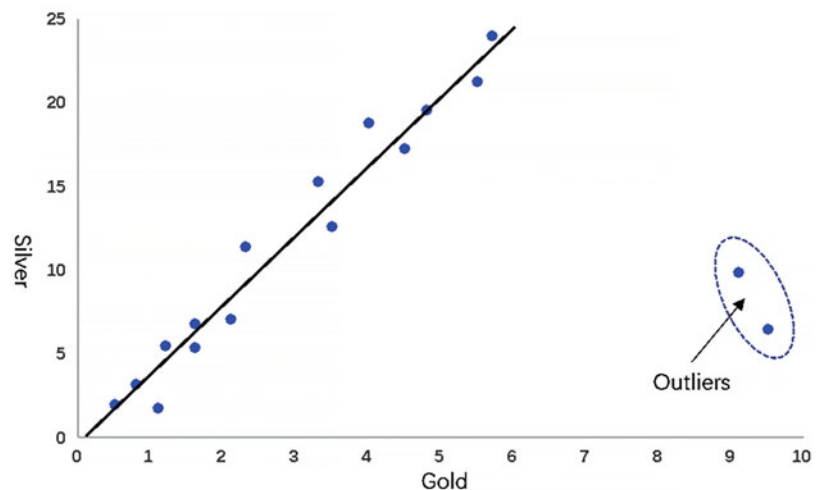
While Eq. (8) provides the means to calculate values of the correlation coefficient, the following form is more convenient in computation:

$$r = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{\sqrt{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \sqrt{n \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2}} \quad (9)$$

This equation is used most often in practice because it does not require prior computation of the means, and the

Correlation Coefficient,

Fig. 2 Scatter graph of observations for variables gold and silver from a drilling program, showing that the two variables are highly correlated except for a couple of outlying points



Correlation Coefficient, Table 1 Observations of gold and silver grades in a mineral deposit

Observation	Gold (gpt)	Silver (gpt)	Note
1	0.8	3.2	Regular sample point
2	2.3	11.4	Regular sample point
3	1.2	5.5	Regular sample point
4	4.5	17.3	Regular sample point
5	1.6	6.8	Regular sample point
6	0.5	2.0	Regular sample point
7	5.5	21.3	Regular sample point
8	5.7	24.0	Regular sample point
9	4.8	19.6	Regular sample point
10	3.5	12.6	Regular sample point
11	4.0	18.8	Regular sample point
12	2.1	7.1	Regular sample point
13	1.6	5.4	Regular sample point
14	3.3	15.3	Regular sample point
15	1.1	1.8	Regular sample point
16	9.5	6.5	Extreme point (outlier)
17	9.1	9.9	Extreme point (outlier)

computational algorithm is easily programmed for a computer (Ayyub and McCuen 2011).

The formula has been devised to ensure that the value of r will lie in the range $[-1, 1]$ with $r = -1$ for perfect negative correlation, $r = 1$ for perfect positive correlation, and $r = 0$ for no correlation at all. Using the previous example in Table 1, the correlation coefficient between gold and silver is calculated to be 0.46. From the scatter plot, two observations (9.5, 6.5) and (9.1, 9.9) are clearly deviated from the general trend of linear association between the two variables. They are called outliers, which can seriously distort the real relationship between the variables. The correlation coefficient is increased to 0.98 when these two outliers are excluded from the sample.

The Pearson correlation coefficient is good to measure the strength of linear associations between a pair of random variables without involvement of other variables. When more than two variables are there in the system, this measure is referred to as the partial correlation between two random variables by ignoring the effect of other random variables. The partial correlation coefficient will give misleading results if there is another random variable that is numerically related to both variables of interest.

Hypothesis Testing

Equations (8) and (9) provide a tool of estimating correlation coefficient between two variables given a sample consisting a set of observations. It is generally true that higher absolute value of correlation coefficient indicates greater correlation between the variables, and vice versa. In many cases,

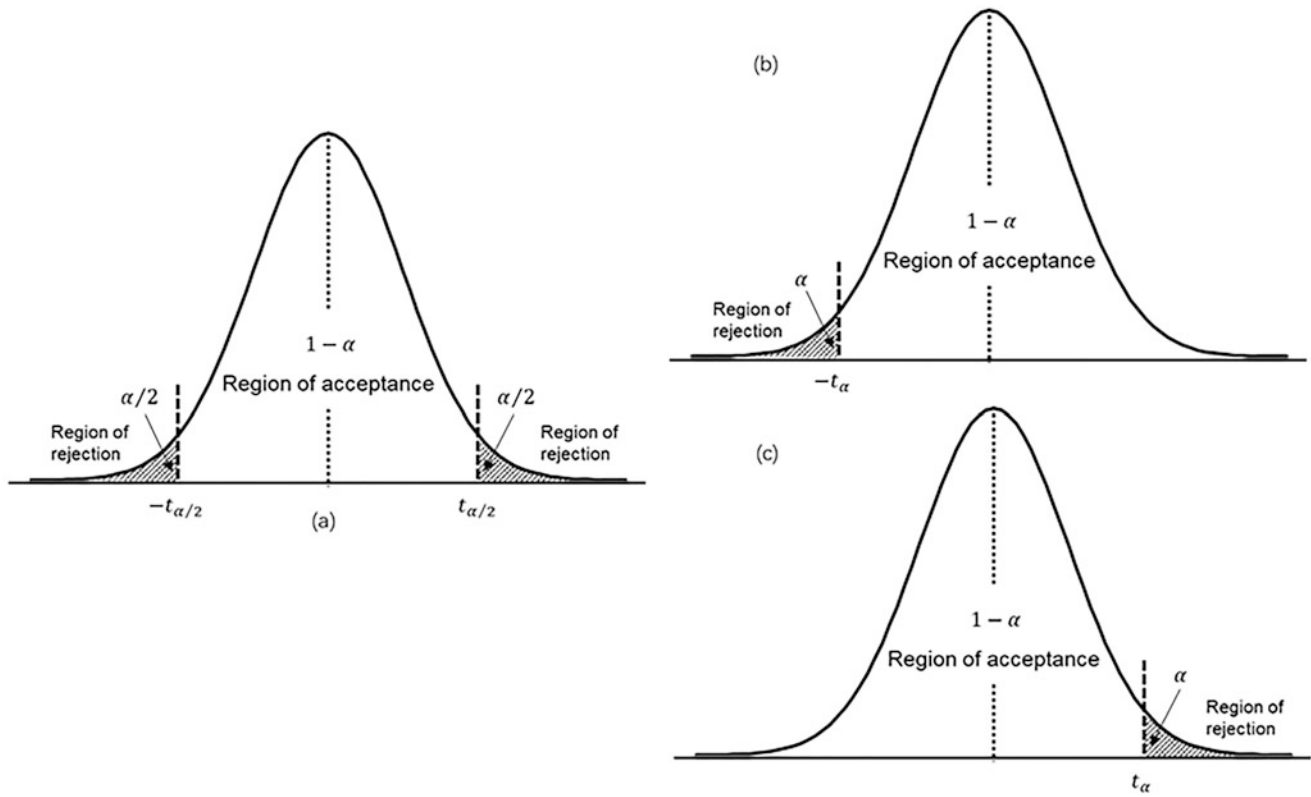
judgment or decision is necessary to make if the correlation of two variables is statistically significant in addition to the estimate of correlation coefficient. This judgment calls for the tool of hypothesis testing, which is the formal procedure for using statistical measures in the process of decision making.

Hypothesis testing represents a class of statistical techniques designed to extrapolate information from samples of data to make inferences about populations for the purpose of decision making (Ayyub and McCuen 2011). Here t -test is chosen for hypothesis testing on the statistical significance of correlation coefficient. Note that use of t -test requires that population follows a normal distribution and samples are obtained by simple random sampling. If samples differ slightly from normal distribution, t -test is still applicable, but its results will be inaccurate. As the deviation from normality increases, the results become less reliable.

The strength of correlation is generally gauged by how far from zero the value of r lies. A correlation coefficient of zero indicates that there is no linear relationship between the two variables. The t -test aims to answer the question: "Is the population correlation equal to 0?" This question can be answered by hypothesis testing through the following steps (Ayyub and McCuen 2011): (1) formulate hypotheses; (2) select an appropriate statistical model that identifies the test statistic; (3) specify the level of significance, which is a measure of risk; (4) collect a sample of data and compute estimate of the test statistic; (5) define the region of rejection for the test statistic; (6) select the appropriate hypothesis. The t - statistic can be used to test the significance of linear correlation when the sample size n is not very large.

The t -test for correlation coefficient focuses on the statistical significance of correlation between two variables. Since correlation can be positive and negative, the region of rejection will consist of values in both tails. This is illustrated in Fig. 3a, where T is the test statistic that has a continuous probability density function. In this case, the test is called a two-sided test. A one-sided hypothesis test is also needed sometime for the so-called "directional" parameters. The region of rejection for directional test is associated with values in only the upper tail or lower tail of the distribution, as illustrated in Fig. 3b, c.

The foremost step in performing the t - test is to form hypotheses, which are composed of statements involving either population distributions or parameters. Note that hypotheses should not be expressed in terms of sample statistics. The first hypothesis is called the null hypothesis denoted by H_0 , which is formulated as an equality. The second hypothesis is called alternative hypothesis denoted by H_1 , which is formulated to indicate the opposite. The null and alternative hypotheses, which are written both mathematically and grammatically, must express mutually exclusive conditions. Thus, when a statistical analysis of sampled data suggests that the null hypothesis should be



Correlation Coefficient, Fig. 3 Sampling distribution of the t -test statistic, showing the region of rejection (cross-hatched area), region of acceptance, and critical value (t_{α}): (a) two-sided test, (b) lower tail one-sided test, and (c) upper tail one-sided test

rejected, the alternative hypothesis must then be accepted. The following is a formal t -test procedure:

Step 1: Propose hypothesis H_0 , which implies that there is no linear correlation between the two variables when the population correlation coefficient is zero. The alternative hypothesis would indicate the opposite.

$$H_0 : \rho = 0 \quad (10)$$

$$H_1 : \rho \neq 0 \quad (11)$$

It can be shown that for $n > 2$ these hypotheses can be tested using a t -statistic that is given by

$$t = |r| \sqrt{(n-2)/(1-r^2)} \quad (12)$$

where t is a random variable that has approximately a t -distribution with $(n-2)$ degrees of freedom.

Step 2: Choose a level of significance α . Determine a threshold value α for the test and find out the critical value $t_{\alpha/2}(n-2)$ for a two-sided t -test at the threshold from the t -distribution table, which can be found in many text books for statistics, such as Chen and Yang (2018).

Step 3: Calculate the t -statistic using Eq. (12). The calculated t value is compared against the critical value obtained from the t -distribution table at the critical value $\alpha/2$ for a two-sided test or α for a one-sided test.

Step 4: The null hypothesis H_0 is rejected and H_1 is accepted if the calculated t value is greater than the critical value from the t -distribution table, implying that the correlation between the two variables is statistically significant at the significance level α ; otherwise, the correlation is not statistically significant at the same level.

As an illustration, assume the threshold $\alpha = 0.05$ or 5% for a two-sided test. Using the data in Table 1 by excluding the two outliers, t -value is calculated by Eq. (12), which is given by

$$\begin{aligned} t_{\alpha/2} = t_{0.025} &= 0.98 \sqrt{(15-2)/(1-0.98^2)} = 17.76 \\ &> 2.16 \end{aligned} \quad (13)$$

The null hypothesis H_0 is then rejected in this two-sided test, implying that the correlation is considered statistically significant at the level of significance of $\alpha = 0.05$, since the computed $t_{0.025} = 17.76$ value is greater than the critical value $T_{0.025} = 2.16$ given a freedoms of degree $n-2 = 13$.

As stated, the t -test is suitable when sample size is not very large. For a large sample size, other testing options are recommended, such as z – test, the chi – square test, and the f -test, etc.

Rank Correlation

All discussions above focus on correlation coefficient for continuous variables. In practice, however, it is not always necessary, or even possible, when investigating correlation, to draw on the continuous measurements that have been presented so far. An alternative is to work only from the rank positions. In such cases, the data are ranked according to their importance, class, quality, or other factors, using integer numbers $1, 2, \dots, n$. The coefficient of rank correlation can be calculated for a pair of ranked variables x and y .

While it would be possible to use Pearson correlation coefficient r , in the ranked data, there is an alternative measure, which is specially designed for ranked data called the rank coefficient of correlation r_s . It is sometimes known as “Spearman’s coefficient of rank correlation,” which is given below (Spearman 1904; Graham 2017):

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad (14)$$

where n is the number of all pairs (x_i, y_i) of ranked data and d is the difference between ranks of corresponding values of x and y .

As with the product-moment correlation coefficient, the coefficient of rank correlation r_s , has been devised to ensure that its value lies within the range $[-1, 1]$. The value for r_s can be interpreted in much the same way as the values for r were in the previous section: $r_s = 1$ for perfect positive correlation, $r_s = -1$ for perfect negative correlation, and $r_s = 0$ for no correlation.

As an illustration in a mineral exploration program, it is insightful to examine the relationship between rock type (x) and copper mineralization (y). Past exploration experience and existing geological maps provide evidence of relevancy between the two variables. Thus, rocks can be ranked according to the degree of their relevancy to the mineralization of copper. Table 2 shows the coded data of rock type and copper mineralization intensity from 12 drillhole intervals, which is explained in Table 3. The scatter plot in Fig. 4 reveals a clear positive correlation between copper mineralization and rock type.

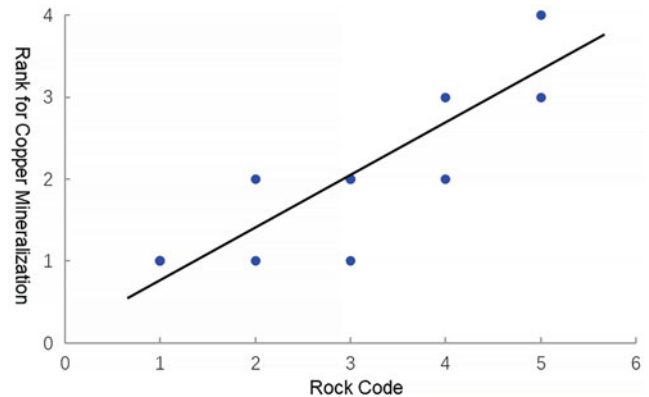
Using Eq. (14), the rank coefficient of correlation between copper mineralization and rock type is calculated to 0.94 using the 12 pairs of rank observations drawn from the drillholes. The result, also shown in Table 2, clearly indicates

Correlation Coefficient, Table 2 Coded rock type and copper mineralization

Observation	Rock code	Cu mineralization	d Square
1	2	1	1
2	1	1	0
3	3	2	1
4	2	2	0
5	3	1	4
6	4	2	4
7	1	1	0
8	1	1	0
9	4	3	1
10	5	4	1
11	3	2	1
12	5	3	4

Correlation Coefficient, Table 3 Code for rock type and rank of copper mineralization intensity

Rock type	Code	Mineralization intensity	Rank
Overburdens	1	Baren	1
Dyke	2	Weakly mineralized	2
Argilized limestones	3	Low-grade ore	3
Altered porphyry	4	High-grade ore	4
Skarn	5		



Correlation Coefficient, Fig. 4 Scatter graph between code of rock type (x) and rank of copper mineralization intensity (y). The graph clearly shows a highly correlated relationship between the two variables

that copper mineralization has a strong association with rock type.

Rock type is represented by numerical code, while copper mineralization is expressed by class numbers representing intensity of mineralization. The definitions of rock type and mineralization class are given in Table 3.

The statistical significance of the rank coefficient can also be judged by hypothesis testing similar to the correlation coefficient. However, testing methods for discrete data differ

from the *t*-test for continuous data. This is not a subject to be discussed here.

Spearman rank coefficient for correlation analysis is a nonparametric method, which has several advantages. The significance test for rank correlation doesn't depend on sample distribution. Another advantage is immune from outliers or extreme values. If the sample size is small, one big outlier can completely distort Pearson's correlation coefficient, leading to wrong conclusions. Spearman rank correlation coefficient is less affected by outliers and its results are more robust. The rank approach provides a natural filter to suppress or remove excessive influence or noises of outliers from the data.

In addition to the Spearman rank correlation coefficient introduced here, other commonly used nonparametric correlation indexes include the contingency coefficient and the Kendall rank correlation coefficient.

Summary

Random variation represents uncertainty. If the variation is systematically associated with one or more other variables, the uncertainty in estimating the value of a variable can be reduced by identifying the underlying relationship. Correlation and regression analyses are important statistical methods to achieve these objectives. They should be preceded by a graphical analysis to determine (1) if the relationship is linear or nonlinear, (2) if the relationship is direct or indirect, and (3) if any extreme events might dictate the relationship.

The Pearson correlation coefficient is defined as a quantitative index of the degree of linear association. Cautions must be taken when the correlation analysis is applied to the real world. The first caution is the preconditions for construction of Pearson correlation coefficient and its *t*-statistic test. Appropriate use of these requires that the distribution of variables involved must be the bell-shaped and that sample is collected from a representative, randomly selected portion of the total population.

Pearson product-moment correlation coefficient only quantifies the degree of linear association between two variables. A high statistical correlation between two variables can give a misleading impression about the true nature of their relationship. A strong Pearson correlation between two variables does not prove that one has caused the other. A strong correlation merely indicates a statistical link, but there may be many reasons for this besides a cause-and-effect relationship. Variables can be associated in temporally, causally, or spatially. For example, gold and silver are often temporally or spatially associated in a mineralization and geological environment. High correlation coefficient values do describe highly physical association between the two metals in this case. The cost of living and wages are a good example of causal relationship.

Pearson correlation coefficient can be readily distorted when outlying observations in the sample are present. A small number of extreme value points can completely alter the statistical significance of linear association, creating spurious correlation outcomes. Outliers or extreme values are common in sampling for variables in the geoscience fields, such as copper in a mineral deposit. Hence, it is wise to perform visual inspection and basic statistical treatments on extreme values in the original data prior to use of Pearson correlation coefficient.

An important assumption of using correlation coefficient and its *t*-test is the bell-shaped or normal distribution of random variables involved. Unfortunately, most metal grades in mineral deposits are distributed in skewed forms with long tails in the upper value side. Instead of following a normal distribution, these variables usually follow some forms of the so-called lognormal distribution. In this situation, it is appropriate to use the log scale of original data in the correlation analysis.

Bibliography

- Aldis M, Aherne J (2021) Exploratory analysis of geochemical data and inference of soil minerals at sites across Canada. *Math Geol* 53: 1201–1221
- Ayyub BM, McCuen RH (2011) Probability, statistics, and reliability for engineers and scientists. CRC Press, 624p
- Chen JH, Yang YZ (eds) (2018) Principal of statistics. Beijing Institute of Technology Press, 291p
- Graham A (2017) Statistics: an introduction. Kindle Edition, 320p
- Illowsky B, Dean S (2018) Introductory statistics. OpexStax, 905p
- Kumar S, Chong I (2018) Correlation analysis to identify the effective data in machine learning: prediction of depressive disorder and emotion states. *Int J Environ Res Public Health* 15(6):2907
- Pan GC, Harris DP (1990) Three nonparametric techniques for the optimum discretization of quantitative geological variables. *Math Geol* 22:699–722
- Pearson K (1920) Notes on the history of correlations. *Biometrika* 13: 25–45
- Rumsey DJ (2019) Statistics essentials for dummies. Wiley, 174p
- Spearman C (1904) The proof and measurement of association between two things. *Am J Psychol* 15:72–101

Coupled Modeling

Mohammad Ali Aghighi and Hamid Roshan
School of Minerals and Energy Resources Engineering,
University of New South Wales, Sydney, NSW, Australia

Definition

Coupled modeling refers to translating our understanding of a process involving two or more interacting physical or

chemical subprocesses into a set of governing equations. These partial differential equations are derived either phenomenologically or from fundamental laws. The governing equation of a subprocess is often extended to account for the effect of other interacting subprocesses through additional terms called coupling terms. Coupled models are solved analytically or numerically through one-way coupling, iterative coupling, or fully coupling approaches. The latter refers to simultaneous solution of all governing eqs. A proper characterization of many of the Earth's processes requires coupled modeling. For instance, the process of geothermal energy extraction involves heat transfer, fluid flow, and rock deformation subprocesses. This process requires cold water to be injected into the natural or induced fractures of a hot rock via a borehole so that the injected water picks up the heat and flows back to the surface through another borehole. Pressurized by the injected water, the rock fractures open, thus their flow characteristics alter. Also, the heat exchange takes place between the injected cold water and the hot rock resulting in rock contraction/expansion as well as rising water temperature. Study of such interacting processes requires coupled modeling. A geothermal energy extraction process constitutes at least three governing equations representing rock deformation, water flow, and heat transfer. Coupled modeling provides more realistic insights into the Earth's developments involving significant interacting subprocesses.

Introduction

Geoscience deals with many natural and artificial processes of which those occurring within the Earth's crust are more of interest from an applied standpoint. Although no process in

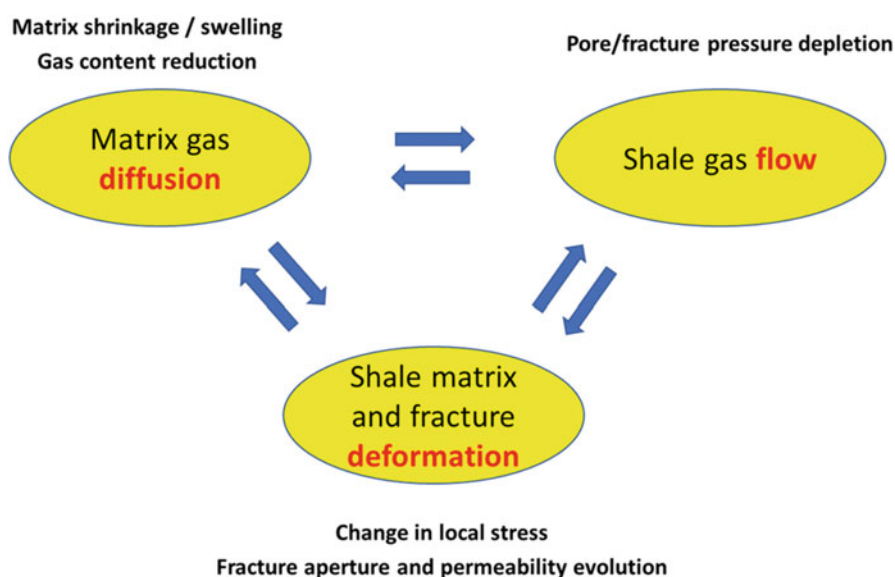
Earth takes place in isolation, some can be sufficiently modeled in an uncoupled manner if the effect of other interacting processes can be neglected within an acceptable error range. For example, the process of groundwater flow can be perceived as a sole fluid flow problem if the effect of other processes such as rock deformation on flow is negligible. Otherwise, groundwater flow is a coupled process thus requiring coupled modeling. This means that the governing equation of fluid flow must somehow incorporate the effect of rock deformation.

Another example from an applied geoscience perspective is the process of gas production/drainage from shales. This process involves several individual processes such as gas diffusion within the shale matrix, flow of gas within fractures, and the deformation of matrix due to sorption-induced shrinkage. The interaction of these processes (Fig. 1) adds to the complexity of the problem. For example, the shale matrix deformation changes the flow characteristics, which in turn affects the diffusion process within the matrix, thus deforming the shale matrix and bulk. Study of these interacting processes therefore requires coupled modeling (Aghighi et al. 2020).

In order to develop a coupled model, every contributing process needs to be first described in the form of a field eq. A field equation for a specific process is derived through combining constitutive models and conservation laws. The set of field equations representing all contributing processes is then solved using analytical or numerical methods.

Among countless physical and chemical processes occurring in Earth, solid deformation and transport phenomena are more of interest particularly in applied geoscience. Table 1 provides more information on the simple form of these processes. Most Earth developments and transformations involve the interaction of these processes.

Coupled Modeling, Fig. 1 An example of coupled subsurface processes is shale gas production where several interacting subprocesses of deformation, fluid flow, and diffusion type are involved



Coupled Modeling, Table 1 Solid deformation and transport processes

Process category	Solid deformation	Transport phenomena			
Process	Linear elasticity	Fluid flow	Heat transfer	Solute (ion) transfer	Charge transfer
Physics law	Hooke's law	Darcy's law	Fourier's law	Fick's law	Ohm's law
Flux/strain	Strain (ϵ)	Pore fluids flux ($\mathbf{q}$)	Heat flux ($\mathbf{q}_H$)	Solute flux ($\mathbf{f}$)	Charge flux ($\mathbf{i}$)
Potential/field	Displacement	Head (h)	Temperature (T)	Concentration (C)	Voltage (V)
Drive or load	Stress (σ)	Hydraulic gradient	Temperature gradient	Chemical potential gradient	Electrical potential gradient
Material property	Young's modulus (E)	Hydraulic conductivity ($\mathbf{K}$)	Thermal conductivity (κ)	Diffusion coefficient (D)	Electrical resistance (R)
Equation	$\epsilon = \frac{1}{E} \sigma$	$\mathbf{q} = -\mathbf{K} \nabla h$	$\mathbf{q}_H = -\kappa \nabla T$	$\mathbf{f} = -D \nabla C$	$\mathbf{i} = -\frac{1}{R} \nabla V$

Modified after Mitchell and Soga (2005)

When examining a geoscience-related process, it is more convenient and cost efficient to deal with only one individual process rather than a combination of some processes; however, coupled modeling is often inevitable. Natural and manmade problems in geoscience such as plate tectonics, erosion, sedimentary deposition, groundwater flow, geothermal energy extraction, oil and gas production, CO₂ capture and sequestration, storage of radioactive waste, subsidence, fault reactivation, and hydraulic fracturing inherently consist of more than one process. While the fundamentals of each individual process in physics or chemistry are somewhat well understood, an emerging challenge is to enhance our understanding of the coupled processes (Acocella 2015). This challenge cannot be tackled without studies involving coupled modeling.

The following sections describe some of the most common coupled modeling that are encountered in Earth sciences. These coupled modeling cases are explained here for relatively simple yet common situations where materials are isotropic and homogeneous, fluids are single phase, laminar and incompressible, flow is Newtonian, and the relationship between stress and strain is linear, so is that between the change of the fluid content of the porous medium and pore pressure. There are numerous extensions to the following coupled models where nonlinear relationships (e.g., plasticity and viscoelasticity), non-Newtonian flow, compressible fluids (e.g., gas), anisotropy, heterogeneity, among other complexities, are taken into account. These models have been increasingly used in applied geoscience such as geophysics and hydrogeology as well as earth-related engineering disciplines such as civil, petroleum, mining, and environmental engineering.

Hydromechanical Coupling

Hydromechanical coupling refers to the interactions of fluid flow and solid deformation. One of the simplest formulations of the hydromechanical coupling is the linear poroelastic

theory. This theory was developed by Biot (1941) as an extension to the Terzaghi's effective stress law (Terzaghi 1923). Poroelasticity has been reformulated (Rice and Cleary 1976), extended (Skempton 1954), and used in many applications (e.g., Detournay and Cheng 1988). The linear poroelastic theory is based on the following assumptions among others (Biot 1941): (1) there is an interconnected pore network uniformly saturated with fluid (2) the total pore volume is small compared to the bulk volume of the rock, and (3) pore pressure and stresses can be averaged over the volume of rock.

The governing equations of poroelasticity consist of two equations: a fluid flow equation incorporating the effect of solid deformation and a solid deformation equation incorporating the effect of fluid flow. These governing equations can be written in terms of different pairs of unknowns; however, it is more common to write and solve these equations in terms of displacement and pore pressure representing the solid deformation and fluid flow model, respectively.

Solid Deformation Model and Its Coupled Form

The combination of Hooke's law and the equilibrium equation leads to (Jaeger et al. 2007):

$$(\lambda + G)\nabla(\nabla \cdot \mathbf{u}) + G\nabla^2 \mathbf{u} = 0 \quad (1)$$

where λ and G are Lamé's constants and $\mathbf{u}$ is the displacement vector. Incorporating the effect of fluid flow in Eq. 1 yields (compressive stresses are positive) (Jaeger et al. 2007):

$$(\lambda + G)\nabla(\nabla \cdot \mathbf{u}) + G\nabla^2 \mathbf{u} + \alpha \nabla p = 0 \quad (2)$$

where α is the Biot's coefficient and p is the pore pressure. The term $\alpha \nabla p$ is the hydromechanical coupling term.

Fluid Flow Model and Its Coupled Form

The diffusivity equation results from substituting the Darcy's law in the continuity equation (Ahmed 2010):

$$\frac{\partial p}{\partial t} = \frac{k}{\mu \phi c_f} \nabla^2 p \quad (3)$$

where t is time, k is the permeability of the porous medium, μ is the fluid viscosity, ϕ is the solid porosity, and c_f is the fluid compressibility. Equation 3 can be extended to account for the effect of solid deformation as follows (Detournay and Cheng 1993):

$$\frac{\partial p}{\partial t} = \frac{kM}{\mu} \nabla^2 p + \alpha M \frac{\partial}{\partial t} (\nabla \cdot \mathbf{u}) \quad (4)$$

where M is the Biot's modulus. The last term in the right-hand side is the coupling term representing the effect of solid deformation on fluid flow.

The set of Eqs. 2 and 4 is a well-posed mathematical problem including four equations with four unknowns (three displacement components and the pore pressure).

Thermomechanical Coupling

Thermomechanical coupling or thermoelasticity considers the mutual effect of solid deformation and temperature. This coupling is very similar to the poroelastic coupling with the temperature replacing the pore pressure. The set of governing equations of thermoelectricity is as follows (Coussy 1995):

$$(\lambda + G) \nabla (\nabla \cdot \mathbf{u}) + G \nabla^2 \mathbf{u} + 3\alpha_T K \nabla T = 0 \quad (5)$$

$$\frac{\partial T}{\partial t} = \frac{k_T}{\rho c_v} \nabla^2 T + \frac{3K\alpha_T T_0}{\rho c_v} \frac{\partial}{\partial t} (\nabla \cdot \mathbf{u}) \quad (6)$$

where T is the difference between the current and the reference temperature, α_T is the coefficient of linear thermal expansion, K is the bulk modulus and k_T is the thermal conductivity, ρ is the density, and c_v is the specific heat at constant strain. Equation 6 is a diffusion type equation for temperature, which is obtained by introducing the Fourier's law equation into the conservation of energy equation. Note that $3\alpha_T K \nabla T$ is the thermoelastic coupling coefficient. It can be shown that the effect of solid deformation on heat transfer is negligible for typical values of the parameters involved. Therefore, Eq. 6 is reduced to:

$$\frac{\partial T}{\partial t} = \frac{k_T}{\rho c_v} \nabla^2 T \quad (7)$$

Equation 7 is independent of solid deformation and has been analytically solved for different geometries and boundary conditions (Carslaw and Jaeger 1959).

Thermo-Hydromechanical Coupling

Where change in temperature and rock deformation both affect the fluid flow in a porous medium, thermo-hydromechanical coupling is required for modeling of the system. In this case, three governing equations are needed to describe the three processes involved (Charlez 1991):

$$(\lambda + G) \nabla (\nabla \cdot \mathbf{u}) + G \nabla^2 \mathbf{u} + \alpha \nabla p + 3\alpha_T K \nabla T = 0 \quad (8)$$

$$\frac{\partial p}{\partial t} = \frac{kM}{\mu} \nabla^2 p + \alpha M \frac{\partial}{\partial t} (\nabla \cdot \mathbf{u}) - \alpha_{sf} M \frac{\partial T}{\partial t} \quad (9)$$

$$\frac{\partial T}{\partial t} = \frac{k_T}{\rho c_v} \nabla^2 T \quad (10)$$

where α_{sf} is a thermic coefficient.

Chemo-Hydromechanical Coupling

In a chemically active saturated porous system such as shales and clay-rich rocks, three processes are in interaction: solid deformation, fluid flow, and ion transfer. The ion transfer can alter the state of stress around boreholes causing instabilities (Roshan and Aghighi 2011). Clays swell or shrink when they adsorb or lose water because of the differences in chemical potentials of the species in pore fluid and other fluids that the rock is exposed to. These developments lead to changes in permeability, hence the flow pattern as well as the stress distribution in the formation. In an isotropic formation, where other assumptions of linear poroelasticity also hold, the following governing equations constitute a chemo-poroelastic model (Ghassemi and Diek 2002):

$$(\lambda + G) \nabla (\nabla \cdot \mathbf{u}) + G \nabla^2 \mathbf{u} + \alpha \nabla p - \alpha_{uc} \nabla C = 0 \quad (11)$$

$$\frac{\partial p}{\partial t} = \frac{kM}{\mu} \nabla^2 p + \alpha M \frac{\partial}{\partial t} (\nabla \cdot \mathbf{u}) - \alpha_{pc} \frac{\partial C}{\partial t} - \beta_{pc} \nabla^2 C \quad (12)$$

$$\frac{\partial C}{\partial t} = \alpha_{cc} \nabla^2 C - \nabla (JC) \quad (13)$$

where α_{uc} , α_{pc} , β_{pc} , α_{cc} , and J are chemical and flow coefficients elaborated in Roshan and Aghighi (2011). Equations 11 and 12 include the effects of other interacting processes through coupling terms; however, Eq. 13 is independent of the two other processes (i.e., rock deformation and fluid flow), thus can be solved using existing analytical solutions.

Other cases such as chemo-thermo-hydromechanical coupling can be found in Roshan and Aghighi (2012).

Solving Coupled Models

Coupled models can be solved analytically, numerically, or in a hybrid form. Analytical methods are used either to solve the set of governing equations in a closed form or simplify them for a numerical analysis. Even after being reduced in terms of spatial dimensions and simplifying boundary and initial conditions, only a small fraction of coupled models can be solved analytically. Nonetheless, analytical solutions are widely used for gaining useful preliminary insights into the problem through initial simulations and analysis. They are also employed for benchmarking and validation of numerical solutions of the real problem in more complex forms. Function transformations such as Laplace, Fourier, and Hankel transforms as well as differential operator methods are commonly used for solving coupled models analytically (Bai and Elsworth 2000).

Analytical solutions are convenient and simple to use; however, numerical methods can provide solutions for complex geometries and conditions representing more realistic replications of the problem. Numerical methods discretize the space domain into elements and turn the continuous governing equations into finite equations. The finite element method (FEM) (Zienkiewicz and Taylor 2000), the finite difference method (FDM) (LeVeque 2007), and the boundary element method (BEM) (Banerjee and Butterfield 1981) are the most common numerical methods. The FEM is the most capable among the foregoing methods in terms of handling complex geometries and boundary conditions as well as heterogeneity and nonlinearity (Bai and Elsworth 2000).

Coupling Approaches

Coupled models can be solved based on different coupling approaches: partially coupled and fully coupled. In partial coupling, governing equations are solved separately in a certain order with results being passed onto the next solution at each time step. In hydromechanical coupling, for instance, pore pressure can be evaluated first by solving the governing equation of fluid flow. The pore pressure results will be then introduced into the governing equation of rock deformation (geomechanics) to calculate displacements and stresses. The procedure is repeated for other time steps. If there is no iteration involved at each time step, the method is called “one-way coupling” (Longuemare et al. 2002). The effect of geomechanics on fluid flow is not taken into account in one-way coupling approach and the governing equations are solved once at each time step.

If the results are exchanged at each time step in an iterative manner until a convergence criterion is satisfied, the coupling is called iterative. Results of the iterative coupling approach

are obviously more accurate than that of the one-way coupling approach.

Since the interactions of subprocesses take place simultaneously, it is sensible to solve their respective equations concurrently. This is what signifies a fully coupled approach, which provides accurate results without any need for iteration. The fully coupled approach requires less time compared to the iterative coupling. The governing equations, however, often need to be simplified to be suitable for full coupling and meet the requirements of solution procedures.

Summary

Many processes in the earth involve two or more interacting subprocesses. Coupled modeling is required to describe such processes. A coupled model is a set of governing equations characterizing all contributing subprocesses. Governing equations are derived by combining constitutive and conservation laws. Some of the widely occurring coupling processes can be characterized using hydromechanical, thermo-mechanical, thermo-hydromechanical, and chemo-hydromechanical couplings. Coupled models are solved using analytical and numerical models. While analytical models are more convenient to use, numerical models can handle more complex geometries and conditions. Common approaches for solving coupled models are one-way coupling, iterative coupling, and fully coupling methods. In the latter approach, governing equations are solved simultaneously. Solving models in a fully coupled manner has the advantage of requiring less time compared to the iterative approach; however, it is not applicable to all governing equations unless some simplifications are made.

Cross-References

- Computational Geoscience
- Earth System Science
- Flow in Porous Media
- Fast Fourier Transform
- Geomechanics
- Laplace Transform
- Mathematical Geosciences
- Partial Differential Equations
- Simulation

Bibliography

- Acocella V (2015) Grand challenges in earth science: research toward a sustainable environment. *Front Earth Sci* 3(68). <https://doi.org/10.3389/feart.2015.00068>

- Aghighi MA, Lv A, Roshan H (2020) Non-equilibrium thermodynamic approach to mass transport in sorptive dual continuum porous media: a theoretical foundation and numerical simulation. *J Nat Gas Sci Eng*:103757. <https://doi.org/10.1016/j.jngse.2020.103757>
- Ahmed T (2010) *Reservoir engineering handbook*. Elsevier Science, San Diego
- Bai M, Elsworth D (2000) *Coupled processes in subsurface deformation, flow, and transport*. American Society of Civil Engineers, Reston
- Banerjee PK, Butterfield R (1981) *Boundary element methods in engineering science*, vol 17. McGraw-Hill, London
- Biot MA (1941) General theory of three-dimensional consolidation. *J Appl Phys* 12(2):155–164. Retrieved from <http://link.aip.org/link/?JAP/12/155/1>
- Carslaw H, Jaeger J (1959) *Conduction of heat in solids*. Clarendon Press, Oxford
- Charlez PA (1991) *Rock mechanics: theoretical fundamentals*. Éditions Technip, Paris
- Coussy O (1995) *Mechanics of porous continua*. Wiley
- Detournay E, Cheng AHD (1988) Poroelastic response of a borehole in a non-hydrostatic stress field. *Int J Rock Mech Min Sci Geomech Abstract* 25:171–182
- Detournay E, Cheng AHD (1993) Fundamentals of poroelasticity. In: Hudson (ed) *Comprehensive rock engineering: principles, practice and projects*, vol 2. Pergamon Press, Oxford, pp 113–171
- Ghassemi A, Diek A (2002) Poroelastoclasticity for swelling shales. *J Pet Sci Eng* 34(1–4):123–135. [https://doi.org/10.1016/S0920-4105\(02\)00159-6](https://doi.org/10.1016/S0920-4105(02)00159-6)
- Jaeger JC, Cook NGW, Zimmerman R (2007) *Fundamentals of rock mechanics*. Wiley, Malden
- LeVeque RJ (2007) *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*. SIAM, Philadelphia
- Longuemare P, Mainguy M, Lemonnier P, Onaisi A, Gerard C, Koutsabeloulis N (2002) *Geomechanics in Reservoir Simulation: Overview of Coupling Methods and Field Case Study Oil & Gas Science and Technology* 57:471–483
- Mitchell JK, Soga K (2005) *Fundamentals of soil behavior*, vol 3. Wiley, New York
- Rice JR, Cleary MP (1976) Some basic stress diffusion solutions for fluid-saturated elastic porous media with compressible constituents. *Rev Geophys* 14(2):227–241. <https://doi.org/10.1029/RG014i002p00227>
- Roshan H, Aghighi MA (2011) Chemo-poroelastic analysis of pore pressure and stress distribution around a wellbore in swelling shale: effect of undrained response and horizontal permeability anisotropy. *Geomech Geoen* 7:209–218. <https://doi.org/10.1080/17486025.2011.616936>
- Roshan H, Aghighi MA (2012) Analysis of pore pressure distribution in shale formations under hydraulic, chemical, thermal and electrical interactions. *Transp Porous Media* 92(1):61–81. <https://doi.org/10.1007/s11242-011-9891-x>
- Skempton AW (1954) The pore-pressure coefficients A and B. *Geotechnique* 4(4):143–147. <https://doi.org/10.1680/geot.1954.4.4.143>
- Terzaghi K (1923) Die brechnung der durchlassigkeitsziffer des tones aus dem verlauf der hydrodynamischen spannungserscheinungen. *Sitz Akad Wissen, Wien Math Naturwiss Kl, Abt Ila* 132:105–124
- Zienkiewicz OC, Taylor RL (2000) *The finite element method – basic formulation and linear problems*, vol 1, 5th edn. Butterworth-Heinemann, Oxford

Cracknell, Arthur Phillip

Kasturi Devi Kanniah

TropicalMap Research Group, Centre for Environmental Sustainability and Water Security (IPASA), Faculty of Built Environment and Surveying, Universiti Teknologi Malaysia, Johor Bahru, Malaysia



Fig. 1 Arthur Phillip Cracknell (1940–2021), © Prof. Kasturi Kanniah

Biography

Arthur P. Cracknell passed away in April 2021. He was an emeritus professor at the School of Engineering, University of Dundee, UK, from September 2002 until April 2021. He was one of the pioneers in the field of remote sensing technology and his contribution to the development and advancement of the technology is remarkable. I regard him as my mentor and a dear friend and I am honored to write a biography about him.

Professor Cracknell started remote sensing works in the late 1970s when the field was still in its infancy in the UK. The establishment, development, and operations of the National Oceanic and Atmospheric Administration's Polar Orbiting Environmental Satellites data receiving station in Dundee University in 1978 was the catalyst for the development of remote sensing in the UK. While he was there, Professor Cracknell initiated research on processing and interpreting remote sensing data that was required by a group of environmental scientists and engineers in Dundee for their environmental consultancy works.

Professor Cracknell established and led a very active research group in remote sensing at Dundee University and trained many foreign PhD students including two of the current UTM professors in remote sensing. Most of the

research conducted by his group focused on the modeling of atmospheric transmission and earth surface conditions, development of software to process remotely sensed data, and the environmental application of remote sensing data.

Besides being actively involved in research projects, he also initiated academic programs via summer schools and postgraduate courses. The summer schools played a great role in advancing the application of remote sensing technology within a teaching and research context in many organizations around the world. The MSc program in remote sensing, image processing, and applications produced many undergraduates and postgraduates over the years.

Professor Cracknell's contribution to remote sensing knowledge dissemination through academic journals is noteworthy. In 1983, he became the chief editor of the *International Journal of Remote Sensing* and later the editor of *Remote Sensing Letters*. He published over 300 research papers and over 30 books covering physics and various remote sensing applications. These publications had significant contributions for innovations and potential future research in remote sensing. He is recognized for these contributions through various prestigious awards that he has received.

Since his retirement from Dundee University, he continued as an active scientist and researcher with universities within and outside of the UK. He was a senior visiting professor at UTM, between 2003 and 2019. At UTM he was engaged with teaching courses to undergraduates, co-supervising PhD candidates, reviewing the curriculum of remote sensing programs, delivering talks, and assisting in research activities.

I became acquainted with him since 2009 when I started to work closely on a project related to carbon storage of oil palm trees through which I had an opportunity to learn a lot from him. I have always noticed that he loved working with students and was always willing to discuss their research projects and manuscripts while imparting his knowledge and values to them. Through his academic network, he also assisted many students to secure scholarships and travel grants to present their research findings in international conferences. He helped to link UTM researchers with scholars from many world-renowned universities such as the Tsinghua University in Beijing. The connection later enabled UTM to expand its research collaboration in many other areas. On behalf of UTM, I thank Professor Cracknell for all his contributions.

Since my early interaction with Professor Cracknell, I have been fascinated by his passion in conducting research and scientific writing even after his retirement. In fact, due to his profound dedication for research, he was always willing to travel with his own expenses. I consider Professor Cracknell to be a principled, responsible, and highly motivated scholar and a gentleman. I feel very honored and thankful to have known such a great person and to have coauthored almost 20 journal articles, conference papers, and to have jointly supervised postgraduate students.

Acknowledgment This chapter is dedicated to honoring the late Professor Cracknell for his contribution to the development of remote sensing. He definitely deserves the honor as a remarkable scientist, educator, and individual.

Cressie, Noel A.C.

Jay M. Ver Hoef

Alaska Fisheries Science Center, NOAA Fisheries, Seattle, WA, USA



Fig. 1 Noel A. C. Cressie, courtesy of Prof. Cressie

Biography

Cressie is best known for his work in spatial statistics. In his widely cited, 900- page book, *Statistics for Spatial Data*, Cressie (1991) developed a general spatial model that unified disciplines using geostatistical data, regular and irregular lattice data, point patterns, and sets in Euclidean space. Since the 1980s, he has been a developer of statistical theory and methods, such as weighted least squares (see article on *ordinary least squares*) for fitting variograms (see article on *variograms*), and he has also written on the history of geostatistics (see articles on *geostatistics*, *kriging*). His book contains his contributions, and those of many others, in the field of spatial statistics.

Since the 2000s, Cressie has been actively researching and applying spatiotemporal models using a hierarchical-statistical-framework. He has made groundbreaking innovations in dimension-reduction to make spatial and spatiotemporal modeling computationally feasible for “big data.” He has developed new classes of nonseparable spatiotemporal covariance models (Cressie and Huang 1999) (see articles on *spatiotemporal analysis*, *spatiotemporal modeling*) and used them in many applications in the geosciences. He synthesized this knowledge in the award-winning books

Statistics for Spatio-Temporal Data (Cressie and Wikle 2011) and *Spatio-Temporal Statistics with R* (Wikle et al. 2019). In statistical theory, Cressie is also highly regarded for his fundamental contributions to goodness-of-fit, which provided a new unified view of this classic research field (Read and Cressie 1988) and Bayesian sequential testing (Cressie and Morgan 1993).

Cressie has published 4 books and over 300 articles/chapters/discussions in scholarly journals and edited books. Among numerous awards and fellowships, in 2009, he received from the Committee of Presidents of Statistical Societies (COPSS) one of the highest awards in statistical science—the R.A. Fisher Award and Lectureship. In 2014, Cressie was awarded the Pitman Medal by the Statistical Society of Australia for his outstanding achievements in the discipline of Statistics, and in 2018, he was elected a Fellow of the Australian Academy of Science. More on Cressie's background and history may be found at https://en.wikipedia.org/wiki/Noel_Cressie and in Wikle and Ver Hoef (2019).

Cross-References

- [Geostatistics](#)
- [Kriging](#)
- [Matheron, Georges](#)
- [Ordinary Least Squares](#)
- [Spatial Statistics](#)
- [Spatiotemporal Analysis](#)
- [Spatiotemporal Modeling](#)
- [Variogram](#)
- [Watson, Geoffrey S.](#)

References

- Cressie NAC (1991) *Statistics for spatial data*. John Wiley & Sons, New York, NY
- Cressie N, Huang H-C (1999) Classes of nonseparable, spatio-temporal stationary covariance functions. *J Am Stat Assoc* 94:1330–1340. (Correction, 2001, Vol. 96, p. 784)
- Cressie N, Morgan PB (1993) The VPRT: a sequential testing procedure dominating the SPRT. *Economet Theor* 9:431–450
- Cressie N, Wikle CK (2011) *Statistics for Spatio-temporal data*. John Wiley & Sons, Hoboken
- Read TR, Cressie NA (1988) *Goodness-of-fit statistics for discrete multivariate data*. Springer, New York, NY
- Wikle CK, Ver Hoef JM (2019) A conversation with Noel Cressie. *Stat Sci* 34:349–359
- Wikle CK, Zammit-Mangion A, Cressie N (2019) *Spatio-temporal statistics with R*. CRC/Chapman and Hall, Boca Raton

Crystallographic Preferred Orientation

Helmut Schaeben

Technische Universität Bergakademie Freiberg, Freiberg, Germany

Synonyms

Crystallographic texture

Definition

Crystallographic preferred orientation is a multidisciplinary topic of crystallography (solid states physics), mineralogy, mathematics, and materials science and geosciences. Its subject used to be the statistical distribution of the spatial orientations of crystallites (grains) within polycrystalline specimen, a distinct pattern of which is referred to as texture. Since spatially resolving electron backscatter diffraction delivers digital orientation map images, contemporary texture analysis comprises the spatial distribution of grains by means of image analysis and includes intragrain orientation distributions and grain boundary phenomena.

Historical Notes

Analysis of crystallographic preferred orientation or crystallographic texture analysis and its application originate in materials science and have been initiated by Günter Wassermann in the 1930s and largely extended together with Johanna Grewen (Wassermann and Grewen 1962). Single crystals are largely anisotropic with respect to almost all physical properties, one of the best-known examples in geosciences is the anisotropy of thermal expansion of quartz crystals which is for the crystallographic a-axis about twice as large as for the crystallographic c-axis. Thus, directional or tensorial physical properties of polycrystalline materials such as metals, rocks, ice, and many ceramics depend on the statistical distribution of orientations of grains assuming that the orientation of a grain is unique. Vice versa, a grain could be defined as a spatial domain of material support of a unique crystallographic orientation.

The first experiments to determine texture of a specimen of polycrystalline material were done with X-ray diffraction in a texture goniometer and provided integral orientation measurements. A texture goniometer applies Bragg's law

$$2d \sin \theta = n\lambda, n \in \mathbb{N}, \quad (1)$$

to select a feasible diffracting crystallographic lattice plane with interplanar distance d when keeping the angle θ and the wavelength λ fixed. Then a texture goniometer measures the diffracted intensities of the selected lattice planes with their common unit normal crystallographic direction $\mathbf{h} \in \mathbb{S}^2$ aligned with a given but variable macroscopic direction $\mathbf{r} \in \mathbb{S}^2$ referring to the specimen. These intensities are eventually modeled by a bi-directional probability density function $P(\mathbf{h}, \mathbf{r})$ of normals $\mathbf{h}$ of specified crystallographic lattice planes and specimen directions $\vec{r}$, and its equal area projections on the unit disk are called pole figures. In 1965, Bunge and Roe suggested simultaneously (Bunge 1965; Roe 1965) a mathematical approach in terms of a harmonic (Fourier) series expansion to determine a crystallographic orientation density function f from several $\mathbf{h}$ -pole figures. However, their approach was flawed by the same erroneous conclusion that the odd-order harmonic coefficients of an orientation density function are identical to 0, i.e., that the orientation density function is an even function because experimental pole figures are even, $P(\mathbf{h}, \mathbf{r}) = P(\mathbf{h}, -\mathbf{r})$ due to Friedel's law implying that the diffraction experiment (excluding anomalous scattering) itself introduces a center of symmetry whether it is included in the actual crystal symmetry class or not. Their model of the relationship of an orientation and pole probability density function was incomplete in that it missed to account for both directions $\mathbf{h}$ and its antipodal $-\mathbf{h}$. The mistake was abetted by their rather symbolic notation. Numerical realizations of their model in various computer codes led to occurrences of false texture components in the computed orientation density functions, which became notorious as terrifying "texture ghosts" for more than a decade. The riddle was resolved by a proper model accounting for $(\mathbf{h}, -\mathbf{h})$ by Matthies (1979).

In the early 1990s, orientation imaging microscopy has been developed (Adams et al. 1993; Schwartz et al. 2009) based on electron backscatter diffraction (EBSD) providing spatially resolved individual orientation measurements. The spatial reference of these measurements provides the essential prerequisite to model the geometry and topology of grains at the surface of the specimen by means of mathematical approaches and their respective assumptions. Thus, EBSD experiments open the venue to realize Bruno Sander's vision of a comprehensive fabric analysis (Sander 1950) beyond axes distribution analysis including grain size distribution and grain shape analysis as well as various misorientation distributions, grain boundary classification and directional grain boundary distribution, and other instructive entities.

Introduction

The subject of texture analysis is crystallographic preferred orientation, i.e., of patterns of crystallographic orientations in

terms of their statistical or spatial distribution within a specimen. The statistical distributions range from a uniform (or random) texture representing the lack of any preferential pattern, to sharp textures concentrated around a preferred orientation with a deviation of a few degrees, referred to as a technical single crystal orientation. Texture analysis is a field of multidisciplinary interaction of crystallography, mathematics, solid state physics, mineralogy, and materials science, on the one hand, and geophysics and geology, on the other hand. Analysis of crystallographic (preferred) orientation is based on integral orientation measurements with X-ray, neutron, or synchrotron diffraction, or individual orientation measurements with electron backscatter diffraction (EBSD). Integral orientation measurements require to resolve the inverse problem to estimate an orientation density function from experimental pole intensities. The key to derive any reasonable method is to recognize that their relationship can mathematically be modeled in terms of a Funk-Radon transform. Estimation of an orientation density function from given individual orientation measurements is done with non-parametric spherical kernel density estimation. Moreover, spherical statistics applies to analyze EBSD data. Since EBSD results in spatially referenced individual crystallographic orientations, they provide spatially resolved information and can be visualized as orientation maps. Then crystallographic grains can be modeled, and texture analysis can be extended to fabric analysis.

Definitions

An orientation is an element of the special orthogonal group $\text{SO}(3)$ of rotations in $\mathbb{R}^3$ thought of a rotational configuration of one right-handed orthonormal 3-frame, i.e., an ordered basis of $\mathbb{R}^3$, say $\mathcal{K}_S$, with respect to another right-handed orthonormal 3-frame, denoted $\mathcal{K}_C$. The orientation of the 3-frame $\mathcal{K}_C$ with respect to the 3-frame $\mathcal{K}_S$ is defined as the element $g \in \text{SO}(3)$ such that

$$g\mathcal{K}_S = \mathcal{K}_C. \quad (2)$$

Then the coordinate vectors $\mathbf{h}_{\mathcal{K}_C}$ with respect to the 3-frame $\mathcal{K}_C$ and $\vec{r}_{\mathcal{K}_S}$ with respect to the 3-frame $\mathcal{K}_S$ of a unique unit vector $\mathbf{v} \in \mathbb{S}^2$ are related by

$$g\mathbf{h} = \mathbf{r}. \quad (3)$$

Spatial shifts between the 3-frames are not considered.

A crystallographic orientation takes crystallographic symmetry into account. Crystallographic symmetry enables to distinguish mathematically different orientations of a crystal which are physically equivalent. Representing crystallographic symmetry in terms of the point group S_{point} of the

crystal, its crystallographic orientation is defined as the (left) coset $gS_{\text{point}} = \{gq \mid q \in S_{\text{point}}\}$ of all orientations crystallographically symmetrically equivalent to a given orientation g . A comprehensive treatise of rotations and crystallographic orientations is found in (Morawiec 2004). In case of diffraction experiments, not only the crystallographic symmetry but also symmetry imposed by the diffraction itself (Friedel's law) has to be considered. This symmetry is generally described by the Laue group S_{Laue} which is the point group of the crystal augmented by inversion. However, the cosets of equivalent orientations can be completely represented by proper rotations. Likewise, when analyzing diffraction data for preferred crystallographic orientation, it is sufficient to consider the restriction of the Laue group S_{Laue} to its purely rotational part $\tilde{S}_{\text{Laue}}$. Then, two orientations g and $g' \in \text{SO}(3)$ are crystallographically symmetrically equivalent with respect to $\tilde{S}_{\text{Laue}}$ if there is a symmetry element $q \in \tilde{S}_{\text{Laue}}$ such that $gq = g'$. Thus, a crystallographic orientation is a left coset. The left cosets $g\tilde{S}_{\text{Laue}}$ define classes of crystallographically symmetrically equivalent orientations. The set of all cosets is called quotient space.

Two crystallographic directions $\mathbf{h}$ and $\mathbf{h}' \in \mathbb{S}^2$ are crystallographically symmetrically equivalent if a symmetry element $q \in \tilde{S}_{\text{Laue}}$ exists such that $q\mathbf{h} = \mathbf{h}'$. The set of all crystallographically symmetrically equivalent directions $\tilde{S}_{\text{Laue}}\mathbf{h} \subset \mathbb{S}^2$ may be viewed as a set of positive normals sprouting from the symmetrical equivalent crystallographic lattice planes.

A fundamental entity of texture analysis is the statistical distribution of crystallographic orientations by volume (as opposed to number), i.e., of the volume portion of a specimen supporting crystallographic orientations within a given (infinitesimally) small range neglecting its spatial distribution. The statistical distribution is usually represented by an orientation probability density function f almost always falsely referred to as orientation distribution function since the early days of texture analysis in materials science and solid state physics, respectively, cf. (Bunge 1965; Roe 1965). Thus, an orientation density function f models the relative frequencies of crystallographic orientations within a specimen by volume. It is properly defined on the corresponding quotient space $\text{SO}(3)/\tilde{S}_{\text{Laue}}$, i.e.,

$$f(g) = f(gq), g \in \text{SO}(3), q \in \tilde{S}_{\text{Laue}}. \quad (4)$$

Crystallographic preferred orientation may also be represented by a pole density function P modeling the relative frequencies that an element of $\tilde{S}_{\text{Laue}}\mathbf{h} \subset \mathbb{S}^2$ coincides with a given specimen direction $\mathbf{r} \in \mathbb{S}^2$, i.e., one of the crystallographically symmetrically equivalent directions $\tilde{S}_{\text{Laue}}\mathbf{h}$ or its antipodal symmetric direction coincides with a given specimen direction $\mathbf{r} \in \mathbb{S}^2$. Thus, a pole density function satisfies the symmetry relationships

$$P(\mathbf{h}, \mathbf{r}) = P(q\mathbf{h}, \mathbf{r}), (\mathbf{h}, \mathbf{r}) \in \mathbb{S}^2 \times \mathbb{S}^2, q \in \tilde{S}_{\text{Laue}}, \text{ and}$$

$$P(\mathbf{h}, \mathbf{r}) = P(-\mathbf{h}, \mathbf{r}),$$

i.e., it is essentially defined on $\mathbb{S}^2/\tilde{S}_{\text{Laue}} \times \mathbb{S}^2$.

Pole density functions thought of as functions of the variable specimen direction $\mathbf{r}$ parametrized by the crystallographic axis $\mathbf{h}$ are experimentally accessible as pole intensities from X-ray, neutron, or synchrotron diffraction for some crystallographic forms $\mathbf{h}$ and displayed as $\mathbf{h}$ -pole figures. Orientation probability density and pole density function are related to each other by the Funk-Radon transform (Bernstein and Schaeber 2005). Once an orientation density function has been estimated from experimental $\mathbf{h}$ -pole intensities (or from individual orientation measurements), its corresponding fitted pole density function may be computed and displayed as function of $\mathbf{r}$ parametrized by $\mathbf{h}$, or vice versa; the latter are referred to as inverse pole figures where the parameter $\mathbf{r}$ is usually chosen to be one of the axes of the specimen coordinate system.

Estimation of an Orientation Density Function

An orientation density function can be estimated from integral orientation measurements as provided by X-ray, synchrotron, or neutron diffraction experiments in terms of pole figures or from individual orientation measurements as provided by spatially resolving electron backscatter diffraction (EBSD) experiments in terms of orientation maps. Once an orientation density function is estimated sufficiently well, characteristics quantifying its features such as Fourier coefficients, texture index, entropy, modes, volume portions around peaks or fibers, and others can be computed.

Estimation with Integral Orientation Measurements

Up to normalization, the pole intensities $P(\mathbf{h}, \mathbf{r})$ as displayed in pole figures are a measure of the mean density of orientations rotating the elements of $\tilde{S}_{\text{Laue}}\mathbf{h}$ to the macroscopic direction $\mathbf{r}$. The key element of a mathematical model for crystallographic pole density functions is the Funk-Radon transform $\mathcal{R}f$ of an orientation density function f defined on $\text{SO}(3)$. Here, the Funk-Radon transform assigns the mean values of f along geodesics $G \subset \text{SO}(3)$ to f . Any geodesic G can be parametrized by a pair of unit vectors $(\mathbf{h}, \mathbf{r}) \in \mathbb{S}^2 \times \mathbb{S}^2$ as

$$G(\mathbf{h}, \mathbf{r}) = \{g \in \text{SO}(3) \mid g\mathbf{h} = \mathbf{r}, (\mathbf{h}, \mathbf{r}) \in \mathbb{S}^2 \times \mathbb{S}^2\}. \quad (5)$$

Since

$$\bigcup_{\mathbf{h} \in \mathbb{S}^2} G(\mathbf{h}, \mathbf{r}) = \bigcup_{\mathbf{r} \in \mathbb{S}^2} G(\mathbf{h}, \mathbf{r}) = \text{SO}(3), \quad (6)$$

the geodesics provide a double (Hopf) fibration of $\text{SO}(3)$. Then the Funk-Radon transform $\mathcal{R}f$ of f defined on $\text{SO}(3)$ is defined on $\mathbb{S}^2 \times \mathbb{S}^2$ as

$$\mathcal{R}f(\mathbf{h}, \mathbf{r}) = \frac{1}{2\pi} \int_{G(\mathbf{h}, \mathbf{r})} f(g) dg. \quad (7)$$

It is a probability density function if the function f is a probability density function. The Funk-Radon transform $\mathcal{R}f$ possesses a unique inverse (Helgason 1999). Thus, the initial function f can be uniquely recovered from its Funk-Radon transform $\mathcal{R}f$ by its inversion, the numerical inversion being notoriously ill-posed. The question appears how can a one-to-one relationship exist between the function f defined on the three-dimensional manifold $\text{SO}(3)$ and its Funk-Radon transform $\mathcal{R}f$ defined on the four-dimensional manifold $\mathbb{S}^2 \times \mathbb{S}^2$. The affirmative answer is that the Funk-Radon transform is governed by a Darboux-type differential equation, the ultra-hyperbolic differential equation

$$(\Delta_{\mathbf{h}} - \Delta_{\mathbf{r}})\mathcal{R}f(\mathbf{h}, \mathbf{r}) = 0, \quad (8)$$

where Δ denotes the (spherical) Laplace-Beltrami operator and the subscript denotes the variable with respect to which the Laplace-Beltrami operator is applied (Savolova 1994).

Taking means of the Funk-Radon transform, e.g.,

$$\overline{\mathcal{R}f}(\mathbf{h}, \mathbf{r}) = \frac{1}{2}(\mathcal{R}f(\mathbf{h}, \mathbf{r}) + \mathcal{R}f(-\mathbf{h}, \mathbf{r})), \quad (9)$$

a one-to-one relationship between the mean $\overline{\mathcal{R}f}$ and the initial function does not exist. To recover an estimate $\hat{f}$ from $\overline{\mathcal{R}f}$ requires additional modeling assumptions which are mathematically tractable and physically reasonable, e.g., the non-negativity of f .

To account for crystal symmetry, the basic model of a crystallographic pole density function is

$$\chi f(\mathbf{h}, \mathbf{r}) = \frac{1}{\#S_{\text{Laue}} \mathbf{h}} \sum_{\mathbf{n} \in S_{\text{Laue}} \mathbf{h}} \mathcal{R}f(\mathbf{n}, \mathbf{r}), \quad (10)$$

and may be referred to as spherical crystallographic X-ray transform. Of course, it does not possess an inverse, and the inverse problem does not have a unique solution. The complete discrete mathematical model of the experimental intensities is much more involved, e.g., intensities are experimentally accessible for a few distinct crystallographic lattice planes only, effects of partial or complete superposition

of diffracted intensities from different lattice planes, e.g., complete superposition of (hkl) and (khl) of quartz, have to be accounted for by their structure coefficients, and the experimental intensities, actually counts, are usually not normalized. Their normalization requires special attention as the experimental intensities do not cover an entire hemisphere, a problem referred to as “incomplete pole figures.”

All approaches to the numerical resolution of the inverse problem involve truncated infinite or finite series expansion of the orientation probability density function f , e.g., into generalized spherical harmonics given rise to Fourier series expansion, into finite elements or splines, respectively, or into appropriate radial basis functions (Hielscher and Schaeben 2008), and their discretization. The Fourier approach requires truncation of the series and allows to estimate the even Fourier coefficients only, i.e., the kernel of the mean $\overline{\mathcal{R}f}$, Eq. 9, comprises the generalized harmonics of odd order (Matthies 1979). Thus the nonuniqueness of the inverse problem of texture analysis can be quantified that all odd Fourier coefficients remain inaccessible without additional efforts. Finite series expansion into nonnegative radial basis function employs the heuristic that their linear combination is apt to approximate any nonnegative function reasonably well. It is computationally feasible only when spherical nonequispaced fast Fourier transform (Kunis and Potts 2003, Potts et al. 2009) is applied.

Estimation with Individual Orientation Measurements

An EBSD experiment yields Kikuchi patterns to be processed by methods of image analysis or image comparison to result eventually in a set of spatially referenced representatives $g(x_\ell) = g_\ell \in \text{SO}(3)$, $x_\ell \in \mathbb{R}^2$, $\ell = 1, \dots, n$, of crystallographic orientations in $\text{SO}(3)/\tilde{S}_{\text{Laue}}$. Then an orientation probability density function is estimated by nonparametric kernel density estimation, i.e., the superposition of a radial basis function centered at the data g_ℓ , $\ell = 1, \dots, n$,

$$\hat{f}_\kappa(g|g_1, \dots, g_n) = \frac{1}{n} \sum_{\ell=1}^n \Psi_\kappa(\omega(gg_\ell^{-1})), \quad \kappa \in \mathbb{A}, \quad (11)$$

where $(\Psi_\kappa, \kappa \in \mathbb{A})$ is a set of nonnegative radial basis functions with shape parameter $\kappa \in \mathbb{A}$, actually an approximate identity for $\kappa \rightarrow \kappa_0$. While the choice of the radial basis function Ψ is not generally critical in practical applications, the choice of its shape parameter κ controlling the width or bandwidth, respectively, is so. If the kernel is given, heuristics can be applied to optimize its shape parameter with respect to various criteria (Hielscher 2013). Kernel density estimation

assumes independent and identically distributed random orientations with finite expectation and finite variance.

Grain Modeling with Individual Orientation Measurements

Even though the orientation probability density function and its estimate, respectively, referring to a set of crystallographic grains are basic entities of polycrystalline materials, there are many more entities rather referring to grain boundaries or individual grains which are instrumental to describe and quantify features of the fabric of polycrystalline materials.

Modeling the geometry and topology of grains applies heuristic modeling assumptions, some kind of classification of individual orientation measurements into classes of similar orientations, a geometrical segmentation (partition) of the orientation map image, and a data model to represent the geometry and topology of the grains and their boundaries. The basic heuristic modeling assumption is the user's definition of an angular threshold of the misorientation angle of two crystallographic orientations. If the actual misorientation angle exceeds the threshold, the orientations are considered to be different, possibly stemming from different grains and indicating a grain boundary if their locations are adjacent. Straightforward thresholding of pixel-to-pixel misorientation angles does not yield unique grains. There are quite a few numerical approaches to grain modeling. They may differ by the order of procedural steps, e.g., commencing with the segmentation of the orientation map image and then proceeding to threshold controlled amalgamation of the initial segments to crystallographic grains, or the other way round, prioritizing the clustering of similar orientations followed by geometrical modeling. An example of the former approach features the Dirichlet-Thiessen-Voronoi partition of the map image into polygonal cells centered at the locations of the measurements. This partition is based on the modeling assumption that grain boundaries are located at the bisectors of adjacent measurement locations. It immediately implies that grains are composed of adjacent Dirichlet-Thiessen-Voronoi cells (Bachmann et al. 2011). An example of the latter approach is generalized fast multiscale clustering resulting in grains composed of adjacent pixels (McMahon et al. 2013). Both methods apply graph theory to represent the grain model.

Orientation Statistics

It appears tempting to apply statistical analysis and parametric model fit to individual crystallographic orientation data. Statistics of orientations may be seen as generalization of statistics of directions and axes (Downs 1972). The generalization seems straightforward when orientations are thought of in

terms of unit quaternions. The Bingham distribution applies to axes in $\mathbb{S}^2$ as well as to pairs of antipodal quaternions in $\mathbb{S}^3$ (Kunze and Schaebe 2004). Likewise, the Fisher distribution of directions in $\mathbb{S}^2$ can be generalized to the Fisher matrix distribution of matrices in $\text{SO}(3)$. However, the distributions known from spherical statistics do not apply to crystallographic orientations as they cannot represent any crystal symmetry. Symmetry could be imposed on a quaternion or matrix distribution by superposition corresponding to the crystal symmetry class, but they would not be exponential and therefore lose all the distinct properties of exponential distributions like the Bingham or Fisher distribution. Nevertheless, they may be useful to distinguish patterns of preferred crystallographic orientation in terms of their parameters. The effects of mixing by superposition of a distribution are small if the distribution is highly concentrated. High concentrations can be expected if the measurements refer to an individual grain. Then inferential statistics is possible to distinguish the geometrical shape of highly concentrated clusters of crystallographic orientations (Bachmann et al. 2010). Otherwise, native models for ambiguous rotations are required (Arnold et al. 2018).

Estimation of Macroscopic Directional Properties

Given an anisotropic antipodally symmetric property of a single crystal $E_c(\pm \mathbf{h})$ depending on the crystallographic axis $\pm \mathbf{h} \in \mathbb{S}^2$, the corresponding macroscopic property of a specimen $E_m(\mathbf{r})$ depending on the macroscopic direction $\mathbf{r} \in \mathbb{S}^2$ can be computed given the orientation density function f or the even pole density function, respectively,

$$E_m(\mathbf{r}) = \frac{1}{4\pi^2} \int_{\mathbb{S}^2} E_c(\mathbf{h}) \chi f(\mathbf{h}, \mathbf{r}) d\mathbf{h}. \quad (12)$$

For a much more detailed presentation of the topic, the reader is referred to (Mainprice et al. 2011).

Applications of Texture Analysis

Typical Applications in General

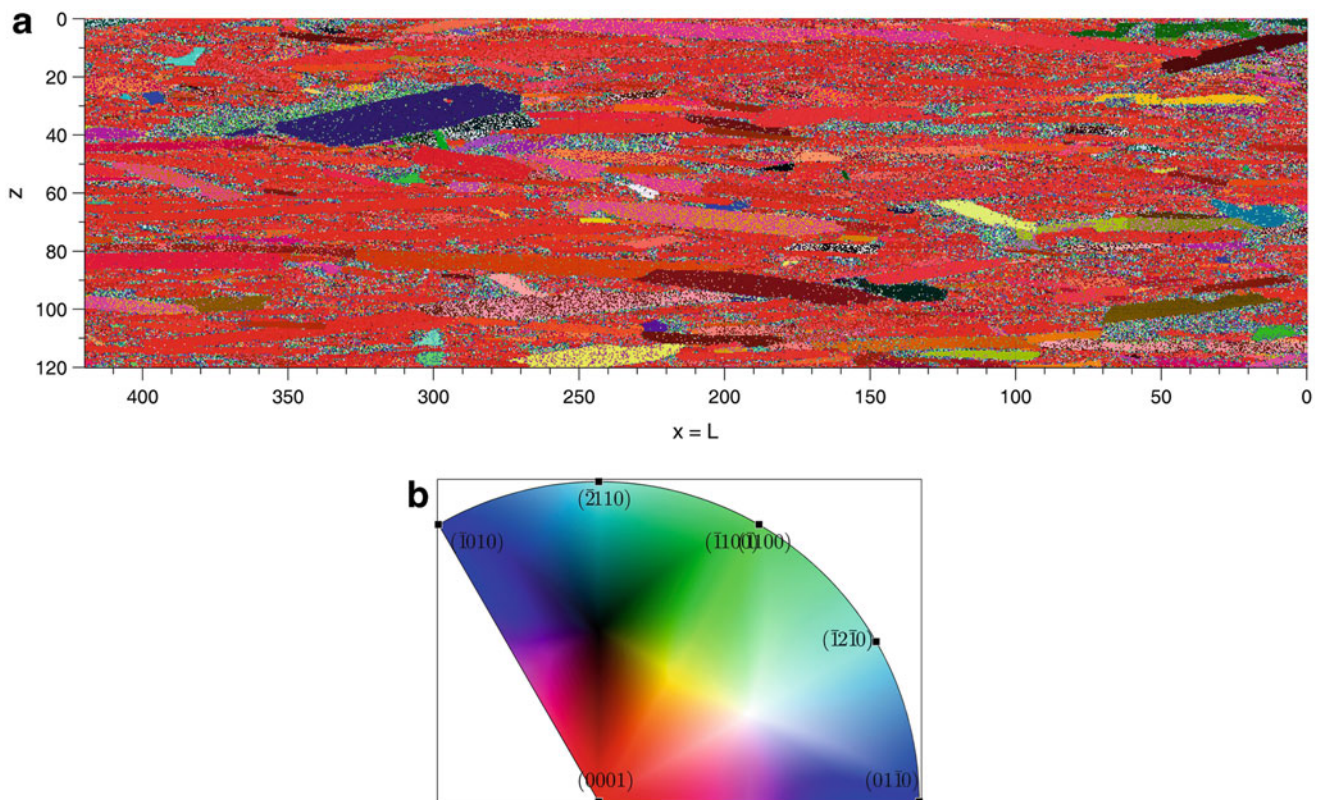
In terms of their objective, applications of texture analysis in materials science or geosciences are very different. The prevailing processes forming texture, i.e., distinct patterns of crystallographic preferred orientation, are plastic deformation, recrystallization, and phase transformations. In materials science, texture analysis typically addresses “forward” problems, like what pattern of crystallographic preferred orientation is caused by a given process, and refers to process control in the laboratory or quality control in production to guarantee a required texture and corresponding macroscopic physical properties, e.g., isotropic steel, i.e., almost perfect uniform

distribution, no crystallographic preferred orientation, or high-temperature semiconductors, i.e., an almost perfect “single crystal” crystallographic preferred orientation.

In geosciences, texture analysis is typically applied to the much more difficult “inverse” problem to identify process(es) of a sequence of geological processes which may have caused an observed pattern of crystallographic preferred orientation in rocks or ice. A result of geological texture analysis could be to exclude a geological process or event. In any case, geological texture analysis aims at an interpretation of the kinematics and dynamics of geological processes contributing to a consistent reconstruction, e.g., of the geological deformation history. Geological texture analysis provides otherwise inaccessible insight and proceeding understanding, e.g., differences in the velocity of seismic waves along or across ocean ridges have been explained with textures changes during mantle convection (Almqvist and Mainprice 2017), varying texture may result in a seismic reflector (Dawson and Wenk 2000), and the texture of marble slabs employed as building facades or tombstone decoration is thought to influence the spectacular phenomena of bending, fracturing, spalling, and shattering of the initially intact slabs (Shushakova et al. 2013).

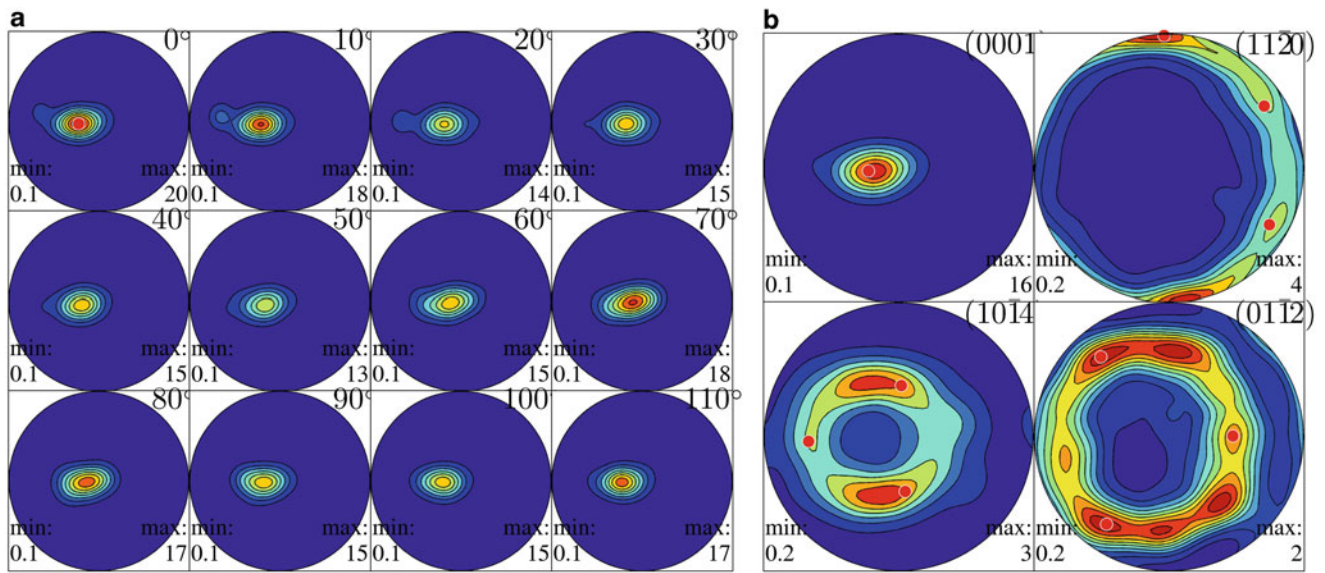
Texture of a Hematite Specimen from the Pau Branco Deposit, Quadrilátero Ferrífero (Iron Quadrangle), Brazil

Some typical steps of an EBSD data analysis have been presented with respect to hematite from a high-grade schistose ore sample from the Pau Branco deposit, Quadrilátero Ferrífero (Iron Quadrangle), Brazil (Schaebe et al. 2012). They include estimation of an orientation probability density function and modeling of grains, and in particular their interplay. A raw orientation map image from EBSD with the Pau Branco specimen is shown in Fig. 1, where crystallographic orientations are assigned to pixels which in turn are color-coded according to the colors defined in the inverse z -pole figure. The initially estimated orientation probability density function and its corresponding pole figures show a pronounced deviation from the familiar symmetric pattern of crystallographic preferred orientation of hematite (Fig. 2). The obvious asymmetry is caused by one single grain, which is peculiar both by size and by crystallographic orientation (Fig. 3). Excluding this particular grain from the orientation probability density estimation yields the expected symmetric pattern (Figs. 4 and 5).



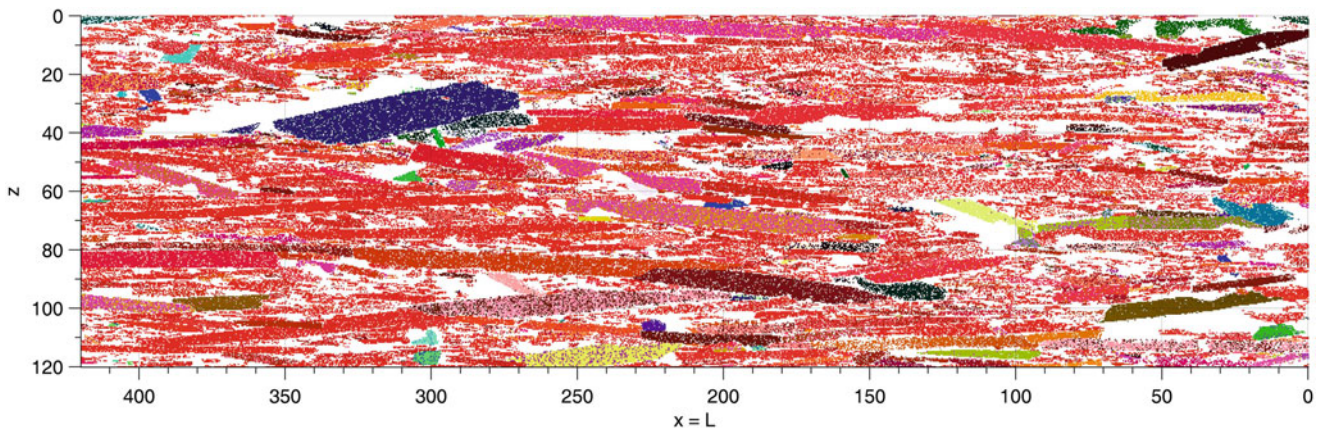
Crystallographic Preferred Orientation, Fig. 1 (a) Raw orientation map image of 316,050 spatially referenced individual orientation measurements with EBSD (left), where pixels supporting orientation g are

(b) color coded according to $g^{-1}z$ in the z -axis inverse pole figure (right), (Schaebe et al. 2012)



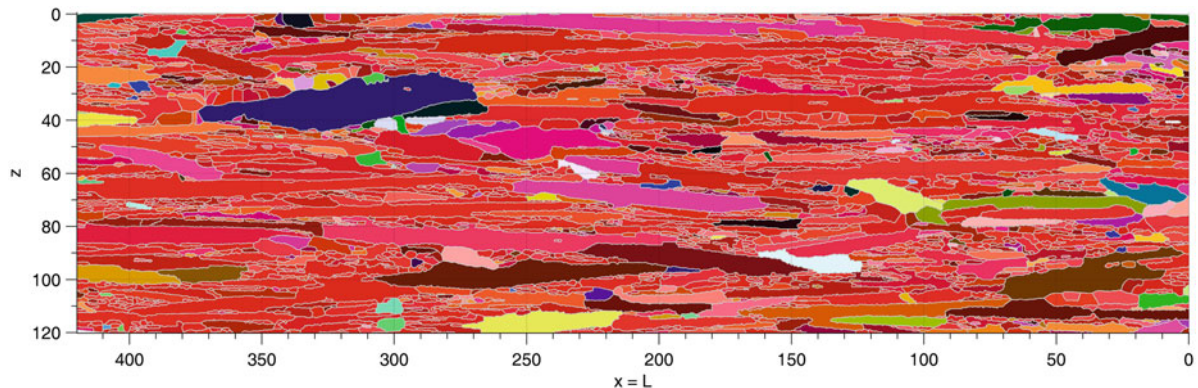
Crystallographic Preferred Orientation, Fig. 2 (a) ODF of individual orientations displayed in σ -sections (left) revealing a conspicuous asymmetry best visible in the $\sigma = 10^\circ$ section (left); (b) corresponding

PDFs for crystal forms of special interest (right). The dots mark the modal orientation g_{modal} and their corresponding $r = g_{\text{modal}}h$, (Schaeben et al. 2012)



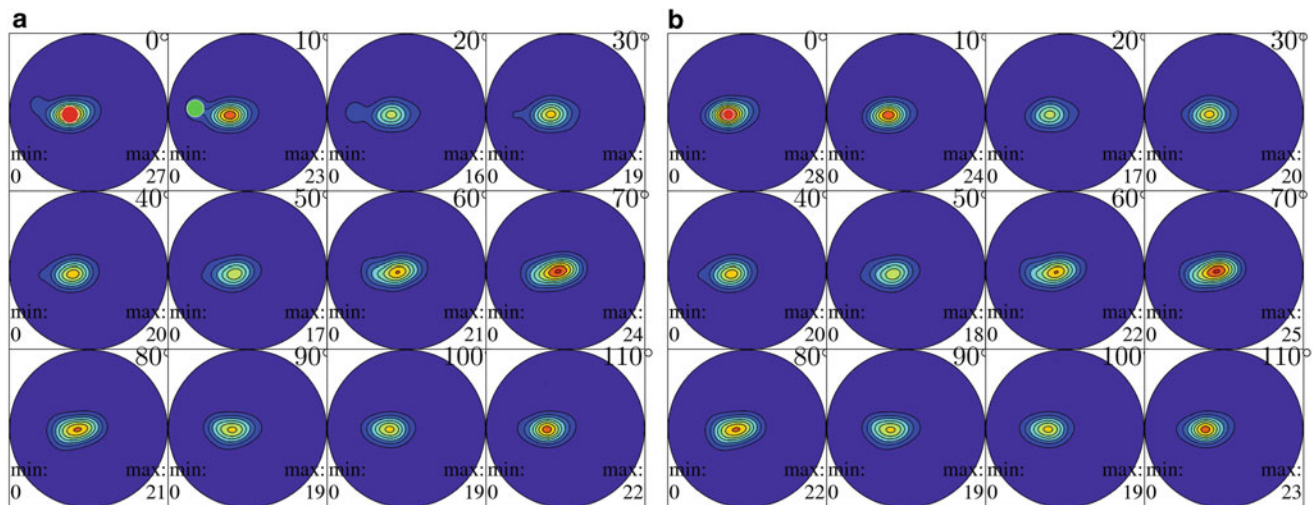
Crystallographic Preferred Orientation, Fig. 3 Orientation map image of spatially referenced individual orientation measurements with EBSD after corrections involving confidence index (CI), forward scatter

detector signal (SEM), and grain size. The vaguely visible pattern of vertical lines may be reminiscent of specimen preparation (left); (Schaeben et al. 2012)



Crystallographic Preferred Orientation, Fig. 4 Orientation map image displaying grains with respect to an angular threshold of 10° of misorientation and their mean orientation. It should be noted that there is

a conspicuous grain in terms of its blue color coding a peculiar orientation and its size in the upper left of the image, which accounts for 2.4% of the total surface area of the specimen (Schaeben et al. 2012)



Crystallographic Preferred Orientation, Fig. 5 (a) ODF of individual orientations displayed in σ -sections augmented by the mean orientation of conspicuous grain 829 as a green dot visible in the $\sigma = 10^\circ$ section

(left); (b) σ -sections of a recalculated ODF excluding grain 829 and all its properties, augmented by the modal orientation visible in the $\sigma = 10^\circ$ section (right) (Schaeben et al. 2012)

Conclusions

Texture, i.e., a pattern of preferred crystallographic orientation, provides the first order relation of a microscopic anisotropic property of a single crystal to the corresponding macroscopic anisotropic property of the polycrystalline material. Thus, texture is essential in the design of high-tech materials as well as in the recognition of texture-forming processes.

Bibliography

- Adams BL, Wright SI, Kunze K (1993) Metall Mater Trans A 24:819
 Almqvist BSG, Mainprice D (2017) Rev Geophys 55:367
 Arnold R, Jupp PE, Schaeben H (2018) J Multivar Anal 165:73
 Bachmann F, Hielscher R, Jupp PE, Pantleon W, Schaeben H, Wegert E (2010) J Appl Crystallogr 43:1338
 Bachmann F, Hielscher R, Schaeben H (2011) Ultramicroscopy 111:1720
 Bernstein S, Schaeben H (2005) Math Methods Appl Sci 28:1269
 Bunge HJ (1965) Z Metallk 56:872
 Dawson PR, Wenk H-R (2000) Phil Mag A 80:573
 Downs TD (1972) Biometrika 59:665
 Helgason S (1999) The radon transform, 2nd edn. Birkhäuser, Boston
 Hielscher R (2013) J Multivar Anal 119:119
 Hielscher R, Schaeben H (2008) J Appl Crystallogr 41:1024
 Kunis S, Potts D (2003) J Comput Appl Math 161:75
 Kunze K, Schaeben H (2004) Math Geol 36:917
 Mainprice D, Hielscher R, Schaeben H (2011) Calculating anisotropic physical properties from texture data using the MTEX open source package. In: Prior DJ, Rutter EH, Tatham DJ, (eds) Deformation mechanisms, rheology and tectonics: microstructures, mechanics and anisotropy, vol 360. Geological Society, Special Publications, London, p 175192. <https://doi.org/10.1144/SP360.10>. <http://sp.lyellcollection.org/content/360/1/175.full>
 Matthies S (1979) Phys Stat Sol A 92:K135
 McMahon C, Soe B, Loeb A, Vemulkar A, Ferry M, Bassman L (2013) Ultramicroscopy 133:16

- Morawiec A (2004) Orientations and rotations. Springer, Berlin
 Potts D, Prestin J, Vollrath A (2009) Numer Algorithms 52:355
 Roe RJ (1965) J Appl Phys 36:2024
 Sander B (1950) Einführung in die Gefügekunde der Geologischen Körper. Zweiter Teil Die Korngefüge. Springer
 Savyolova TI (1994) In: Bunge HJ (ed) Proceedings of the 10th international conference on textures of materials, materials science forum, vol 15762, p 419
 Schaeben H, Balzuweit K, León-García O, Rosière CA, Siemes H (2012) Z Geol Wiss 40:307
 Schwartz AJ, Kumar M, Adams BL, Field DP (2009) Electron backscatter diffraction in materials science. Springer, New York
 Shushakova V, Fuller ER Jr, Heidelbach F, Mainprice D, Siegesmund S (2013) Environ Earth Sci 69:1281. <https://doi.org/10.1007/s12665-013-2406-z>
 Wassermann G, Grewen J (1962) Texturen metallischer Werkstoffe, 2nd edn. Springer, Berlin/Heidelberg

Cumulative Probability Plot

Pravan Danda¹, Aditya Challa¹ and B. S. Daya Sagar²
¹Computer Science and Information Systems, APPCAIR, Birla Institute of Technology and Science Pilani, Zuari Nagar, Goa, India
²Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, Karnataka, India

Definition

A cumulative probability plot or a cumulative distribution function (see Rohatgi and Saleh 2015) is defined on real-valued random variables. For a univariate real-valued random variable, it is given by the function $F_X: \mathbb{R} \rightarrow [0, 1]$ such that

$F_X(x) = P(X \leq x)$ for each $x \in \mathbb{R}$. In the case of two or more random variables, i.e., a multivariate real-valued random variable, it is given by the function $F_{X_1, \dots, X_n} : \mathbb{R}^n \rightarrow [0, 1]$ such that $F_{X_1, \dots, X_n}(x_1, \dots, x_n) = P(X_1 \leq x_1, \dots, X_n \leq x_n)$.

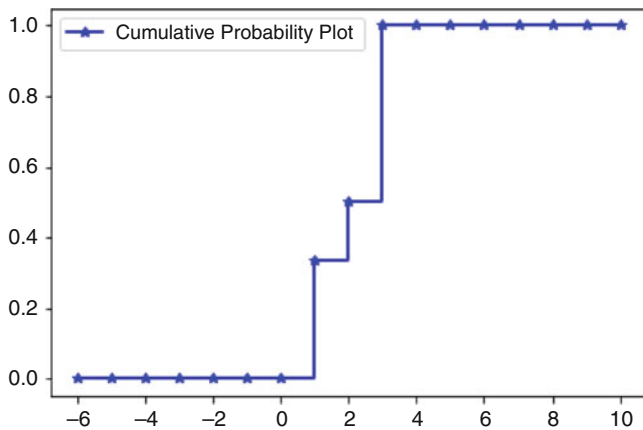
Illustration

As the name suggests, for every real number, a cumulative probability plot accumulates the probability that the random variable takes any value less than or equal to the real number. For example, let X be a univariate random variable that takes values 1 with probability $\frac{1}{3}$, 2 with probability $\frac{1}{6}$, and 3 with probability $\frac{1}{2}$. The cumulative distribution function of X is illustrated by Fig. 1. In the case of multivariate distributions, a similar assumption holds, i.e., the accumulation occurs in each random variable separately.

Relation to Probability Density and Probability Mass Functions

In the case of discrete univariate probability distributions, the cumulative distribution function can be written as a possibly infinite summation (as the support of the probability distribution, i.e., the number of values which the random variable attains with positive probability can be countably infinite) of probabilities or the probability mass function (see Rohatgi and Saleh 2015) as follows:

$$F_X(x) = \sum_{x_i \leq x} p(x_i) \quad (1)$$



Cumulative Probability Plot, Fig. 1 The cumulative distribution function of the discrete distribution given in the text is plotted here. The function is discontinuous at three points viz. at $x = 1$, and $x = 3$. The corresponding y values of the cumulative distribution function at these points are identified by marking with a star

where $p(x_i)$ denotes the probability that $X = x_i$ and the summation is over all $x_i \leq x$ such that $p(x_i) > 0$.

Similarly, the cumulative distribution function can be written as an improper integral of the probability density function (see Rohatgi and Saleh 2015) in the case of continuous univariate probability distributions:

$$F_X(x) = \int_{-\infty}^x f_X(t) dt \quad (2)$$

where $f_X(\cdot)$ denotes the probability density function of X .

It is important to note that there exist probability distributions that are neither discrete nor continuous. Such distributions do not possess either probability mass function or probability density function. For example, consider a real-valued random variable Y whose cumulative distribution function is given by:

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ 1 - \frac{e^{-y}}{2} & \text{otherwise} \end{cases} \quad (3)$$

Y is neither a discrete random variable nor a continuous random variable. Fortunately, such distributions are rarely used in applications.

Some Important Properties

In this section, some important properties of a cumulative distribution function are mentioned. These properties hold for cumulative distribution function of every probability distribution irrespective of whether the probability distribution is a discrete or continuous or neither.

1. For a univariate random variable, the probability that the random variable takes values between two real numbers can be expressed as a difference of the cumulative probability plot at the end points. Formally, if $a \leq b$ are real numbers, we have

$$P(a < X \leq b) = F_X(b) - F_X(a) \quad (4)$$

In the case of multivariate distributions, a similar property holds. Let $1 \leq i \leq n$ then

$$P(X_1 \leq x_1, \dots, a < X_i \leq b, \dots, X_n \leq x_n) = F_X(x_1, \dots, b, \dots, x_n) - F_X(x_1, \dots, a, \dots, x_n) \quad (5)$$

2. For a univariate random variable, the cumulative distribution function is a nondecreasing function, i.e., if $b \geq a$ are real numbers,

$$F_X(b) \geq F_X(a) \quad (6)$$

Similarly, for a multivariate distribution, the cumulative distribution function is nondecreasing in each of its variables, i.e.,

$$F_X(x_1, \dots, b, \dots, x_n) \geq F_X(x_1, \dots, a, \dots, x_n) \quad (7)$$

3. The cumulative distribution function is right continuous in each of its variables, i.e., if $1 \leq i \leq n$ then

$$\lim_{x \rightarrow a+} F_X(x_1, \dots, x, \dots, x_n) = F_X(x_1, \dots, a, \dots, x_n) \quad (8)$$

4. The cumulative distribution function is bounded by 0 and 1. More specifically, let $1 \leq i \leq n$ then the following hold:

$$\lim_{x_1, \dots, x_n \rightarrow +\infty} F_X(x_1, \dots, x_n) = 1, \quad (9)$$

$$\lim_{x_i \rightarrow -\infty} F_X(x_1, \dots, x_n) = 0 \quad (10)$$

Uses in Statistics

A cumulative distribution function can be used to simulate data from continuous probability distributions. Suppose one can simulate a continuous uniform distribution on the closed interval $[0, 1]$ using a computer, i.e., $u_1, \dots, u_k$ are realizations of $U[0, 1]$ using a pseudorandom number generation process. Let X be a continuous random variable with a theoretically known distribution. Data from X can be generated using the inverse transformation $F_X^{-1}()$ on $u_1, \dots, u_k$, where $F_X()$ denotes the cumulative distribution function of X . In other words, $F_X^{-1}(u_1), \dots, F_X^{-1}(u_k)$ are realizations of X .

Often, in absence of reasonable assumptions on the distribution of data, nonparametric tests based on empirical estimates of the cumulative distribution function are performed to test statistical hypotheses (see Lehmann and Casella 2006;

Lehmann and Romano 2006). For example, Kolmogorov-Smirnov test (see Wasserman 2006) is widely used in applications.

Summary

In this chapter, a cumulative probability plot or a cumulative distribution function is defined for univariate and multivariate probability distributions. These definitions are then illustrated on a discrete probability distribution with finite support. The relation of cumulative probability plot to probability mass function and probability density is described. This is followed by mentioning some important properties of a cumulative probability plot. The chapter is then concluded by mentioning some of its uses in statistical analyses.

Cross-References

- [Hypothesis Testing](#)
- [Probability Density Function](#)
- [Simulation](#)

Bibliography

- Lehmann EL, Casella G (2006) Theory of point estimation. Springer Science & Business Media
- Lehmann EL, Romano JP (2006) Testing statistical hypotheses. Springer Science & Business Media
- Rohatgi VK, Saleh AME (2015) An introduction to probability and statistics. Wiley
- Wasserman L (2006) All of nonparametric statistics. Springer Science & Business Media

D

Dangermond, Jack

Lowell Kent Smith
Emeritus, University of Redlands, Redlands, CA, USA



Fig. 1 Jack Dangermond, courtesy of J. Dangermond

Biography

Jack Dangermond was born in 1945 and grew up in Redlands, California, where his parents owned a plant nursery at which he worked from an early age. His education included a B.S., Landscape Architecture, California Polytechnic College – Pomona, 1967; M.S., Urban Planning, Institute of Technology, University of Minnesota, 1968; and M.S., Landscape Architecture, Graduate School of Design, Harvard University, 1969. He studied at Harvard's Laboratory for Computer Graphics and Spatial Analysis (LCGSA) which influenced his founding in 1969, with his wife Laura, of Environmental Systems Research Institute, (now Esri) and also influenced the later development of Esri's ARC/INFO software.

The Dangermonds' vision for Esri was to develop Geographic Information Systems (GIS) software, demonstrate its usefulness by applying it to real problems, and then make it freely available. When this model proved financially infeasible, Esri was incorporated as a privately held, profit-making business; but Esri continues to be the means to realize their vision.

In the 1970s, Esri developed PIOS, the Polygon Information Overlay System, and GRID for raster data, and applied these in project-focused consulting. Esri next developed ARC/INFO (1982); since its release, Esri has focused primarily on GIS software development, leading to ArcGIS (1999 to the present). In 50 years, Esri has grown from a small research institute to a company with a staff of 4000 in 11 research centers and 49 offices around the world. Jack is a leading advocate for GIS, making thousands of presentations to diverse audiences all over the world.

Supporting geoscience, Esri provides research grants and inexpensive software licenses to higher education, and in 2014 donated a billion dollars of software to US schools as part of the ConnectEd initiative. Esri Press publishes GIS-related technical and educational titles. Esri offers online courses and teaching materials. More than 7000 colleges and universities use and teach ArcGIS.

Jack's professional service has included: Committee on Geography, US National Academy of Sciences; National Geographic Society Board of Trustees; National Geospatial Advisory Committee, NGAC; Earth System Science and Applications Advisory Committee, NASA; Science and Technology Advisory Committee, NASA; National Steering Committee, Task Force on National Digital Cartographic Standards; and Executive Board, National Center for Geographic Information and Analysis (NCGIA).

Among the recognitions of his accomplishments are Officer in the Order of Orange Nassau, Netherlands, 2019; Audubon Medal, The National Audubon Society, 2015; Alexander Graham Bell Medal, National Geographic Society, 2010; Patron's Medal, Royal Geographical Society, 2010; Carl Mannerfelt

Medal, International Cartographic Association, 2008; Fellow of ITC, The Netherlands, 2004; Brock Gold Medal for Outstanding Achievements in the Evolution of Spatial Information Sciences, International Society for Photogrammetry and Remote Sensing, 2000; Cullum Geographical Medal of Distinction for the advancement of geographical science, American Geographical Society, 1999; and James R. Anderson Medal of Honor in Applied Geography, Association of American Geographers, 1998. Jack has received 13 honorary doctorates, including: University of Massachusetts, Boston, Massachusetts, 2013; University of Minnesota, Minneapolis, 2008; California Polytechnic University – Pomona, 2005; State University of New York – Buffalo, 2005; and City University of London, 2002.

Data Acquisition

Rajendra Mohan Panda¹ and B. S. Daya Sagar²

¹Geosystems Research Institute, Mississippi State University, Starkville, MS, USA

²Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, Karnataka, India

Definition

Data acquisition is a process to collect and store information for production use. Technically, data acquisition is a process of sampling signals of real-world physical phenomena and their digital transformation.

Terminology

Analog-to-digital converter (ADC) prepares sensor signals in the digital form.

Data loggers are independent data acquirers.

Digital-to-analog (D/A) converter produces analog output signals.

Electromagnetic spectrum is the wavelengths or frequencies extending from gamma to radio waves.

Resolution is the smallest incremental unit of a signal in the data acquisition system. Resolution is either expressed in bits or percent of full scale, where one part in 4096 resolution represents 0.0244% of full scale.

Radiometers are the instruments used to measure the intensity of electromagnetic radiation in select bands.

Sample rate is the speed at which a data acquisition system gathers data. Unit: sample/s

Sensors or transducers convert physical phenomena to electric signals.

Spectrometers are devices designed to detect, measure, and analyze the spectral phenomena of reflected electromagnetic radiation.

Signal-molding hardware shapes sensor signals to digital forms in an analog-to-digital converter.

Signal conditioning or transmission is an optimization process that modifies sensor signals to a usable form.

Introduction

Data acquisition is a process to collect and store information for production use. Data are newly collected, legacy data, shared, and purchased data. These data include sensor-generated data, empirical data, and other sources. This data acquisition concept started around the early seventies, with the development of software by the International Business Management (IBM) company solely dedicated to data collection. The computing abilities for data acquisition have been improved immensely with better storage capacities and faster processing. Data acquisition involves signal reception and digitization for storage and analysis using computers (Meet et al. 2018). The data acquisition (DAQ) system has three components (OMEGA):

Sensor → Analog-to-digital-conversion → Transmission

Sensors transduce physical values to electric signals, and the sensor-molding hardware converts them to digital values. These digitally modified values undergo noise correction and optimize for several implications. Several programming languages used for obtaining such signals are Assembly, BASIC, C, C++, Fortran, Java, LabVIEW, Lisp, and Pascal. There is also stand-alone DAQ hardware that can process electric signals without a computer, e.g., data loggers and oscilloscopes. The DAQ software processes data from the DAQ hardware to outputs for meaningful uses.

Data acquisition devices are equipped with signal conditioning and analog-to-digital converter. These devices can be wired or wireless and need a computer for data transmission. The wired data acquisition process can be single-ended or differential inputs. Interface buses like RS232 and RS485 are popular DAQ devices. RS232 is suitable for small businesses but has limitations of supporting one device for serial communication with the transmissibility of 50 ft. Alternately, RS485 has options for multiple device connectivity and transmission support for a distance of 5000 ft. General Purpose Interface Bus (GPIB) or IEEE 488 (as per ANSI protocol) are standard interface buses. The Universal Serial Bus (USB), Ethernet, and PCI are other data acquisition devices, with USB being advantageous for its higher bandwidth and power supply capability. Data acquisition devices vary depending on the data collection speed of the analog-to-digital

converter. This speed is the number of channels sampled per second and expressed as the *sample rate*. DAQ cards or boards increase the sample rate by directly connecting to the bus. In addition, modular data acquisition systems, designed for a high channel count, integrate and synchronize multiple sensors with multiple applications to increase the speed of data collection. The PXI is a popular modular system with tremendous flexibility. It has options of channel count and is cost-effective. This modular system is easy to use and is capable of fast data acquisition. Some data acquisition tools for specific uses by the users are WINDAQ, ActiveX, DATAQ, Flukecal, MSSATRLABS, DAQami, TracerDAQ, DASYPAL, and LabVIEW.

Remote Sensing Data Acquisition

Principles

Remote sensing is a process of data acquisition on earth or other planetary systems. It involves data capture and interpretation of the electromagnetic spectrum without making physical contact with the object. It works on the principle of the law of conservation of energy, i.e., total energy to be conserved over time (Feynman 1970):

$$\text{Incident energy} = \text{Reflected energy} + \text{Absorbed energy} + \text{Transmitted energy}$$

Types

Usually, remote sensing satellites follow three orbital paths: polar, nonpolar, and geostationary (Zhu et al. 2018). *Polar satellites* are placed about 90° above the equatorial plane to procure information about the entire globe and polar regions. These are sun-synchronous satellites appearing at the exact location at the same time for every cycle. They move from south to north and north to south, which explains their ascending and descending nature.

Nonpolar satellites are at lower orbits, i.e., 2000 km from the earth's ground. These satellites provide partial coverage, for example, the Global Precipitation Measurement (GPM) mission satellite covers 65°N to 65°S latitude.

Geostationary satellites are placed above 36,000 km of the earth's surface, very powerful in weather forecasting. These satellites follow the earth's rotation at the same speed the earth rotates. Due to the same reason, these satellites capture the same view of earth's particular area with each observation in a regular fashion.

Sensors

The chief component of remote sensing data acquisition is the sensors, which can be active or passive. *Passive sensors* are capable of detecting solar radiation that is reflected or emitted by the object or scene being observed (NASA) and have radiometers and spectrometers operating in the visible

(400–700 nm), infrared (700–3000 nm), thermal infrared (3000 nm–1 mm), and microwave (1 mm–300 cm) portions of the electromagnetic spectrum (Elert 1998). These sensors are useful in procuring information from land and sea surface, vegetation, cloud, aerosol, and other physical phenomena. Passive sensors are not efficient in penetrating dense clouds, which active sensors can fulfill.

Active sensors do not depend on sunlight for emissivity. These sensors primarily operate in the microwave portion of the electromagnetic spectrum having radio detection and ranging (radar) sensors, altimeters, and scatterometers. Active sensors measure the vertical profiles of aerosols, forest structure, precipitation and winds, sea surface topography, and snow cover. Synthetic aperture radar (SAR) is a well-known active sensor, which provides cloud-free data. Its sensors are capable of collecting data day and night (Woldai 2004).

Resolution

The remote sensing data include multispectral, hyperspectral, optical, Infrared, thermal, and microwave imageries, and their quality depend on four resolution categories: radiometric, spatial, spectral, and temporal (NASA). The satellite resolution varies with the orbits and the type of sensors.

The radiometric resolution provides energy information of a pixel in bits. The higher the bit value, the higher the pixel information, and the higher the radiometric accuracy of the sensor (Liang and Wang 2019). For example, an 8-bit resolution has an extent of digital numbers from 0 to 255 against an 11-bit from 0 to 2047 and so on.

Spatial resolution represents the pixel size of a digital image that explains the minimum area coverage of the satellite sensor on the ground. For example, a satellite sensor at a 1-km resolution has a pixel size of 1 km × 1 km. The pixel is a function of orbital altitude, radiation angle, sensor size, and focal length of the optical system. *Swath width* corresponds to the total field of view on the ground. Satellites have different spatial resolutions to accomplish specific purposes (Table 1). Based on pixel size, satellite sensors have four major spatial resolution categories: (a) coarse (>1 Km), (b) medium (100 m–1 km), (c) high (5–100 m), and (d) very high (<5 m). The lesser the pixel size, the higher the image information and vice-versa. The image quality reduces from a finer to a coarser spatial resolution and vice versa (Fig. 1).

Spectral resolution represents the number of spectral bands captured by a sensor. It helps users to differentiate objects by identifying their spectral signature. Spectral signature is the unique feature of an object expressed as a function of its reflectance or emittance to the wavelengths of the electromagnetic spectrum. The digital images are assigned separate classes of similar pixels by comparing the known spectral signatures in *supervised classification*. Some popular algorithms used for supervised classification are maximum likelihood, minimum distance, principal component analysis, support vector machine, and random forest. Alternately,

Data Acquisition, Table 1 List of popular remote sensing satellites, their sensors, and resolution. The bold letters represent radar satellites, and the numbers in parentheses indicate the radiometric resolution of the satellite

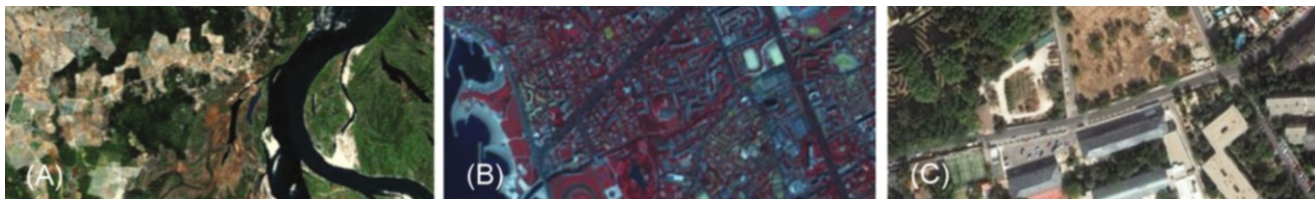
Satellite	Sensor	Spatial resolution (m)	Temporal resolution	Operator
AISat-1B	ALITE	12 (PAN), 14 (MS)	7	ASAL/CNES
AISat-2A/2B	NAOMI	2.5 (PAN), 10 (MS)		ASAL/CNES
ALOS-2 or DAICHI2	PALSAR2	1 (PAN), 3 (PAN), 6 (PAN), 10 (PAN), 100 (PAN), HH, HV, VV, HH/HV, VV/VH HH/HV/VV/VH	14	JAXA
ASNARO-2 (8)	XSAR	1 (PAN), 2 (PAN), 16 (PAN), HH/VV	14	NEC/USEF
ASTER (8, 12)	ASTER	15 (MS), 30 (MS), 90 (MS)	16	NASA
NOAA 19 (10)	HIRS	1000 (MS)	Daily	NOAA
Capella 1/Denali	SAR	0.5 (PAN), 1.7 (PAN), 7 (PAN), HH, VV	Hourly	Capella Space of California
Cartosat-1 (10)	PAN-FORE, PAN-AFT	2.5 (PAN)	5	ISRO
Cartosat 2A, 2B	PAN	<1 (PAN)	4	ISRO
Cartosat 2C, 2D, 2E, 2F	PAN, HRMS	0.65 (PAN), 2 (MS)	4	ISRO
Cartosat-3 (11)	PAN, MSI	0.28 (PAN), 1.12 (MS)	4	ISRO
CBERS-4	IRMSS, MUX, PANMUX, WFI	5 (PAN), 10 (MS), 20 (MS), 40 (MS), 64 (MS), 80 (MS)	26	CNSA/NISR
COSMO-SkyMed 1-4	SAR 2000	1 (PAN), 5 (PAN), 15 (PAN), 30 (PAN), 100 (PAN), HH, VV, HV, VH, VV, HH/VV, HH/HV, VV/VH	16	Thales Alenia Space
Dove Pioneer, Flock 2k, 2p, 3m, 3p	Planet Scope-2	4 (PAN)		Planet
EROS-B (11)		0.7 (PAN)	4	ImageSat International
Gaofen-2	PMC-2	0.8 (PAN), 3.2 (MS)		CNSA
Gaofen-3	SAR C	1 (PAN), 3 (PAN), 5 (PAN), 8 (PAN), 10 (PAN), 25 (PAN), 50 (PAN), 100 (PAN), 500 (PAN)		CNSA
GeoEye1 (11)	GIS	1.41 (PAN), 1.65 (MS)	2.1–8.3	DigitalGlobe (Maxar)
GRUS-1A		2.5 (PAN), 5 (MS)		Axelspace
IKONOS (11)		0.82–1 (PAN), 3.2–4 (MS)	3	DigitalGlobe (Maxar)
KOMPSAT-2	MSC	1 (PAN), 4 (MS)	14	KARI
KOMPSAT-3	AEISS	0.7 (PAN), 2.8 (MS)	14	KARI
KOMPSAT-3A	AEISS-A, IIS	0.55 (PAN), 2.2 (MS)	14	KARI
KOMPSAT-5	COSI	1 (PAN), 3 (PAN), 20 (PAN), HH, HV, VH, VV	14	KARI
Landsat-7 (8)	ETM+	15 (PAN), 30 (PAN)	16	NASA/USGS
Landsat-8 (12)	OLI, TIRS	15 (PAN), 30 (MS), 60 (MS)	16	NASA/USGS
Landsat-9 (14)	OLI2, TIRS2	15 (PAN), 30 (MS), 100 (MS)	16	NASA/USGS
MODIS (Terra or Aqua) (12)	MODIS	250 (MS), 500 (MS), 1000 (MS)	Daily	NASA
NovaSAR-S1	S-SAR	6 (PAN), 20 (PAN), 30 (PAN), 33 (PAN), 35 (PAN), 40 (PAN), 45 (PAN), 400 (PAN)	14	Surrey Satellite Technology Ltd.
PAZ	SAR-X	1 (PAN), 2 (PAN), 3 (PAN), 6 (PAN), 16 (PAN)	11	
Pléiades 1A & 1B	HiRI	0.5 (PAN), 2 (MS)	26	CNES
Quickbird (11)		0.65 (PAN), 2.62 (MS)	2.5	DigitalGlobe (Maxar)
RADARSAT-2	SAR	1 (PAN), 3 (PAN), 8 (PAN), 12 (PAN), 18 (PAN), 25 (PAN), 30 (PAN), 40 (PAN), 50 (PAN), 60 (PAN), 100 (PAN), HH, HV, VH, VV, HH/HH, HH/HV, HV/HH, HV/HV, Alt. HV/HV	24	CSA
RapidEye	REIS	5 (MS)	5.5	Planet

(continued)

Data Acquisition, Table 1 (continued)

Satellite	Sensor	Spatial resolution (m)	Temporal resolution	Operator
Sentinel-1	C-SAR	5 (PAN), 20 (PAN), 40 (PAN), HH, VV, HH/HV, VV/VH	6	ESA
Sentinel-2	MSI	10 (MS), 20 (MS), 60 (MS)	6	ESA
Sentinel-3	OLCI	300 (MS)	2	ESA
Sentinel-3	SLSTR	500 (MS), 1000 (MS)	Daily	ESA
SkySat 1-7	SS-MSI	0.5 (PAN), 2 (MS)	2	Planet
SPOT 5 (8)	HRG	2.5 (PAN), 5 (MS)	26	EADS Astrium/ Azercosmos
SPOT 6, 7 (8)	NOAMI	1.5 (PAN), 6 (MS)	26	EADS Astrium/ Azercosmos
SuperView1		0.2 (PAN), 0.5 (MS)		Beijing Space View Technology
TanDEM-X, TerraSR-X	SAR	1 (PAN), 3 (PAN), 16 (PAN), HH, HV, VH, VV	11	Cosmodrome, Kazhakstan
TripleSat 1-3, SSTL- S1	VHRI-100	0.8 (PAN), 3.2 (MS)	1	DMCii
Vivid-i 1-5	HRI	0.6 (PAN)	0.5	Earth-i
VNREDSat-1A (12)	NAOMI	2.5 (PAN), 10 (MS)		VAST
VRSS-1	PMC	2.5 (PAN), 10 (MS)	57	ABAE
VRSS-1	WMC	16 (PAN), 16 (MS)	57	ABAE
VRSS-2	IRC	30 (PAN), 60 (MS)	101	ABAE
VRSS-2	PMC-2	1 (PAN), 4 (MS)	101	ABAE
WorldView-1 (11)	WV-60	0.5 (PAN)	1.7	DigitalGlobe (Maxar)
WorldView-2 (11)	WorldView- 110	0.46 (PAN), 1.85 (MS)	1.1	DigitalGlobe (Maxar)
WorldView-3 (11)	WV-3 Imager	0.31 (PAN), 1.24 (MS)	1	DigitalGlobe (Maxar)

Abbreviations: *USEF* Japanese Institute for Unmanned Space Experiment Free Flyer, *JAXA* Japan Aerospace Exploration Agency, *NASA* US National Aeronautics and Space Administration, *NOAA* US National Oceanic and Atmospheric Administration, *ESA* European Space Agency, *NISR* Brazilian National Institute of Space Research, *ISRO* Indian Space Research Organization, *CNSA* Chinese National Space Administration, *ASAL* Algerian Space Agency, *KARI* Korean Space Research Institute, *VAST* Vietnam Academy of Science and Technology, *EADS* European Aeronautic Defenses and Space Company, *CNES* French Centre for Space Studies, *ABAE* Bolivarian Agency for Space Activities, *USGS* US Geological Survey

**Data Acquisition, Fig. 1** (a) Coarse resolution; (b) moderate resolution; (c) fine resolution. (Source: L3HARRIS)

images are classified based on their properties without supervision in *unsupervised classification*. Both pixel-based image classification techniques differ from the *object-based image analysis* (OBIA) which groups pixels into representative shapes to their size and geometry.

Temporal resolution is the revisit time of remote sensing satellite to acquire data for the exact location. It depends on the orbital characteristics of the satellite. The frequency of revisiting a location helps to collect information on changes in

forest cover, forest-fire events, glacier melting, flooding, deforestation, and seasonal crop pattern.

Buczowski (2020) provides a comprehensive list of remote sensing satellites and sensors including information about radiometric, spatial, and spectral resolution. More information is gathered using a google search engine and Wikipedia (Table 1). The remote sensing satellite information is categorized based on optical and radar satellites, including their operating organizations or companies. Panchromatic

Data Acquisition, Table 2 List of remote sensing data sources

Data sources	Data acquisition	Availability	Source link
USGS EarthExplorer	Landsat, spy satellites, hyperspectral	Free	https://earthexplorer.usgs.gov/
Sentinel Open Access Hub	Sentinel-2, Sentinel-1 (C-band)	Free	https://scihub.copernicus.eu/dhus/#/home; https://www.copernicus.eu/en
NASA EarthData Search	Derived data: Lland cover and land use	Free	https://search.earthdata.nasa.gov/search
NOAA Data Access Viewer	Aerial, LiDAR	Free	https://coast.noaa.gov/dataviewer/#/
Maxar Open Data Program	DigitalGlobe	Partly Free	https://www.maxar.com/open-data
Geo-Airbus Defense	SPOT, Pleiades, RapidEye	–	https://www.intelligence-airbusds.com/imagery/
NASA WorldView	Wildfires, icebergs, earthquakes	Free	https://worldview.earthdata.nasa.gov/
NOAA Class	Products for ocean, atmosphere, climate	Free	https://www.avl.class.noaa.gov/saa/products/welcome
INPE Catalog	CBRS-2	Partly Free	http://www.dgi.inpe.br/CDSR/
Bhuvan	IMS-1 (hyperspectral), Cartosat, OceanSat, ResourceSat	Partly Free	https://bhuvan-app3.nrsc.gov.in/data/download/index.php
Bhuvan	NDVI, CARTODEM	Free	https://bhuvan-app3.nrsc.gov.in/data/download/index.php
ZAXA	ALOS	Free	https://www.eorc.jaxa.jp/ALOS/en/index_e.htm
VITO Vision	PROBA-V, SPOT Vegetation, and METOP	Partly Free	https://www.vito-eodata.be/PDF/portal/Application.html#Home
NOAA Digital Coast	Benthic, elevation, land cover, socio-economic	Free	https://coast.noaa.gov/digitalcoast/
UNAVCO	SAR	Free	https://www.unavco.org/data/sar/sar.html
LAND COVER	Landsat, MODIS, and AVHRR	Free	http://maps.elie.ucl.ac.be/CCI/viewer

(PAN) and multispectral (MS) resolution is highlighted for each satellite, including revisit time. A list of data sources, available in the public domain, is summarized (Table 2).

Conclusions

Data acquisition is an important part of research and data analytics. Understanding its process and use provide valuable inputs for research and management. This is essential for efficiency enhancements in evaluating capacity development and monitoring. With rising demand and complexities in data, the different organizations, industries, and business houses continue to adopt new technological advancements in data acquisition systems. The scope is enormous with the use of big data and simultaneous multitasking.

Bibliography

- Buczowski A (2020) The most epic, detailed, and complete list of all Earth Observation satellites. <https://geoawesomeness.com/the-most-epic-detailed-and-complete-list-of-all-earth-observation-satellites/>. Accessed 2 Dec 2021
- Elert G (1998) The electromagnetic spectrum, the physics hyper-textbook. Hypertextbook.com
- Feynman R (1970) The Feynman lectures on physics, vol I. Addison Wesley. ISBN 978-0-201-02115-8

- Liang S, Wang J (2019) Advanced remote sensing: terrestrial information extraction and applications. Academic, 1–57 Pages
- Meet R, Sharma V, Patel A (2018) Remote data acquisition. Int Res J Eng Technol (IRJET) 5:12. e-ISSN: 2395-0056
- NASA. .What is remote sensing? <https://earthdata.nasa.gov/learn/backgrounders/remote-sensing>. Accessed 6 Dec 2021
- OMEGA. A complete guide to Data Acquisition (DAQ) systems. <https://www.omega.com/en-us/resources/daq-systems>. Accessed 6 Dec 2021
- Woldai T (2004) Electromagnetic energy and remote sensing. In: Norman K, Lucas LFJ, Gerrit CH (eds) Principles of remote sensing. ITC Educational Textbook Series. ITC
- Zhu L, Suomalainen J, Liu J, Hyyppä J, Kaartinen H, Haggren H (2018) A review: remote sensing sensors. In: Multi-purposeful application of geospatial data. IntechOpen, pp 19–42

Data Dictionary

Madhurima Panja, Tanujit Chakraborty and Uttam Kumar
Spatial Computing Laboratory, Center for Data Sciences,
International Institute of Information Technology Bangalore
(IIITB), Bangalore, Karnataka, India

Definition

Data dictionary stores catalog information about schemas and constraints, design decisions, usage standards, application program description, user information, etc. that can be

accessed directly by users or the database administrators when needed (Elmasri and Navathe 2000). Such a system is also called an information repository and is a record of the objects in the database (Raschka and Mirjalili 2019). In technological domain, these objects are referred to as metadata. As per the IBM Dictionary of Computing, data dictionary is defined as “centralized repository containing information of the data in the database such that the meaning, relationship, source of the data, where it will be used and the format is clearly mentioned or specified” (McDaniel 1994). In other words, a data dictionary is often termed as a data definition matrix, which is a textual description of data objects and their interrelationships. It is commonly used in confirming data requirements and for database developers to create and maintain a database system. More precisely, a data dictionary describes the physical attributes of the elements of a database such as meaning, relationships to other data, responsibility, origin, usage, allowable values, and format. A data dictionary utility is similar to the database management system (DBMS) catalog, but it includes a wider variety of information and is accessed mainly by users rather than by the DBMS software. Modern businesses rely on database systems that have built-in active data dictionaries and documentations generated with rational DBMSs (e.g., SQL, Oracle, MySQL).

Introduction

In the era of digitization, data is a strategic resource used for scientific research, business management, design of public policies, and many forms of human organizational activity (Batini et al. 1990). A large volume of data generated in the real-world activities falls under the umbrella of big data due to their volume, velocity, and variety (also known as 3Vs of big data). Data collected from the space falls under the category of Earth and astronomical data. This large volume of dataset is creating new opportunities, revolutionizing the innovation of methodologies and thought patterns in the geographic information science (GIS) domain. It is forming the foundation of basic and applied geoscience research and has the potential to underpin industry programs to discover and develop domestic natural resources in order to fulfill the world’s energy and mineral requirements. By modeling geoscience data, scientists can provide improved predictions of both immediate and long-term potential hazards. The geoscience community has gathered an enormous amount of valuable data, much of which is irreplaceable. These data are a trove of untapped resources awaiting scientists, engineers, educators, and policymakers to consolidate and use the information. The potential usefulness of these data has continued to expand, so has the need for space/storage to support the data preservation. Since the ability to preserve the spatial data has not been done at the same pace as that of its acquisition, various

nations are facing the problem of irretrievability of such valuable information. To overcome this problem of data loss, majority of the data are now stored in a centralized DBMS. The metadata of these DBMS, which include supporting information about the database, are then provided through a well-documented data dictionary. Data dictionary is a repository of all types of data produced, managed, exchanged, and maintained in a database represented in a way that allows all users of the information system to understand their meanings.

In other words, database system maintains a description of all the data that it contains. For example, a relational DBMS maintains information about every relation and index that it contains. Similarly, a DBMS also maintains information about views, for which no tuples are stored explicitly. Actually, a definition of the view is stored and is used to compute the tuples that belong in the view when it is queried. This information is stored in a collection of relations maintained by the system called *catalog relations* or *system catalog* or *data dictionary* or *metadata*, that is, not data, but descriptive information about the data. The information in the data dictionary that is metadata about the structure and schema of the database is used for query optimization (Ramakrishna and Gehrke 2000). The data dictionary may also note the storage organization (sequential, hash or heap) of relations, and the location where each relation is stored (Silberschatz et al. 2006). So, a data dictionary is a system database in its own right rather than a user database containing data about the data. All of the various schemas and mapping (external, conceptual, etc.) and various security and integrity constraints are stored in the metadata (Date 2000).

These dictionaries make users aware of various aspects of the data and reduce their dependencies on the database managers/administrators. In the absence of a data dictionary, a user with no prior knowledge might lose the data indefinitely as they may not know the precise location of the data in the database, and ironically this makes it an inevitable part of the database. Therefore, a data dictionary consists of names, definitions, and attributes of data elements to be used or captured in a database, information system, or part of a scientific research. A data dictionary also provides metadata about data elements. The usefulness of data dictionary could be the following among many others:

1. Assisting in providing data consistencies during the project.
2. Using data by multiple users or members of the project.
3. Making data easier to handle and analyze.
4. Enforcing the use of data standards.
5. Cataloging and communicating the structure and content of data among all the users.

Types of Data Dictionary

Data dictionaries can be broadly classified into two main categories – active data dictionary and passive data dictionary – with different purposes.

Active Data Dictionary

Sometimes the structure of the database has to be changed; these changes include removing some old attributes and adding some new ones. If these changes are updated automatically in the data dictionary by DBMS, then the data dictionary is said to be active. Active data dictionary is also known as integrated data dictionary.

Passive Data Dictionary

The data dictionary that is not part of and managed by DBMS is called a passive data dictionary. In this dictionary, any changes in the database are updated manually or by a directed software. These are also known as nonintegrated data dictionary. There are different types of passive data dictionary, which are described next.

Document or spreadsheet: The simplest database technology, spreadsheet, is often used for creating a data dictionary. The Excel data dictionaries are mostly considered to be a passive one due to the lack of built-in technology that can automate database-to-data dictionary encoding. However, one can automate this encoding for some sections of the dictionary using a user-defined function.

Data catalog: A data catalog provides an illustration of the metadata. It is quite user-friendly than any active data dictionary, however, building such a dictionary is quite laborious.

ETL metadata repositories: This data dictionary is created by extract, transform and load (ETL) technique, i.e., extracting data from multiple sources, transforming them to a similar format and finally, loading them to a central location. Since the abovementioned procedure requires human effort, so it falls under the category of passive data dictionary.

Data modeling tools: Similar to the ETL technique, data modeling also requires data manipulation, however, the way it comes together is different. The data modeler uses various tools to process the data and get the result.

There are certain advantages as well as disadvantages of using passive data dictionary as listed below:

1. Central metadata repository: Unlike the active data dictionaries, the passive ones are not integrated into a database or server. Hence, a single dictionary is sufficient for managing metadata of different databases located on different servers or engines.
2. Unified metadata from various databases: Various DBMS like Oracle, SQL Server, Redshift and others use different catalog tables. Thus, the active data dictionaries for these

databases are quite program-specific and deviate from each other. In case of a passive data dictionary, this problem can be surpassed by creating an ETL metadata repository.

3. Additional metadata: The autogenerated metadata present in the active data dictionaries comprises of trivial information. However, in a passive data dictionary, the user is free to add more relevant information regarding the database.
4. Flexibility: In order to update an active data dictionary, one needs to impact the schema of the database, which is often not possible. Passive data dictionaries overcome this drawback because the metadata can be easily updated without impacting the database.
5. User-friendly access: Since passive data dictionaries are maintained separately outside the database, it does not require the user to access the database for acquiring the metadata, hence they often provide user-friendly interface with search functionalities.

However, there are several downsides to implementing a passive data dictionary, which are stated below:

1. Manual maintenance: The most pronounced limitation of passive data dictionary is that it requires manual, or independent from DBMS, perpetuation of the metadata to apply all the changes in the database structure.
2. Metadata can be out of sync from database: An immediate consequence of the previous drawback is that the metadata in a passive data dictionary can be at times out of sync from the database.
3. Additional costs: Passive data dictionaries require additional work, which is mostly done by a third-party software; this incurs additional costs.

Now, an overarching question from practitioners' point of view would be: "*Which data dictionary should we use?*" The answer to the question is not straightforward. As we infer from the above-stated discussion, both active and passive data dictionaries have some strengths and weaknesses and should not be treated as contender to each other. In reality, passive dictionaries acquire metadata from their active counterparts, hence, the former can be regarded as an extension of the later.

Components of a Data Dictionary

Data dictionaries are almost always stored in a tabular format similar to that of a database itself. The attributes or fields of the data are listed as rows in the dictionary, and the columns comprise an element of useful information about the attributes. A detailed description of these components that can be included in a data dictionary is listed below:

- *Attribute name*: A string that identifies the attribute uniquely.
- *Optionality*: It specifies whether information is required in an attribute before a record can be saved.
- *Data type*: It describes what type of data is allowable in a file. Usually, the frequently used data types include numeric, time, string, Booleans, look-ups, and unique identifiers.
- *Data source*: It provides information regarding the origin of the data.
- *Maximum length*: It specifies the cardinality of the attribute.
- *Default values*: In any database, a user might change the data type of any attribute. A usual practice is to keep a track of the original data type to maintain data integrity.
- *Last updated on*: It indicates the date on which the data was last updated.
- *Ownership*: Multiuser-friendly database usually contains details about the person(s) who is(are) responsible for maintaining and updating the data.
- *Definition*: This element makes the database more user-friendly by providing the descriptive information about the attribute.
- *Calculations*: For derived attributes, this column includes the technical information on how the values have been calculated.
- *Acceptable values*: This column provides some values that are accepted by the attribute.

The above-stated list contains the core elements that are usually available in a data dictionary. However, it is not an exhaustive list, and sometimes a data dictionary might include more additional information based on its domain.

Significance of Data Dictionary

In today's data-driven world, big data are generated by every digital process, gadgets, systems, and sensors. Data dictionary is indispensable for tackling a huge volume of data. According to the International Data Corporation (IDC), the global spent on data initiatives in 2021 was estimated to be \$215.7 billion (Needham [n.d.](#)). In spite of significant investments in data lakes, there is no easy way for users to discover, access, and share data. As an immediate consequence of this, a vast amount of data is left unanalyzed. Usually, the data warehouse is primarily handled by a database administrator who tracks the data collection. However, in the absence of such a moderator, it becomes difficult for the data modelers to simultaneously maintain the database and analyze it. In such a scenario, a data dictionary is appropriate for the users as it provides detailed information regarding the data and assists in planning data collection, project development and other

collaborative efforts. Even though the data dictionary is primarily utilized as a metadata repository, there are several other contributory factors toward the usefulness of a data dictionary that are stated below.

1. **Anomaly detection**: Data dictionary summarizes the minimum or maximum values, count of distinct values, and spots duplicates or questionable data for all the attributes of a database. This summarization is useful for effortlessly identifying data irregularities or missing data in a database.
2. **Evaluate data quality**: Data dictionaries make it easier to create a standard set of variable names and descriptions. This helps operators to understand overall system design and data flow, and to find where data interact with various processes or components.
3. **Accurate data**: Data dictionary furnishes information about the source, ownership, and collection procedure of the data; data become more reliable and trustworthy.
4. **Data transparency**: Data dictionary presents detailed information regarding every entity within a dataset, thus reducing dependencies of its users and helping them use the data in the same way.
5. **Application design**: Application developers can create forms and reports with proper data types and controls, and ensure that navigation is consistent with data relationships using a data dictionary.
6. **Data integration**: The notions of various data elements provided in a data dictionary furnish the contextual understanding needed for deciding how to map one data system to another, or whether to subset, merge, stack, or transform data for a specific use.

Data Dictionary in NLP

With the growth of textual big data, the use of Artificial Intelligence (AI) technologies such as natural language processing (NLP) and machine learning becomes even more imperative. The initial steps of any text mining analysis comprise understanding the data examples. To start with, a word cloud is created by spotting the most frequently occurring words. These words are further mapped with the lexicons of the data dictionary for further analysis. The preciseness of these data dictionary plays a vital role in improving the accuracy of any NLP model. Usually, dictionaries are at the heart of any text mining analysis. In the context of NLP, dictionaries are termed as corpus and are majorly available on the web world. For instance, for a sentiment analysis problem where the aim is to automatically classify text according to the emotion they express, a sentiment analysis dictionary can be useful (Medhat et al. [2014](#)). Such a dictionary comprises of information about the emotions or polarity expressed by words, phrases, or concepts. In practice, a

dictionary usually provides one or more scores for each word. These scores are then used to compute the overall sentiment of an input sentence based on individual words. Some of the popular sentiment analysis dictionaries include SentiWordNet (Baccianella et al. 2010), SentiWords (Gatti et al. 2015), VADER (Hutto and Gilbert 2014), etc.

Data Dictionary in Geoscience

In the field of geoscience, several data dictionaries have been created to study land-cover recognition, remote sensing, and climate change and weather prediction problems (Long and Xinqing 2014). From daily temperatures to the movement of the Earth, drone images to satellite images, Google Maps to solar energy data, all relevant information are captured constituting a large volume of data by private and public organizations (Barclay et al. 2020). Some key examples of data dictionaries from geoscience and Earth science fields widely used in GIS are listed below:

- [Earth Explorer USGS Landsat Data Dictionary](#)
- [U.S. Geological Survey Open-File Report 03-001: Data Dictionary](#)
- [Aerial Photo Single Frames Data Dictionary](#)
- [National Hydrography Dataset Data Dictionary](#)
- [Planetary Science Dictionary \(NASA\)](#).
- [MODIS Level 1B Products Data Dictionary \(NASA\)](#).
- [Data Dictionary for Organic Carbon Sorption and Decomposition in Selected Global Soils \(ORNL\)](#).
- [Human Health Risk Assessment Data Dictionary \(ORNL\)](#).
- [Climate and Forecast Conventions Standard Name Table](#)
- [Data Dictionary for the National Database of Deep-Sea Corals \(NOAA\)](#).
- [National Elevation Dataset \(NED\) Data Dictionary](#)

Summary

In this chapter, we have explained how data dictionary is an inevitable part of a DBMS and modern data science. A data dictionary is actually a miniature database stored using special-purpose data structure and code. It is generally preferable to store the data about the database in the database itself that facilitates fast access to the system data. One can refer to a data dictionary in order to understand where a data item would fit in the structure of DBMS. The data dictionary gives vital information about the database in a well-structured manner. Using a data dictionary, any data redundancy (e.g., data duplication) can be easily identified. For multiple members of a research team working with similar data, having a data dictionary facilitates standardization by documenting common data structures and providing the precise vocabulary needed for discussing specific data elements. Shared

dictionaries ensure that the meaning, relevance, and quality of data elements are same for all the users. Data dictionaries also provide information needed by those who build systems and applications that support the data. Among various other applications, data dictionary in geoscience and natural language processing helps modern data scientists to understand the data flow and make robust predictions about the future. The current arena of research in building data dictionary still focuses on completeness that can explain the structure and content of data lakes. Future research on spatial data management system would be more focused on the creation of good quality data.

Cross-References

- [Big Data](#)
- [Database Management System](#)
- [Geographical Information Science](#)
- [Machine Learning](#)

References

- Baccianella, S., Esuli, A. & Sebastiani, F., 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. *In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.
- Barclay J, Starn J, Briggs M, Helton A (2020) Improved prediction of management-relevant groundwater discharge characteristics throughout river networks. *Water Resour Res*
- Batini, C., Di Battista, G. & Santucci, G., 1990. A methodology for the design of data dictionaries. *Ninth Annual International Phoenix Conference on Computers and Communications (pp. 706–707)*. IEEE Computer Society
- Date CJ (2000) *An introduction to database systems*, 7th edn. Pearson Education Inc
- Elmasri R, Navathe S (2000) *Fundamentals of database systems*. Addison-Wesley
- Gatti L, Guerini M, Turchi M (2015) SentiWords: deriving a high precision and high coverage lexicon for sentiment analysis. *IEEE Trans Affect Comput* 7(4):409–421
- Hutto C, Gilbert E (2014) Vader: a parsimonious rule-based model for sentiment analysis of social media text. *Proc Int AAAI Conf Web Soc Media* 8(1):216–225
- Long Z, Xinqing W (2014) General geo-spatial database construction method based on data dictionary. *Remote Sens Land Resour* 26(1): 173–178
- McDaniel G (1994) *IBM dictionary of computing*. McGraw-Hill, Inc., New York
- Medhat W, Hassan A, Korashy H (2014) Sentiment analysis algorithms and applications: a survey. *Ain Shams Eng J*:1093–1113
- Needham, M., n.d. International data corporation. [online] Available at: <https://www.idc.com/getdoc.jsp?containerId=prUS48165721>
- Ramakrishna R, Gehrke J (2000) *Database management systems*, 2nd edn. McGraw-Hill
- Raschka S, Mirjalili V (2019) *Python machine learning: machine learning and deep learning with python, scikit-learn, and TensorFlow 2*, 3rd edn. Packt Publishing Ltd.
- Silberschatz A, Korth HF, Sudarshan S (2006) *Database system concepts*. McGraw-Hill, Singapore

Data Life Cycle

Fang Huang
CSIRO Mineral Resources, Kensington, WA, Australia

Definition

Data life cycle, or data management life cycle, refers to the whole procedure of data management, starting from the conceptualization of a research question to creating a data product (Fig. 1). In different scenarios, data life cycle models can vary. For example, data sharing is crucial to normal research projects but not so to some sensitive and confidential ones. Based on the comprehensive DCC curation life cycle model (Higgins 2008) and the DDI conceptual model (DDI Structural Reform Group 2004), my proposed generic data life cycle model for mathematical geosciences data includes conceptualization, data collection, data processing, data analysis, data archiving, data sharing, and data product (Fig. 1). These steps are not linear – some steps can be repeated iteratively, and some steps can be further divided into smaller steps. All discussions in this paper mainly focus on research-related scenarios.

Introduction

In the big data era, with the development of scientific instruments and collection methods, the volume, variety, and velocity of data production are increasing at a speed never seen before. It is increasingly important for scientists to come up with tools and data management protocols for conducting research and building scientific collaborations. For example, in high energy physics, the Large Hadron Collider (LHC) accelerator at CERN in Geneva takes 40 million pictures (40 Tb of raw data) per second, which creates an extremely challenging data management problem (Editorial 2018). To deal with data of such a big size, in addition to fast data processing algorithms, a good data management plan is also crucial. A good data management plan will help (1) preserve data provenance and ensure the reproducibility and replicability of scientific projects; (2) make it easier for researchers to access, use and reuse of previously generated data; and (3) enable computers and machines to process data more accurately and efficiently, which becomes increasingly important in the big data era. An overview by Boulton et al. (2012) gives an overview of different data life cycle models.

Conceptualization

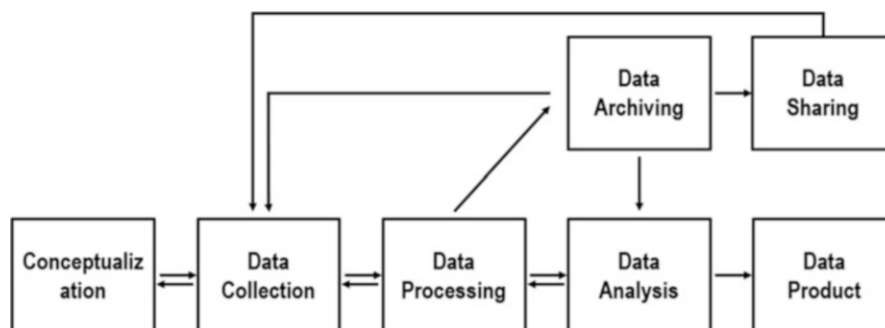
Research projects at different scales could emphasize on different steps of data life cycle models. Therefore, before starting data collection, we need to define the scientific goals, the desired results, and expected impacts of the research project. By doing this, we can better design the tools for data collection and processing and find efficient data types for storage. Obviously, these tools are very different among a research group of a few people and an international collaboration program of thousands of people. However, in some cases, we are dealing with problems of so-called “unknown unknowns” – we do not know the problem before some exploratory analysis. Hence, as the “first” step, data conceptualization can always be revisited and revised during the whole data life cycle.

Data Collection

There are generally two ways, usually combined, to collect data – produce data directly and gather data produced by others. Many geoscientists perform instrumental analysis on samples collected by themselves. This is more straightforward because researchers usually know what to expect from sample analysis. Another type is to gather data produced by others. Many published papers, especially old ones, do not have data readily for reuse. Usually, research data will be put as tables in the main text or provided in the supplementary documents. We must use computer programs or scripts to extract and process data from those papers and perform the process later. Another way to get previously published data is to retrieve them directly from data repositories. This way is more efficient with better data quality. A data repository is a set of databases that stores and curates data and provides easy access to users. More discussion can be found in “Data Archiving” and “Data Sharing” sections.

There are existing online tools to help researchers find datasets and data repositories of interest. For example, Google dataset search (datasetsearch.research.google.com) is a tool to discover datasets of interest, both scientific and nonscientific. Google processes and manages metadata from open data repositories and organize them in a way more accessible and findable by users. Another example is Re3data (www.re3data.org), which is a searchable registry of more than 2000 scientific data repositories. The repository registry entries are curated by a group of professionals and update weekly. Datacite (datacite.org) is the global leading provider of persistent identifiers (DOIs) for research data and outputs, where researchers can publish, find, and cite research data. Google dataset search and Datacite lead you directly to the datasets of interest while Re3data provides links to data repositories related to your search keywords. The collected data will be fed into the next step for processing.

Data Life Cycle, Fig. 1 The data life cycle in separate steps and relations



Data Processing

Data processing can have different definitions (Huang 2019), but in this chapter, data processing is defined as the process to clean, combine, and change raw data into the desired form for future analysis. It is said that data processing and preparation usually take 80% of the time of data analysis. After data are collected, we need to examine the dataset and clean up the errors and fix missing values in the data. The examination is sometimes achieved by performing exploratory data analysis, especially for large complex datasets. Depending on the results, we may need to go back to the data collection step to collect more data or even to data conceptualization to redesign the research plan. Data processing is very important and time-consuming. It should not be considered as a single step, but a continuous process that can be revisited during almost all steps throughout the whole data life cycle.

Data Archiving

Data archiving refers to the procedure of moving processed and organized datasets into a data repository, maintaining data provenance and making the datasets accessible to researchers. Data archive takes care of datasets, just like libraries take care of books. There are many existing well-curated data repositories for geoscience research. For example, EarthChem library (www.earthchem.org) hosts a large amount of geoscience analytical data, data syntheses, and even data models. Each dataset will be assigned a DOI for access and citation uses. The GeoBiodiversity database (www.geobiodiversity.com), or GBDB, provides data on stratigraphy and paleontology and enables researchers to conduct research on quantitative stratigraphy, biodiversity dynamics, and paleoecology. The Mindat (www.mindat.org) and Ruff (ruff.info) are both mineralogy repositories, which hosts data including basic mineral properties, origin of samples, formation ages, and even pictures of minerals.

Other than simply being a data curator and provider, data archiving also includes data curation and stewardship, which are key components ensuring the accuracy of the data and enabling the reproducibility and replicability of research projects. More materials can be found in previous studies

(Treloar and Klump 2019). One guideline called “the FAIR data principles” is widely recognized by many research organizations all over the world, including American Geophysical Union (AGU), National Institution of Health (NIH) of the US, European Commission, and countless universities and research institutions. The FAIR principles include the requirements to make data findable, accessible, interoperable, and reusable. Research data that follow these principles will be more easily used and reused by humans, and more machine friendly in terms of automatically finding and processing those data. A comment paper (Wilkinson et al. 2016) gives you a better understanding of the FAIR principles.

Data Analysis

Data Analysis in this chapter refers to the process of using statistical methods and algorithms to extract information from the data. After data processing, we can perform some preliminary analysis to examine the data to determine the next step to take. The data analysis results are then summarized in a data product like a research paper or a report. Typically, data analysis can be divided into exploratory analysis, descriptive analysis, and predictive analysis.

1. **Exploratory analysis:** It is an approach to analyzing datasets to summarize their characteristics, often with visual methods. The primary goal of exploratory analysis is to see if the data can tell us some information before the formal modeling. For example, check the completeness, examine for obvious errors and outliers.
2. **Descriptive analysis:** It usually summarizes the main features of the data. Correlation, classification, and regression analysis can be used to find the patterns and relationships among the data variables. Descriptive analysis, usually aided by data visualization, can help us get a better idea of the data and build hypotheses for next step analysis.
3. **Predictive analysis:** It uses a variety of statistical methods and machine learning algorithms to make predictions based on the trends and patterns found in existing data. In machine learning models, datasets are usually divided into training data and test data, which usually take 80 and

20%, respectively. After the selection of algorithm(s), we can use the train data to tune the parameters of the machine learning models. The test data should be held back from the model training. The performance of trained models can be evaluated using test data to avoid overfitting. There are many well-established ways to randomly select training and test data for unbiased results. With the trained model, we can make predictions on new data collected in the future.

Data Sharing

There are two key reasons for sharing data: (1) for reproducible and replicable scientific research and (2) for future large-scale scientific research by integrating existing research data. According to the “Reproducibility and Replicability in Science” report (National Academies of Sciences and Medicine 2019), reproducibility is “obtaining consistent results using the same input data; computational steps, methods, and codes; and conditions of analysis” and replicability is “obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data,” in which data are the key component.

Other than reproducibility and replicability, data sharing also plays an important role in enabling new sciences in the future (Editorial 2018). In astronomy and astrophysics, the Large Synoptic Survey Telescope (LSST) being built in Chile will capture images of the night sky, and each night will produce 20TB of data, which will be open to the public (Chahin 2017). Similar situation in which large scientific programs share data publicly also happens in research fields like seismology and meteorology. For economics and biology, there is also a well-developed culture of data sharing – researchers working on genetic sequencing cannot publish papers without submitting related data into specific repositories. Compared to these fields, the situation in geoscience communities is improving but not there yet. The shared data can be feed into data collection by other researchers.

Having said that, research projects with different goals can have distinct requirements on data openness: most published data are completely open while some sensitive projects must be kept confidential. Researchers need to find a balance between confidentiality and public interest or benefits.

Data Product

In the current chapter, data product refers to the product produced by the insights gained from previous data analysis, which can be a scientific paper, a report, or simply a dataset, or a tool, an application, or a model that is useful to the community. As a main data product in research, data publication has been proposed and discussed to be a way to give credits to data creators/collectors (Klump et al. 2006). Due to historical reasons, though they are very time-consuming and labor

intensive, data collecting, processing, and curating are often not as well acknowledged and appreciated as scientific papers. In traditional tenure track evaluation systems, journal papers, scientific grants, community service contributions, student guidance, and teaching performance are all important metrics but not for datasets creation and maintenance. Therefore, data publications become an important data product to fill in this gap. However, unlike static scientific publications, datasets can be evolving constantly as new data are being generated and added, which will cause dataset version issues. The discussion is still on going to find better ways to give dataset creator credits, and this discussion should involve all stakeholders – researchers, funding agencies, universities and institutions, and research organizations.

Summary

In the big data era, data become an increasingly important component in scientific research. A data management model that follows the FAIR principles can increase the efficiency of research collaboration and communication by making data more findable and accessible. By making research more transparent, it can also improve reproducibility and replicability of published research. With the help of machines to automate the collecting, processing, analyzing, and sharing data, novel research with larger scale of data will be made possible. Though many examples in this chapter focus on geoscience data, the discussions should be applicable to other research fields as well.

Cross-References

- [Big Data](#)
- [Data Mining](#)
- [Database Management System](#)
- [FAIR Data Principles](#)
- [Geoinformatics](#)
- [Open Data](#)

Bibliography

- Boulton G, Campbell P, Collins B, Elias P, Hall W, Laurie G (2012) Science as an open enterprise. The Royal Society Science Policy Centre report 02/12. The Royal Society, London
- Chahin M (2017) How big data advances physics. Elsevier Connect. <https://www.elsevier.com/connect/how-big-data-advances-physics>
- DDI Structural Reform Group (2004) DDI Version 3.0 conceptual model. DDI Alliance. <http://www.ddialliance.org/sites/default/files/Concept-Model-WD.pdf>
- Editorial (2018) Data sharing and the future of science. Nat Commun 9(1):2817. <https://doi.org/10.1038/s41467-018-05227-z>

- Higgins S (2008) The DCC curation lifecycle model. *Int J Digit Curation* 3(1):134–140. <https://doi.org/10.2218/ijdc.v3i1.48>
- Huang F (2019) Data processing. In: Schintler LA, McNeely CL (eds) *Encyclopedia of big data*. Springer, pp 1–4. https://doi.org/10.1007/978-3-319-32001-4_314-1
- Klump J, Bertelmann R, Brase J, Diepenbroek M, Grobe H, Höck H, Lautenschlager M, Schindler U, Sens I, Wächter J (2006) Data publication in the open access initiative. *Data Sci J* 5:79–83
- National Academies of Sciences and Medicine, E (2019) *Reproducibility and replicability in science*. National Academies Press, Washington, DC
- Treloar A, Klump J (2019) Updating the Data Curation Continuum: not just data, still focussed on curation, more domain-oriented. *Int J Digit Curation* 14(1):87–101
- Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, ... Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3(1):160018. <https://doi.org/10.1038/sdata.2016.18>

Data Mining

Tao Wen

Department of Earth and Environmental Sciences, Syracuse University, Syracuse, NY, USA

Synonyms

Data dredging; Data science; Knowledge discovery from data; Knowledge extraction; Machine learning; Pattern analysis

Definition

Data mining is the process of utilizing computer algorithms to discover knowledge and information from large amounts of data. Data mining, often interchangeably used with knowledge discovery from data (KDD), is sometimes viewed merely by others as one critical step of KDD (Fig. 1) (Han et al. 2011). The three essential elements of data mining are data, algorithms, and information (or knowledge). Data mining is interdisciplinary and employs many techniques from statistics, machine learning (ML), database, high-performance computing, and domain disciplines (e.g., geosciences). Here, the domain discipline refers to the field in which data mining is applied to gain knowledge and information. In geosciences, it is also not uncommon that data mining is treated as a synonym for machine learning, although some might argue that machine learning is more focused on the development of computer algorithms that are applied during the workflow of data mining. The major types

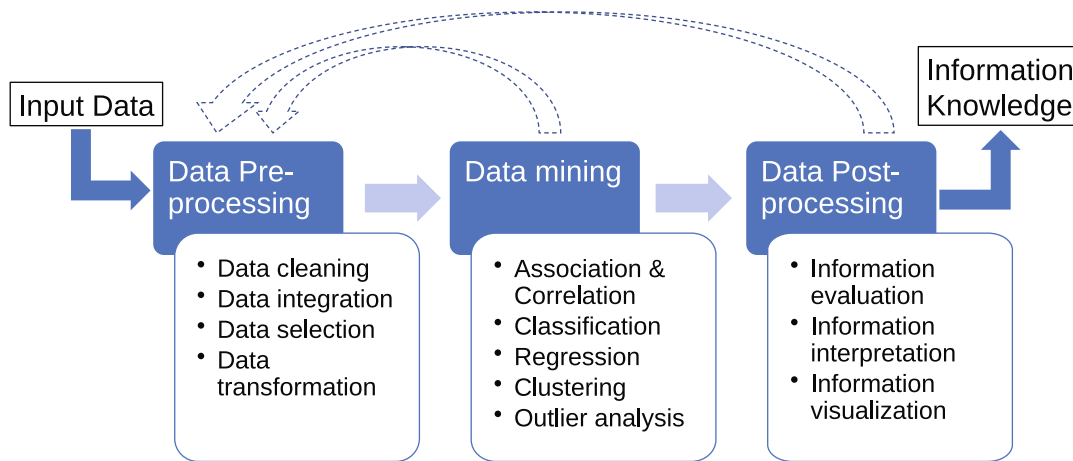
of geoscience data that can be mined for information include matrix data, time-series data, network-based data, and geospatial data. Based on the type of information to be mined and the technologies to be applied, data mining tasks can be divided into five categories: association and correlation, classification, regression, clustering, and outlier analysis (Han et al. 2011). Benefiting from the ever-increasing stream of large geoscience data and the improvement of computational resources, the widespread application of data mining in geosciences becomes inevitable. The most promising near-future development of data mining in geosciences is likely the integration of data mining and physics-based modeling (Reichstein et al. 2019).

What Geoscience Data Can Be Mined?

The explosive growth and ever-increased diversity of geoscience data in the past few decades are at least partially attributed to the development of cheap automatic sensor-based monitoring devices that are deployed in the soil, water, and air. Among those large geoscience datasets, major types of data that can be mined include matrix/tabulated, time-series, geospatial, and network/graph-based. There are overlaps among these data types, e.g., all of the latter three types of data can be formatted as matrix/tabulated data.

The matrix/tabulated data is probably the most common data format in geosciences. Matrix data often has two dimensions in which each row represents one sample while each column contains a feature or characteristic of the sample. For example, Tesoriero et al. (2017) compiled a large set of groundwater quality samples including nitrate, iron, arsenic concentrations, and other bedrock and soil characteristics. In this example, each row represents one groundwater sample. Matrix data is intuitive to read. However, it becomes less convenient to use matrix data to illustrate the spatiotemporal trends of data or the intersample relationship. This is why we also need to organize geoscience data in other formats, e.g., time-series, geospatial, and network-based data. These additional data types allow the incorporation of additional spatiotemporal context and the information of the intersample relationship into the stored data.

Time-series data refers to a set of data points indexed in time order. The time when the data point is collected is often used as the timestamp, although other types of timestamps have also been used (e.g., age of rock samples). Time-series data can be formatted as matrix data in which each row represents a sample from a time point and an additional column(s) should be added to show the timestamp of each sample. Time-series data is often used to describe the temporal trend of geoscience parameters. For example, the National Water Information System from the United States Geological Survey (<https://waterdata.usgs.gov/nwis>) hosts time-series



Data Mining, Fig. 1 The view of knowledge discovery from data (KDD) and data mining from the perspective of machine learning and statistics communities

data of water quantity and quality of the water body in the United States.

Geospatial data refers to the geoscience data for which we have information on the geographic location. Similar to time-series data, geospatial data can be formatted as matrix data as well when the location information is added as additional column(s) (e.g., latitude and longitude) in the matrix. In particular, an important source of geospatial data is the remote sensing data from satellites. For example, Google Maps and government agencies provide public access to low- and high-resolution satellite and multispectral imagery.

The last common data type in geoscience is network-based data. Unlike matrix data, network-based data is represented by nodes (vertices) and edges (links; connecting nodes), which can effectively illustrate the relationship between samples. For example, Agarwal et al. (2020) reported a dataset of five geochemical analytes including chloride, bromide, barium, magnesium, and sodium of >80,000 river water samples in a complex river network. Agarwal et al. used the river network-based data to inform the impact of human activities on river water chemistry. Barbier et al. (2020) visualized the compiled data from multiple experimental serpentinization studies into a network with each node representing each experiment and edges indicating the similarity between different experiments.

More and more geoscience data, in particular, the compiled datasets in large size, becomes available nowadays because more geoscience researchers are following the findability, accessibility, interoperability, and reusability (FAIR) principles when sharing data. To comply with FAIR principles, geoscience data should be transformed into a machine-readable format (Wen 2020). Geoscience data can be shared in either discipline-specific or general data repositories. A data repository is a place storing data and providing

access to users (Wen 2020). For example, the HydroClient database (<https://data.cuahsi.org/>) is a popular database in which the hydrology community publishes time-series and geospatial data. Data in the HydroClient database can be downloaded as matrix data. In addition, the Shale Network database (<https://doi.org/10.4211/his-data-shalenetwork>) that is published in the HydroClient compiles water quality data for water samples collected from the regions of oil and gas production. Each of these Shale Network samples has geographic coordinates reported. The PANGAEA database (<https://www.pangaea.de/>) is an example of more general data repositories hosting geospatial data from many subdisciplines of geosciences.

What Methods Are Used in Data Mining?

Data mining tasks include the characterization of the data properties (descriptive task) and predictions for the past or the future by learning from existing data (predictive task). Major data mining tasks introduced below include association and correlation, clustering, outlier detection, and classification and regression. The first three tasks are descriptive while the last one is a predictive task.

The goal of association and correlation (AC) analysis, as a descriptive task, is to assess the association or correlation between parameters. In geosciences, a common correlation task is to evaluate the spatial correlation of two parameters. For example, Wen et al. (2018) reported a compiled ground-water quality dataset from the region of oil and gas production in the state of Pennsylvania in the United States. With the rapid development of oil and gas drilling in the United States, the public is getting concerned about the associated environmental impact, e.g., the groundwater contamination. Wen

et al. applied a geospatial analysis tool (<https://github.com/jaywt/SWGT>) to assess the spatial correlation between methane (the major component of natural gas) in groundwater and the proximity to oil and gas wells. They successfully identified a few localities where some natural gas wells are potentially problematic. In paleoclimatology, scientists have striven to find geochemical proxies that correlate with the level of atmospheric oxygen, and air and sea temperatures to reconstruct the ancient climate.

Cluster analysis is the data mining task of classifying a set of samples in a way that a sample is more similar to samples in the same class than to those in other classes. Cluster analysis is one type of unsupervised learning technique. Unsupervised learning refers to those data mining techniques which discover information in a dataset with no preexisting classes. Unlike unsupervised learning, supervised learning techniques rely on datasets with preexisting classes which are often generated by humans. Cluster analysis is used to generate these previously unknown classes for a dataset. The principle of cluster analysis is to maximize intra-class similarity while minimizing inter-class similarity (Han et al. 2011). Typical clustering techniques include K-Means, spectral clustering, Gaussian mixture, self-organizing map (SOM), and affinity propagation. For example, SOM was previously used to infer distinct but spatially contiguous classes within footwall and hanging, basalts, and andesites (Cracknell et al. 2014).

Outlier analysis (OA) is the task of detecting rare or anomalous sample(s) in a dataset. These rare or anomalous samples often fail to comply with the general rules or patterns of the majority of the dataset. Such samples are called outliers. Two major applications of outlier detection in geosciences include the removal of anomalous data before performing other data mining tasks and the detection of ecosystem disturbances. Zheng et al. (2017) developed an outlier detection algorithm and applied it to water quality data along with relevant contextual features from the regions of oil and gas production to successfully detect groundwater samples likely contaminated by nearby problematic oil and gas well(s). Some examples of these contextual features include the location of water samples, the information about nearby oil and gas facilities, and geologic features (e.g., the distance to faults).

Unlike the above tasks, regression analysis (RA) and classification analysis (CA) are supervised learning techniques and predictive tasks. In RA, data mining algorithms are first used to evaluate the relationship between a target variable and often a set of predictor variables in a dataset (i.e., training data) for which values of both target variable and predictor variables are known, and then applied on another dataset (testing data) for which only values of predictor variables are known to predict the values of target variable of the testing data. Here, the target variable refers to the parameter to be

predicted while predictor variables denote the set of features used in the RA/CA algorithms to make predictions for the target variable. Similar to RA, CA also requires both training data and testing data, while the goal of CA is to predict the classes of the target variable instead of its values. Typical algorithms used in RA and CA include decision trees, random forest, support vector machines (SVM), naïve Bayesian, rule-based classification, logistic/linear regression, and neural networks. For example, Wen et al. (2021) applied the logistic regression method on compiled water quality data of groundwater samples to predict the probability of dissolved methane in groundwater being originated from nearby oil and gas production activities.

Summary and Conclusions

Data mining refers to the process of discovering information from a large amount of data. Driven by the increasing volume, variety, and velocity (3Vs) of geoscience data as well as the advancement in data science techniques, data mining methods have been widely applied in geosciences. The most promising near-future development of data mining in geosciences is likely the integration of data mining and physics modeling (Reichstein et al. 2019).

Cross-References

- [Machine Learning](#)
- [Spatial Data](#)

Bibliography

- Agarwal A, Wen T, Chen A, Zhang AY, Niu X, Zhan X, Xue L, Brantley SL (2020) Assessing contamination of stream networks near shale gas development using a new geospatial tool. *Environ Sci Technol*. <https://doi.org/10.1021/acs.est.9b06761>
- Barbier S, Huang F, Andreani M, Tao R, Hao J, Eleish A, Prabhu A, Minhas O, Fontaine K, Fox P, Daniel I (2020) A review of H₂, CH₄, and hydrocarbon formation in experimental serpentinization using network analysis. *Front Earth Sci* 8:209. <https://doi.org/10.3389/feart.2020.00209>
- Cracknell MJ, Reading AM, McNeill AW (2014) Mapping geology and volcanic-hosted massive sulfide alteration in the Hellyer-Mt Charter region, Tasmania, using Random Forests™ and Self-organising Maps. *Aust J Earth Sci* 61:287–304. <https://doi.org/10.1080/08120099.2014.858081>
- Han J, Pei J, Kamber M (2011) Data mining: concepts and techniques, the Morgan Kaufmann series in data management systems. Elsevier Science, Saint Louis
- Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, Prabhat (2019) Deep learning and process understanding for data-driven earth system science. *Nature* 566:195–204. <https://doi.org/10.1038/s41586-019-0912-1>

- Tesoriero AJ, Gronberg JA, Juckem PF, Miller MP, Austin BP (2017) Predicting redox-sensitive contaminant concentrations in groundwater using random forest classification. *Water Resour Res* 53:7316–7331. <https://doi.org/10.1002/2016WR020197>
- Wen T (2020) Data sharing. In: *Encyclopedia of big data*. Springer International Publishing, Cham, pp 1–3. https://doi.org/10.1007/978-3-319-32001-4_322-1
- Wen T, Niu X, Gonzales M, Zheng G, Li Z, Brantley SL (2018) Big groundwater data sets reveal possible rare contamination amid otherwise improved water quality for some analytes in a region of marcellus shale development. *Environ Sci Technol* 52:7149–7159. <https://doi.org/10.1021/acs.est.8b01123>
- Wen T, Liu M, Woda J, Zheng G, Brantley SL (2021) Detecting anomalous methane in groundwater within hydrocarbon production areas across the United States. *Water Research* 200:117236. <https://doi.org/10.1016/j.watres.2021.117236>
- Zheng G, Brantley SL, Lauvaux T, Li Z (2017) Contextual spatial outlier detection with metric learning. In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining – KDD '17*. ACM Press, New York, pp 2161–2170. <https://doi.org/10.1145/3097983.3098143>

Data Visualization

Jaya Sreevalsan-Nair

Graphics-Visualization-Computing Lab, International
Institute of Information Technology Bangalore, Electronic
City, Bangalore, Karnataka, India

Definition

As data has become abundant, data visualization has evolved as one of the most universally used data exploration and analytic methodology. There has been a recent uptake in availability of the open-source libraries and browser-based applications, leading to democratization of the use of visualization among data enthusiasts. Data visualization is formally defined as the visual representation of data based on specific mapping between visual constructs and aspects of data (Munzner 2014). Hence, visualization in itself is a multi-disciplinary pursuit involving design, data sciences, computer graphics, human-computer interaction, and required domain knowledge. From a geospatial or a geoscientific perspective, the commonly used terms for data visualization are geovisualization (MacEachren and Kraak 2001) and geovisual analytics (Andrienko et al. 2007). The goal of geovisualization has been to enable viewers or users to “explore, analyze, synthesize, and present” the data that defines the domain.

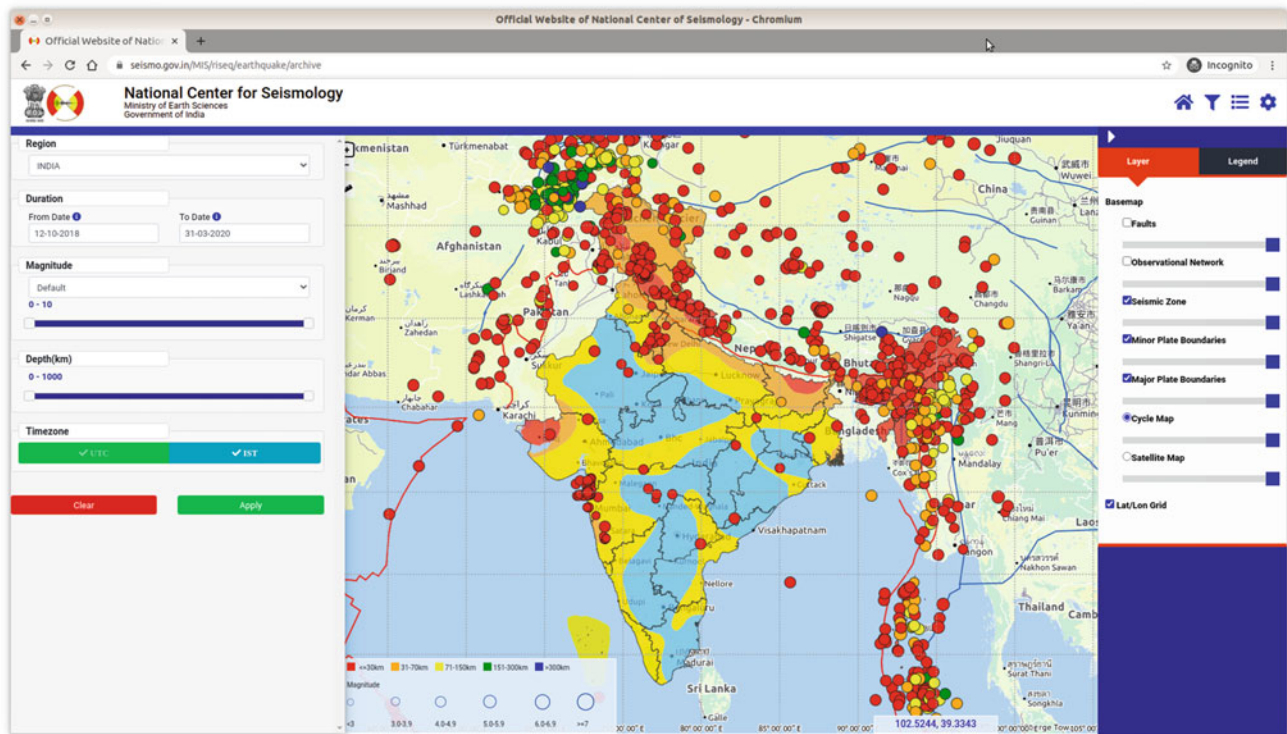
Overview

The usage of data visualization caters to a spectrum of requirements, spanning from data summarization to data

exploration. Each methodology has a combination of summarization and exploration goals in varying degrees. With the advent of high-throughput graphics processing units and interactive display technologies, most visualization systems or tools facilitate real-time rendering and effective human-in-the-loop workflows. Thus, what used to be infographics and static visualization have made way to dynamic visuals. The degree of interactivity in today’s graphical user interfaces varies from text pop-ups for mouse over to data transformations by slicing and dicing. Thus, data visualization and human-computer interaction have now evolved to be inseparable with respect to system requirements and subsequent implementation.

The graphical content of conventional visualizations in geosciences, just like another scientific domains, is determined by the data types. In geosciences, cartography has played a major role in visualization since the human quest for global exploration in maps. The use of cartography emphasizes the importance of visual embeddings, e.g., color schemes (Brewer 1994). The need for visual embeddings, mappings, or correspondences is as old as mankind, owing to its perceptual characterization. The cave paintings are lasting evidence of daily habits of the early Homo sapiens, which can be classified as spatiotemporal visualization, in today’s parlance. Spatiotemporal data is undeniably a key element in geoscientific understanding. Here, we first list out the types of visualizations one would encounter when looking for geosciences-related visualizations. We classify some of the commonly used visualizations into three broad types: (i) map-based visualizations using geo-referenced data, (ii) scientific visualizations using data with physical dimensions, and (iii) information visualizations using other kinds of data which exclude positional information, e.g., statistical measures or frequency data, categorical data, etc. These three types of visualization techniques are usually used in combination, e.g., in the form of overlays. A simple example would be a surface representing bathymetric data, with a map as a reference. The methods we discuss in this article also tend to have variants specific to the geoscientific subdomain where they are applied. For instance, marine geosciences require the use of ocean models and several multiscaled entities for data analysis including visualization (Xie et al. 2019). In that regard, Hovmöller plots or diagrams are unique to ocean data visualization.

Commonly used map-based visualizations include choropleth map, hotspot visualizations, network visualization, symbol map or glyph/icon-based visualizations, abstract visualizations such as cartograms, and three-dimensional projected visualizations. Choropleth map is a thematic map where regions in the cartographic map are colored based on a collective or aggregate value, e.g., population density in the USA, or a finer-grained one at the county level. In Fig. 1, we observe the seismic regions in India colored using a



Data Visualization, Fig. 1 Preliminary data of recent earthquakes in India in 2019, available in the form of a map-referenced point data visualization, rendered as glyphs. This website is an example of open

data dissemination by a governmental agency, e.g., National Center for Seismology, Government of India. Image courtesy: https://seismo.gov.in/MIS/riseq/earthquake/recent_earthquake

qualitative color scheme of blue-yellow-orange-red corresponding to Zones II-III-IV-V, respectively, indicating the increase in damage risk in the zones. Hotspot visualizations are similar to choropleth map in appearance, but the input to the visualization is point data as opposed areal data. From the point data, the kernel density estimation is done to identify regions of varying density from the hotspots, or highly dense clusters of points. Hotspot visualizations are used in the case of studying disease outbreak, traffic congestion, earthquake propagation, etc. Network visualizations include node-link diagrams to demonstrate connections and relationships. Ship trajectories, as shown in Fig. 2, is a good example. The node-link diagrams suffer from visual clutter caused by several crossing edges of the network. A way to minimize the clutter is to use curved edges and/or transparency in rendering the edges, as shown in Fig. 2. Edge bundling also helps in reducing visual clutter by emphasizing the density and directionality of edges.

Glyph-based visualization uses icons or glyphs, such as circles, to indicate points (Fig. 1), Chernoff faces for population value, etc., at specific points on the map. The glyph-based visualizations facilitate the use of glyph properties, such as its shape, size, and color, as visual marks for mapping point-wise variables. For instance, in Fig. 1, the size of the circular glyphs indicates the strength of the earthquake, the color,

and its depth in km. The abstract visualizations include cartograms, which are distorted shapefiles used in maps. These distortions are an outcome of constraining the shape and area of the shapefiles to meet specific criteria, e.g., area proportional to the value of a variable, or a frequency count for the variable. A cartogram is generated from the choropleth map. Contiguous, noncontiguous, and Dorling-type cartograms are generally used. Three-dimensional projections involve mapping onto surfaces of three-dimensional objects, such as sphere or cylinder. Thus, the three-dimensional map-based visualizations are essentially 2.5-dimensional ones of the globe and its variables. Google Earth is an example of such three-dimensional projections.

The scientific visualization techniques are used on two-, three-, and four-dimensional datasets, including varying combinations of three spatial and one temporal dimensions. Three common techniques for two-dimensional data are: (i) feature lines, e.g., streamlines and contours, (ii) colormaps, e.g., gradient colormaps and contour fills, and (iii) glyph visualization, such as arrow plot for air current visualization.

Contours or isolines have different names depending on what variable they represent, e.g., elevation contours for height, isobaths for ocean depth, isochrones for time needed to travel, isotherms for temperature, isobars for (atmospheric) pressure, and so on. An isoline for a specific variable and a



Data Visualization, Fig. 2 Data of worldwide ship routes in 2012, curated, archived, and represented as a map-referenced network visualization, rendered as node-link diagram, colored by ship type. This

website is an example of a publicly available exploration tool for archived data. Image courtesy: <https://www.shipmap.org/>

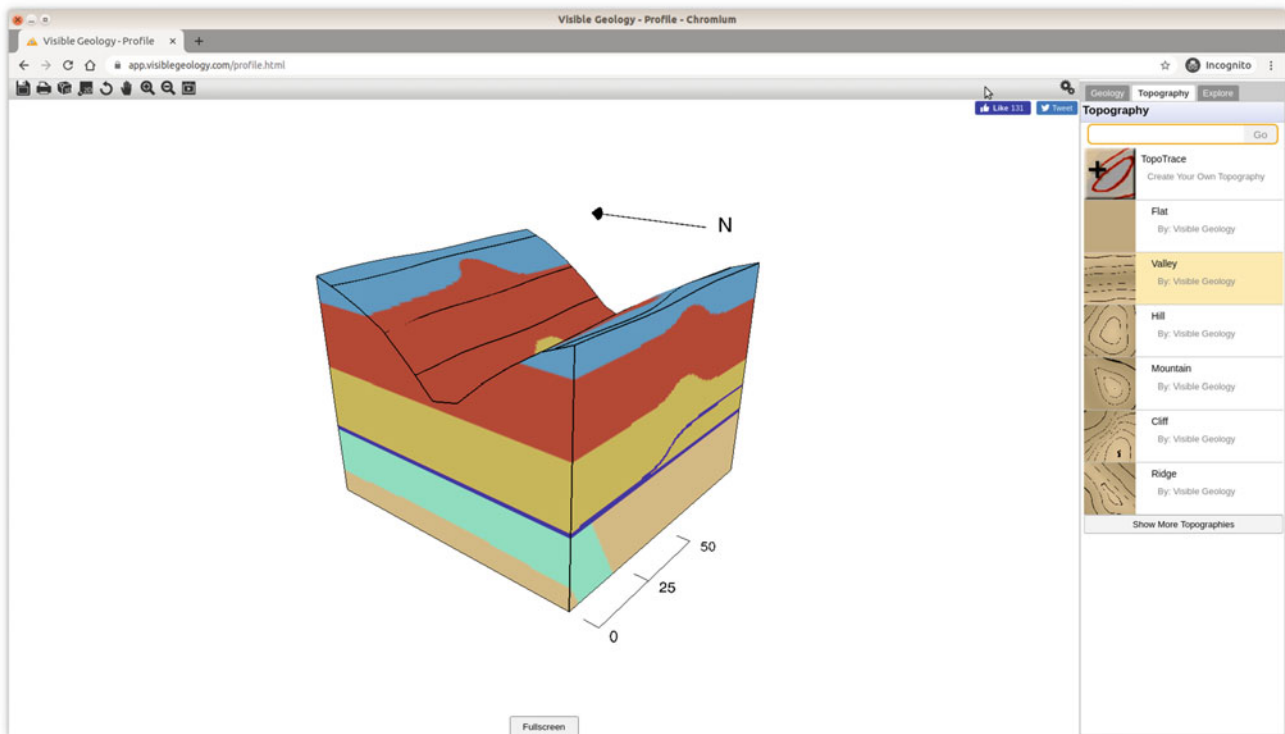
specific value is defined as the locus of points where the variable has the value, and the points are joined together piecewise by line segments. Isolines are used to partition space based on the variable value. As an example, an elevation line for a certain height, h , divides a two-dimensional surface data to a region with height greater than h , and otherwise. Apart from isolines, the feature lines also include vector field lines, such as streamlines, streaklines, pathlines, and timelines, used in vector fields, e.g., ocean currents. Streamlines are widely used vector field lines to understand “flow” orientation, as the vector direction at each point on the line is tangential to the field line. As an extension to three-dimensional volumetric datasets, these feature lines are to surfaces, such as isosurfaces, and visualized as colored or semitransparent surfaces. Figure 4 demonstrates surface constructed from bathymetric data and isosurfaces of salinity. Another technique used for volumetric visualization is direct volume rendering, which uses transparencies mapped to variable values to visually represent interfacial material boundaries. Direct volume rendering is a method which exploits the optical properties of the materials in the volume to enable a see-through visualization into the volume. This type of rendering works the best for data with layers of material, e.g., geological data, as in the simulation in Fig. 3.

Colormapping is the use of a color palette to map a color to a variable using interpolation functions, and subsequently

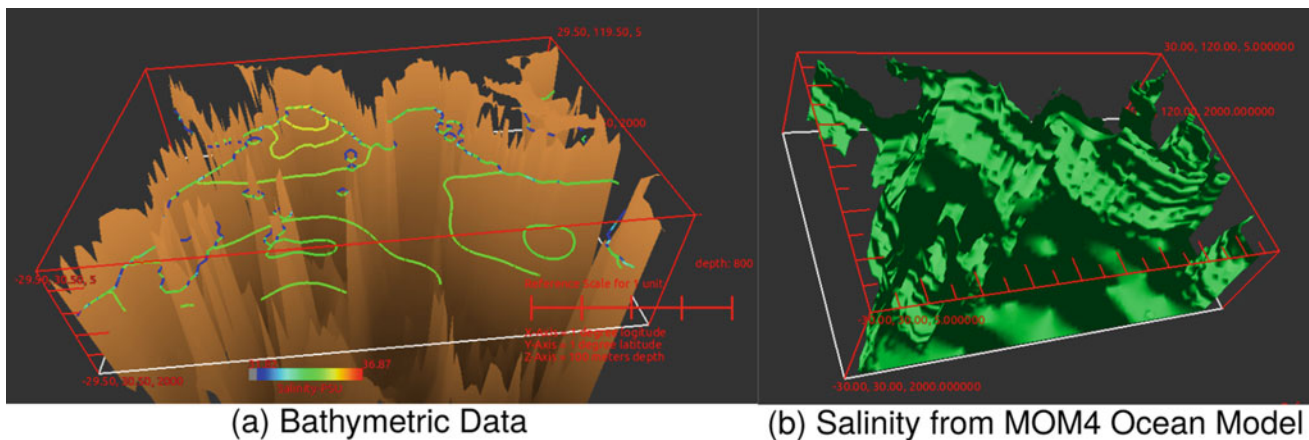
using the appropriate color fill for rendering surfaces. This is commonly used for continuous physical variables, such as temperature, pressure, particle concentration, etc. If intervals are used instead of interpolated functions, we get contour fills when using the color maps. Figure 3 shows contour fills of soil types in a simulated volumetric space. Interestingly, determining the appropriate gradient color palette for colormaps and qualitative color schemes for contour fills requires understanding of color perception and visual mapping (Brewer 1994).

The glyph-based visualization is the same as has been explained for map-based visualizations, and the similar concept can be used for continuous fields. A well-known example of glyphs is the use of arrows for vectors for encoding the magnitude, direction, and orientation of the vector at each point. Glyph-based visualization can also be extended to higher-dimensional space. Point cloud rendering of three-dimensional laser-scanned topographical data, as shown in Fig. 5, is considered to be a special case of glyph-based visualization.

The information visualization community categorizes map-based visualizations as abstract data visualizations. However, bearing in mind the presence of geo-referenced positional information in maps, we mention them separately here, in the context of geosciences. Without the abstraction of position, the design space of information visualizations is



Data Visualization, Fig. 3 Volumetric visualization of a simulation model of geology using a feature-rich browser-based graphical editor. This website is an example of a teaching tool for geology. Image courtesy: <https://app.visiblegeology.com>



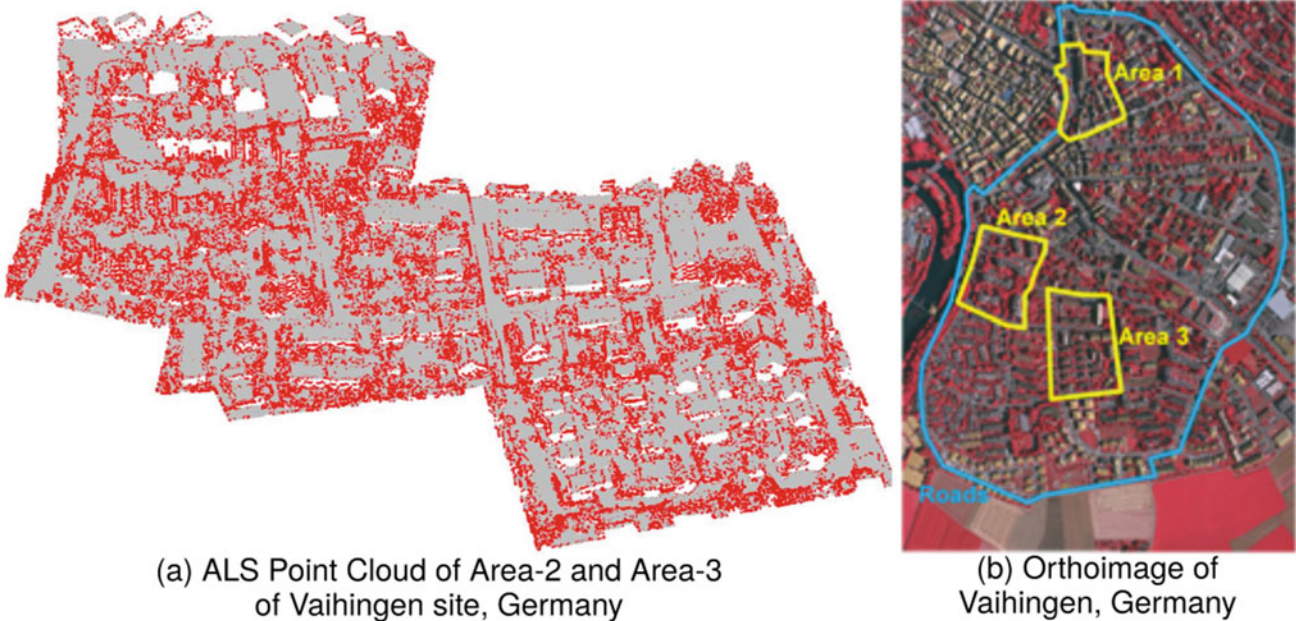
Data Visualization, Fig. 4 (a) This is a combination of surface rendering of bathymetric data (brown) and contour extraction of salinity data rendered using the color spectrum in the legend, for salinity data on a slice at depth 800 m below sea surface in the Indian Ocean dataset. (b)

Isosurface of salinity, rendered in green, at 34.7995 PSU on 15 January 2004, in the volumetric data from the MOM4 ocean model of the Indian Ocean. Image source: Author's own; data source: <https://las.incois.gov.in/las/UI.vm>

large. This design space includes all the commonly used visualizations, such as statistical plots, fractals, diagrams, etc. We direct the reader to the recent handbook of mathematical geosciences (Sagar et al. 2018) for finding several examples of such visualizations. Apart from these, there are commonly used multivariate visualization techniques such as parallel coordinates plots, scatter plot matrices, matrix visualizations, etc., and hierarchical data visualization such

as treemaps, dendrograms, sun burst, etc., which are useful for understanding relationships between data entities.

The presence of time in the data gives rise to different kinds of visualization depending on its role as a dimension versus a variable. When time is a dimension, animation is the best method which mimics the passage of time. A good example of map animation is the use of time-lapse to demonstrate the changes on the earth during 1984–2016, using the



Data Visualization, Fig. 5 (a) Point rendering of Aerial Laser Scanned (ALS) point clouds of topographic data from Vaihingen site, Germany, using data provided by ISPRS as benchmark data. The red points are linear features, delineating structures, e.g., building roofs, and

localizing objects, e.g., foliage, and gray are remaining points. (b) Corresponding orthoimage. Image source: Author's own; data source: <https://www2.isprs.org/commissions/comm2/wg4/benchmark/3d-semantic-labeling/>

Google Earth Engine. Scientific visualization of time series also conventionally uses animation as a more effective ways to understand temporal changes. In information visualization, line graphs have been commonly used for time-varying one-dimensional signals. When time duration is a variable, e.g., geologic age of rock layers, then it is visualized just as any other variable using colormap, glyphs, etc.

Geovisualization and Geovisual Analytics

The aforementioned visualizations are generally integrated to form visualization systems for complex multivariate geospatiotemporal data. Different kinds of data and model scenarios involved in scientific disciplines tend to be multifaceted (Kehrer and Hauser 2012). Heterogeneity in data requires multiple coordinated views for visualizing the multiple facets, and workflows which can flexibly fuse elements at different levels, e.g., data level, visualization level, etc., for making inferences. Consequently, automated geoscientific analytic systems have visualizations in the graphical user interfaces using a combination of different visualizations in composite views to provide holistic insights to the data. For instance, the system for automated geoscientific analyses (SAGA) open-source geographic information system (GIS) provides users several options for visualization of varied kinds of data for different applications, such as soil mapping, climatological studies, etc. (Conrad et al. 2015). Visualization systems have

gradually and steadily paved the way for visual analytics, a relatively new terminology used for visualization systems that implement the endless feedback loop in data science workflow with human in the loop. This feedback loop allows inferential analysis from data using hypotheses built as consequences of data science workflows which predominantly involves visualization. Geovisual analytics (Andrienko et al. 2007) pertains to data from geosciences. Geovisual analytics for spatial decision support has been the singular research agenda for the community. This agenda is different from other visual analytics-driven work based on three aspects, namely the complex spatiotemporal nature of the concerned data, multiple stakeholders, and tacit criteria and knowledge. A motivating example of geovisual analytics is in the use of visualizations for emergency response in a disaster-affected region, say landslides. Synchronized dashboards of interconnected data of weather, road network, hospital locations, rescue camps, etc. enable quick decision-making, planning, and implementation by administration. Such dashboards use design patterns of user interactivity, such as brushing, linking, slice and dice, etc., which are also useful for researchers to perform in-depth study of the situation.

The evolution of visualization systems is strongly tied to the change in technologies involved. With the widespread usage of information and communication technologies in the late 1990s and early 2000s, the habits of both expert and casual users in working with data have also changed. There are more browser-based visualizations catering to several

needs, today, than ever before. Government agencies use them for data dissemination for public awareness (Fig. 1), data providers allow users to explore archived data (Fig. 2), and education experts use them as learning aids (Fig. 3). Also, visualization has increasingly become a way of open data sharing (Figs. 4 and 5).

Challenges, Opportunities, and Future Scope

While there has been technical advancement, the challenges pertaining to visualization have also changed with the technologies involved. The geovisualization research community have identified three persistent challenges, i.e., challenges which are long standing, rather than being of high priority (Çöltekin et al. 2017). Firstly, data visualization requires understanding of the disciplines involved, such as scope of the domains, inter-domain dependencies, and interdisciplinary collaboration. While the first issue is agnostic of the application domain, the second issue of gaps in systematic understanding of human factors is unique to human-in-the-loop computing. The third issue lies in the lack of practical guidelines that matches the visualization types to task types, and design guides for effective and appropriate geovisualizations for the user, thus bringing back the focus to the geoscientific domain. The third issue is, thus, specific to the geosciences. These research challenges of appropriate visual designs of spatiotemporal data, effective data representation for human consumption, and efficient system/workflow integrations have dogged the community for two decades (MacEachren and Kraak 2001).

The aforementioned challenges pave way to opportunities, such as democratization of visualization, which means there are a slew of open-source resources for visualization tools, systems, and libraries, to cater to the needs of the geovisualization community. “Democratization” refers to the extension of accessibility of such tools to a larger audience, not limiting to the visualization experts. Democratization, by design, improves the usability of visualization systems by more stakeholders. The browser-based tools especially have an appeal for reaching a wider audience, thus popularizing the scientific discipline, without the weight of the need of mathematical or programming knowledge to build the visualizations. The pandemic of 2019–2020 caused by novel coronavirus saw the rise of the browser-based choropleth map visualizations for knowledge dissemination of the progression of the pandemic at different spatial scales, namely states, countries, continents, and the proportion of the population being affected worldwide. The rise of such geo-referenced visualizations is also tangibly attributed to the concerted efforts in data gathering, and availability of cloud computing facilities. These visualizations organically grew from the popular consumption of data from browser-based

map applications, e.g., Google Maps, and crowd-sourcing data platforms. In tandem, computing systems using the graphics processing unit (GPU) facilitating browser-based rendering using HTML, WebGL, and Python programming constructs have enabled the implementation.

The future of geovisualizations lie in using them in more complex usage scenarios, with a renewed focus on geoscientific education. Visualizations are consumed using augmented and virtual reality which can be used effectively, e.g., augmented reality sandbox kit for students to interactively build landscape and understand water flow modeling on such landscapes. Current visualizations also focus a lot on the raw data, and there is a large space for seamless integration of visualizations and data transformations. One of the immediate applications is in understanding uncertainty in the data itself, which needs to be studied further, given that “data is the new oil.”

Overall, data visualization is an analytic process which provides insights to data using human-computer interaction, in applications providing data summarization and data exploration in varying degrees. In this entry, we have introduced the topic of data visualization in the context of mathematical geosciences, described common visualization techniques, discussed about its current state in the context of usability and computations, and, finally, impressed on the challenges and opportunities for geovisual analytics practitioners.

Bibliography

- Andrienko G, Andrienko N, Jankowski P, Keim D, Kraak MJ, MacEachren A, Wrobel S (2007) Geovisual analytics for spatial decision support: setting the research agenda. *Int J Geogr Inf Sci* 21(8):839–857
- Brewer CA (1994) Color use guidelines for mapping. In: *Visualization in modern cartography*, pp 123–148. Pergamon
- Çöltekin A, Bleisch S, Andrienko G, Dykes J (2017) Persistent challenges in geovisualization—a community perspective. *Int J Cartogr* 3: 1–25
- Conrad O, Bechtel B, Bock M, Dietrich H, Fischer E, Gerlitz L, Wehberg J, Wichmann V, Böhner J (2015) System for automated geoscientific analyses (SAGA) v. 2.1. 4. *Geoscientific Model Development* 8(2), pp 1991–2007. Copernicus GmbH
- Kehrer J, Hauser H (2012) Visualization and visual analysis of multifaceted scientific data: a survey. *IEEE Trans Vis Comput Graph* 19(3): 495–513
- MacEachren AM, Kraak MJ (2001) Research challenges in geovisualization. *Cartogr Geogr Inf Sci* 28(1):3–12
- Munzner T (2014) *Visualization analysis and design*. CRC Press, Boca Raton, FL, USA
- Sagar BSD, Cheng Q, Agterberg F (2018) *Handbook of mathematical geosciences: fifty years of IAMG*. Springer Nature, Cham, Switzerland
- Xie C, Li M, Wang H, Dong J (2019) A survey on visual analysis of ocean data. *Vis Inform* 3(3):113–128

Database Management System

Rajendra Mohan Panda¹ and B. S. Daya Sagar²

¹Geosystems Research Institute, Mississippi State University, Starkville, MS, USA

²Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, Karnataka, India

Definition

Data management is a process to acquire, store, process, and apply data effectively without compromising the safety and integrity of data use and distribution.

Terminology

An archive is a storage space to keep important information, documents, or objects.

Archiving is the process of data duplication to create new copies for Pen Drive, Hard Disk, other external storage, cloud servers like iCloud, Google Drive, OneDrive, etc.

Compression ratio is the ratio of uncompressed file size to the compressed file size

Data Analysis is the process of understanding data to derive useful information.

Data Archiving is a data backup process.

Data are a collection of facts or figures or information.

Metadata is the information about the data

Introduction

Data management started with punch cards. It came to the limelight with the establishment of the Association of Data Processing Service Organizations (ADPSO) in 1960 (SAS 2022). The data management systems were strictly operational. With the internet, digital data use exponentially increased and these voluminous data needed systematic monitoring in a fast and flexible way without compromising safety. This data monitoring process involves the acquirement, storage, processing, and applications effectively. It represents the overall control policies or protocols of an organization for the productive use of data with focuses on data integrity, governance, quality, and security and storage (USGS 2022). Different organizations have different policies to use and share data and hence, separate data management plans for wise use data. Overall, data management, which involves the collection, secure execution, and monitoring of

data, has a life cycle starting from its inception to final retirement including its progress and failures. Data management has several benefits in multiple applications:

- Strategic decision making
- Good governance
- Greater collaboration
- Increase visibility
- Marketing and networking
- Operation management
- Improving organizational agility
- Prioritizing privacy
- Safeguard piracy
- Secure use
- Sustainable development
- Organizational growth
- Systematic reporting
- Timely compliance

The best data management practices focus on:

- Metadata Management
- Data Quality
- Data Security
- Data Integration
- Data Governance
- Data Storage

Metadata Management

Metadata is information about data. It describes the form, structure, types, attributes, records, summaries, and other key features of data. Its management involves generating, filtering, and searching metadata, which helps us to understand data for better usability. Metadata management facilitates consistent data enhancement processes. Broadly, metadata management is classified into operational, technical, and business categories.

Quality

Data quality describes the correctness and consistency of data. It helps in identifying data deficiencies or completeness and availability. Data quality management assesses data correctness, consistency, integrity, and frequency of distribution. It includes the completeness, collection time, spatial extent, and background information. Completeness checks about the missing information and accuracy. The data collection details about time, extent, location, coordinates, background information, and any other specification that may be essential for proper data assessment. Data quality management has a significant role in better data integrity, quality checking, and also, its improvement.

Security

Data security deals with the protection of data use and release processes. Its management focuses on conditions of access, use, and sharing data. Data security management complies with the General Data Protection Regulations (GDPR). It includes access control, auditing, masking, and tokenization. *Access* is a secure login process with two- or three-factor authentication to access data with permissions. DUO authentication is a common practice in secure login, common practice in cloud computing. Auditing deals with record keeping and documentation to trace data on demand. Unique identifiers are assigned for data for easy auditing. Data security management has significant control over the deletion or substitution of data based on preset rules. It helps organizations streamline data use conveniently and effectively with protection.

Integration

Data integration is data consolidation from different sources into a single dataset providing users with consistent access and delivery of data across multiple spectrums to justify the information needs of all applications and stakeholders (Lenzerini 2002).

Governance

Data governance includes analysis, classification, and maintenance of data. It covers data structure, its use, and distribution. It is required for the smooth function of an organization. A well-defined data management system helps in the effective development of an organization.

Storage

Storage is a significant part of data management, which deals with frequent data export and backup of newly created or modified data files. Data compression is a common method to store more data using less space.

Data Compression

Compression reduces the storage space of files or folders by compression. A JPEG file can reduce a little as 80% and MP3 by 90%. But single file compression suffers the loss of data transparency, color, sound, etc. Alternately, folder compression reduces data storage space without the loss of a single bit. The data compression procedures for multiple files to a single file are called “archive” and formats used for storage space reduction are called “archive formats.”

Archive Formats

Several archive formats are used for data compression. The efficiency of different data archiving formats is measured using two terms: compression ratio and compression speed. Data compression ratio is the relative reduction in data size by

applying a compression algorithm, expressed as a ratio of uncompressed size to compressed size:

$$\text{Compression ratio} = \frac{\text{Uncompressed data size}}{\text{Compressed data size}}$$

$$\text{And, \% space savings} = \left(1 - \frac{\text{Compressed data size}}{\text{Uncompressed data size}}\right) \times 100$$

$$= 1 - \frac{\text{Number of bits after compression}}{\text{Number of bits before compression}}$$

During compression and decompression, some memory is always consumed, which after completing regains the memory (IBM 2022). The speed is a simple ratio of energy bits before or after compression to time (s):

$$\text{Compression speed} = \frac{\text{Compressed bits}}{\text{Compression time}}$$

$$\text{Decompression speed} = \frac{\text{Decompressed bits}}{\text{Decompression time}}$$

The major formats include Zicxac Inline Pin (ZIP), Roshal Archive (RAR), Tape Archive (TAR), GZIP, and 7Z. ZIP is the most commonly used archive format with the highest compression speed and minimum compression ratio compatibility with Windows and Linux systems. ZIP uses DEFLATE for data compression, and WinZip, WinRAR, and 7-Zip for extraction. It stores and extracts files separately. RAR is another popular data compression method, compatible with Windows and macOS. It can compress files by using the WinRAR application but extraction is possible both by WinRAR and WinZip. It has a better compression feature than ZIP. 7Z has the highest compression ratio and minimum compression speed. It is compatible with Windows, Linux, and macOS. It supports multiple applications for compression (LZMA, PPMD, BCJ, Bzip2, Deflate), extraction (7-ZIP), and encryption (AES-256). Its program and file formats are available for free. TAR.BZ2 and TAR.GZ also have open standard licenses like 7Z. TAR.BZ2 is not better than 7Z although its compression rate and speed are better than ZIP and TAR.GZ. BZIP2 and GZIP applications are used to compress TAR archive files to TAR.BZ2 and TAR.GZ. TAR.BZ2 has an intermediate compression speed and compression ratio. TAR.GZ has the highest compression speed and low compression rate. Both support storage of files in Windows, Linux, macOS. It uses GZIP to compress files, but with high compression speed and compression ratio.

Storage Devices

Storage devices are used for data backup or recovery. Magnetic tapes or disks, CDs, DVDs, and Universal Serial Bus (USB) are used for such backup and storage. USB is now common for its higher bandwidth and power supply capability. External hard drives and other drives of the same computer are often used to copy data.

Backup Types

System backup or file backup are two major backup types. The system backup stores everything including operating system and applications, whereas file backup keeps copies of files. The file backup includes differential or full backup and incremental or delta backup. As the name suggests, differential backup copies selected files, unlike full file backup. The incremental file backup appends the selected files without replacing the previous files, but delta backup stores the change portion of the data. The automatic storage of data is often required to transfer data from one storage medium to another based on some predetermined criteria: age, category, etc. Solid State Drives (SSDs) are popular devices for such storage activities. It is preferred to hard disks for having better holding capability to the frequently used datasets, high storage speed, greater reliability.

Cloud Storage

Cloud computing is a recent phenomenon in the internet world. It is a replacement for in-house operations offering service online. It started as a software service for providing online data management in a user-specific computer. Cloud servers are capable of managing huge data traffic with provisions for expansion and contraction. Several service providers: Amazon (Amazon Drive, Flickr), Google (Chromebook, Google Drive, Google Docs, G Suite, Google I, Google Photos), Microsoft (Azure, OneDrive), Dropbox (Dropbox), IBM (IBM Cloud), and Apple (iCloud, iCloud Drive) have revolutionized these online applications replacing large dependencies of in-house data management. These remote internet services for data storage, processing, and management combined with the infrastructure arrangements are *cloud computing*.

Cloud computing uses artificial intelligence (AI) and machine learning to automate multiple data management tasks simultaneously. It offers benefits for data backups, security, and governance such as

- Complexity reduction
- Accuracy improvement
- Increased reliability
- Secured use
- Operational efficacy
- Reduced operational cost
- High-performance

Big Data Management

Big data management is complex and involves more sophisticated technology for smooth and automated acquisition,

storage, transfer, and usage. Lyko et al. (2016) have described the following steps for big data management:

Data acquisition $\Rightarrow$ Data Analytics $\Rightarrow$ Data Curation $\Rightarrow$ Data Storage $\Rightarrow$ Data Usage

Data can be structured or unstructured, simulated or satellite based. These data are systematically verified by analytical approaches: machine learning, stream mining, and semantic analyses. Data analytics involves ecosystem, community, cross-sectoral, or global analyses. Data curation is one of the most significant steps in big data management. It includes data quality, data accuracy, automation, and interpretability. The next important step is data storage, which takes care of performance, security, consistency, and availability. The access and availability are determined by certain standardized protocols designed by each organization. Finally, data helps in wise decision making, predictions, simulations, explorations, and visualization, etc. Cloud computing is believed to be the best available solution for big data management.

Remote Sensing Data Management

Remote sensing data are huge and their acquisition, processing, storage, analysis, and visualization are critical for monitoring soil properties and crop production. Its management is significant for decision support in fertilization, irrigation, and pest control (Huang et al. 2018). The significance of certain bands of electromagnetic spectrum, e.g., infrared and microwave in vegetation and soil moisture (Moran et al. 1997; Mulla 2013) and to other features necessarily demands systematic handling of big data. This involves storage, processing, extraction, and visualization in a time-specific, fast, and realistic manner, essential for precision agriculture.

In general, remote sensing data acquired from different satellites or other sources (e.g., aircraft, drones) have noise. These noises appear due to interactions between the atmosphere, terrain features, and sensors. These errors are corrected before further data processing. Sometimes these data are resampled or rescaled to make some adjustments to the requirements. Image fusion is often needed to create a common coverage for images from multiple sources and scales or resolutions. The fusion that enhances image quality is commonly applied at a pixel level. The images are classified applying various algorithms such as maximum likelihood, machine learning, and object-based classification (Walter 2004; Huang et al. 2018). Deep learning has also been applied to remote sensing data classification (Mohanty et al. 2016).

Classified images are visualized by using the GIS platform, which has immense capabilities of voluminous data, collaboration, integration, and systematic application. However, problems of data integration, validation, and accuracy assessment need further research for complex remote sensing data management.

Conclusions

Data management is critical for organizations to find data reliability and its optimal use. It is a standard process essential for understanding data for good governance and meaningful decision making. It becomes more complex with big data acquisition. Its curation, storage, and usage need systematic management practices to achieve trusted data acquisition, distribution, and automation. The Internet revolution has solved these problems of big data management by replacing in-house operations. Cloud computing is the recent phenomenon that allows third-party management of these huge data resources. Remote sensing data management is complex with issues related to accuracy, validation, and integration. Machine learning and deep learning algorithms in GIS platforms are continuously adding new features. The technology development for error minimization and improvements in future data management is not so far.

Bibliography

- Huang Y, Chen ZX, Tao YU, Huang XZ, Gu XF (2018) Agricultural remote sensing big data: management and applications. *J Integr Agric* 17(9):1915–1931
- IBM (2022) <https://www.ibm.com/docs/en/spectrum-symphony/7.3.0?topic=performance-data-compression>
- Lenzerini M (2002). Data integration: a theoretical perspective (PDF). PODS 2002, pp 233–246
- Lyko K, Nitzschke M, Ngomo AC (2016) Big data acquisition. In: *New horizons for a data-driven economy*. Springer, Cham, pp 39–61
- Mohanty SP, Hughes D, Salathe M (2016) Using deep learning for image-based plant disease detection. *Front Plant Sci* 7:1–10
- Moran MS, Inoue Y, Barnes EM (1997) Opportunities and limitations for image-based remote sensing in precision crop management. *Remote Sens Environ* 61:319–346
- Mulla DJ (2013) Twenty-five years of remote sensing in precision agriculture: key advances and remaining knowledge gaps. *Biosyst Eng* 114:358–371
- SAS (2022) https://www.sas.com/en_us/insights/data-management/data-management.html
- USGS (2022) <https://www.usgs.gov/products/data-and-tools/data-management/data-acquisition-methods>
- Walter V (2004) Object-based classification of remote sensing data for change detection. *ISPRS J Photogramm Remote Sens* 58: 225–238

David, Michel

Denis Marcotte

Department of Civil, Geological and Mining Engineering,
Polytechnique Montreal, Montreal, QC, Canada



Fig. 1 Michel David (right) accepts Krumbein Medal from IAMG President John Davis (Source: JAMG Newsletter 38)

Biography

Michel David was born in France in 1945 and received his first degree in 1967 as *Ingénieur civil des mines* from *Université Lorraine* in Nancy. Trained in geostatistics by George Matheron, Michel David obtained his Ph.D. at Montreal University (1973) in the department of computer sciences and operational research. He was professor of geostatistics (1970–1987) in the mineral engineering department at Polytechnique Montreal, the first university in the Americas to include a geostatistical course in a regular curriculum. He was coorganizer and coeditor of the first two International Geostatistical Congresses in Frascati, Italy, in 1975 and Lake Tahoe, Nevada, in 1983.

Despite the relative brevity of his academic career, he is recognized as one of the key actors for the initial spreading of geostatistical methods among practitioners of the mining industry. In particular, Michel David wrote the first English language textbook on geostatistics applied to mining problems, *Geostatistical Ore Reserve Estimation* (David 1977). A second book, *Handbook of Applied Advanced*

Geostatistical Ore Reserve Estimation, was published in 1988. He was a leader in international geostatistical consulting for mining companies. Furthermore, he became director (1976–1981) of the *Mineral Exploration Research Institute*, and he founded *Geostat Systems International*, a geostatistical consulting company having antennas in Denver (1979) and Montreal (1981). Michel David was proud of defining himself as a teacher, etymologically “a person who shows.” Hence, in addition to his regular courses, this love for teaching lead him to give numerous short courses on geostatistics reaching over a thousand mining professionals around the world.

Michel David’s research was applied to and mainly rooted in typical mining problems. Illustrative examples are the statistical ore dilution problem (the “vanishing tons”) due to support effect and smoothing effect, and the importance of conditional bias encountered with many estimation methods. He was sensitive to the importance of communicating simply the implications, for mining practice, of abstract geostatistical concepts. He also proposed useful empirical tools for practitioners, such as the relative variogram (general and pairwise) and the lognormal shortcut method for estimation of local recoverable reserves. Among other subjects worth mentioning that attracted his attention are the automatic mapping with intrinsic random functions of order k and various geological applications of multivariate classification and of correspondence analysis, a factorial method particularly well suited for the analysis of ubiquitous composition data in geology.

Michel David died in 2000. The importance of his role and contribution to geostatistics and resource evaluation was recognized by the attribution, in 1987, of the W.C. Krumbein medal of the International Association for Mathematical Geology and by his nomination as Fellow of the Academy of Science of the Royal Society of Canada in 1988. Moreover, an international Forum was held in his honor in 1993 in Montreal and a special issue of *Mathematical Geology* was devoted to Michel David in 2005 as a tribute to his legacy.

Michel David (right) receiving the W.C. Krumbein International Association of Mathematical Geosciences (IAMG) medal from J. Davis, IAMG Newsletter No. 38 (used with permission from the IAMG)

Bibliography

- David M (1977) *Geostatistical ore reserve estimation*. Elsevier, Amsterdam. 364 p
- David M (1988) *Handbook of applied advanced geostatistical ore reserve estimation*. Elsevier, Amsterdam. 217 p

Davis, John Clements

Ricardo A. Olea

Geology, Energy and Minerals Science Center,
U.S. Geological Survey, Reston, VA, USA



Fig. 1 John Clements Davis. (Courtesy of Jo Anne Degraffenreid)

Biography

John Davis was among those who first used computers to address geological problems, thus helping to transform geology from a purely descriptive to a more quantitative science. His early work on computer mapping and application of multivariate statistics to geological data led to Davis’s most influential publication, his textbook *Statistics and Data Analysis in Geology* describing these and other quantitative methodologies. The book was widely accepted and was revised through three editions paralleling the evolution of information technology. The first edition (Davis 1973) included simple FORTRAN routines for methods discussed in the book. Enhanced versions came on a diskette in the 1986 second edition, for the third edition in 2002 files could be downloaded over the Internet. Davis was also active in the probabilistic assessment of petroleum prospects and evaluation of geologic hazards, and authored over 100 papers.

John joined the Kansas Geological Survey (KGS) in 1967 and remained there through 2003. As Head of the Mathematical Geology Section he had a lasting influence locally and worldwide. He maintained a “Visiting Scientist” position that allowed his group to host for one or two years some of the most renowned specialists in geomathematics and computer applications, including several of the biographees in this encyclopedia. This cooperation was critical in the development and testing of many techniques that today are routinely used in the earth sciences. Davis also served as professor of petroleum engineering at the University of Kansas and later at the Montanuniversität in Austria, teaching several generations of young geologists and engineers.

John Davis was born in Neodesha, Kansas, on 21 October 1938. He graduated from the University of Kansas with a BS in geology, then moved to the University of Wyoming as an

NDEA Fellow, earning an MS in 1963 and a PhD in 1967, both in geology. John was generous with his time and talents by being active in several professional organizations. He served on several early committees promoting computer applications to geology and taught industry short courses for the American Association of Petroleum Geologists (AAPG). John was Western Treasurer of the International Association for Mathematical Geosciences (IAMG) for 8 years, then was elected secretary-general in 1980, and served as president from 1984–1989. The IAMG presented Davis with the Krumbein Medal in 1986 and the Geological Survey of Austria bestowed on him the Wilhelm Ritter von Haidinger Medal in 1999, both for outstanding contributions to the profession.

Bibliography

- Davis JC (1973) *Statistics and data analysis in geology*. Wiley, New York. 1st edn. 550 pp; 2nd edn., 1986, 646 pp. and a diskette; 3rd edn., 2002, 638 pp
- Harbaugh JH, Davis JC, Wendebourg J (1995) *Computing risk for oil prospects*. Pergamon Press, New York. 452 pp
- Ohlmacher GC, Davis JC (2003) Using multiple logistic regression and GIS technology to predict landslide hazard in Northeast Kansas, USA. *Geol Eng* 69(3–4):331–343

De Wijs Model

Shuyun Xie

State Key Laboratory of Geological Processes and Mineral Resources (GPMR), Faculty of Earth Sciences, China University of Geosciences, Wuhan, China

Definition

De Wijs model is a typical multifractal model, which was introduced by De Wijs (1951, 1953) to study the spatial distribution, redistribution, or enrichment/depletion patterns of elements based on multiplicative cascade processes, and has now been widely applied in different fields and especially played a pivotal role in modern geostatistics (Diggle et al. 2010; Mondal 2015).

Introduction

In the early 1950s, H. J. De Wijs (De Wijs 1951, 1953) discovered a well-known distribution model in studying the element distribution of minerals in rocks, which was later called De Wijs model. The self-similarity criterion contained in the model makes it considered as a fractal model (Mandelbrot 1975), more precisely, a typical multifractal model (Evertsz and Mandelbrot 1992). The data published

by De Wijs (1951) is also called De Wijs data, which has been regarded as the classical fractal data set in Geostatistics. Since 1960s, De Wijs model and data have been widely and deeply studied (Krige 1952, 1958; Agterberg, 1974, 1981, 2005, 2007, 2012; Xie and Bao 2004). Matheron (1962) proved that De Wijs model data would produce logarithmic variogram. The data set is proved to obey multifractal distribution, and the multifractal variation function model in theory is proposed based on the De Wijs data (Cheng and Agterberg 1996). Cheng (2000, 2015a) first studied the local singularity analysis and multifractal properties of De Wijs data.

De Wijs model and data has also been used to test the efficiency of various multifractal methods before dealing with different geochemical data sets (Chen et al. 2007; Gao, 2010; Cheng 2015a). Using the binomial De Wijs model (1951) as an example, the issues of microcanonical versus canonical conservation, algebraic (“Pareto”) versus long tailed (e.g., lognormal) distributions, multifractal universality, conservative and non-conservative multifractal processes, and codimension versus dimension formalisms were discussed deeply (Lovejoy and Schertzer 2007). The relatively simple one-dimensional De Wijs multiplicative cascade model has also been used to discuss the lacunarity for fractals, multifractals, and some nonfractals, introduce the relation between multiplicative cascade processes and information integration processes, and demonstrate the formation of fractal density (Cheng 1997, 2012, 2015b). The De Wijs scheme has also been followed to analyze mineral production data and to simulate the distribution of deposits and for mineral resource appraisal (Ford and Blenkinsop 2009; Gumiel et al. 2010; Arias et al. 2011; Agterberg 2018).

De Wijs Data

Analytical assay data for 118 Zn contents from a sphalerite-quartz vein near Pulacayo in Bolivia, as shown in Table 1 (De Wijs 1951), were published. These 118 channel samples were cut every 2 m in a 240-m-long vein crossing horizontal block of sphalerite-quartz veins (446 m underground). The massive veins were only 0.5 m wide on average and contained disseminated sphalerites in the surrounding rocks on both sides (Agterberg 1974; Chen et al. 2007). These channel samples provided unbiased estimates of zinc concentrations in 2-m-long blocks analyzed along the vein direction. Thus, the 118 geochemical data of Zn contents represent a standard data set which was measured from the samples collected at standard equidistant distances. Although the number of sampling points is very small, as a first example of an in situ horizontal analysis in ore-forming areas, the famous zinc concentration assay values (de Wijs 1951) have been studied by scholars many times and also reanalyzed again and again (Lovejoy and Schertzer 2007).

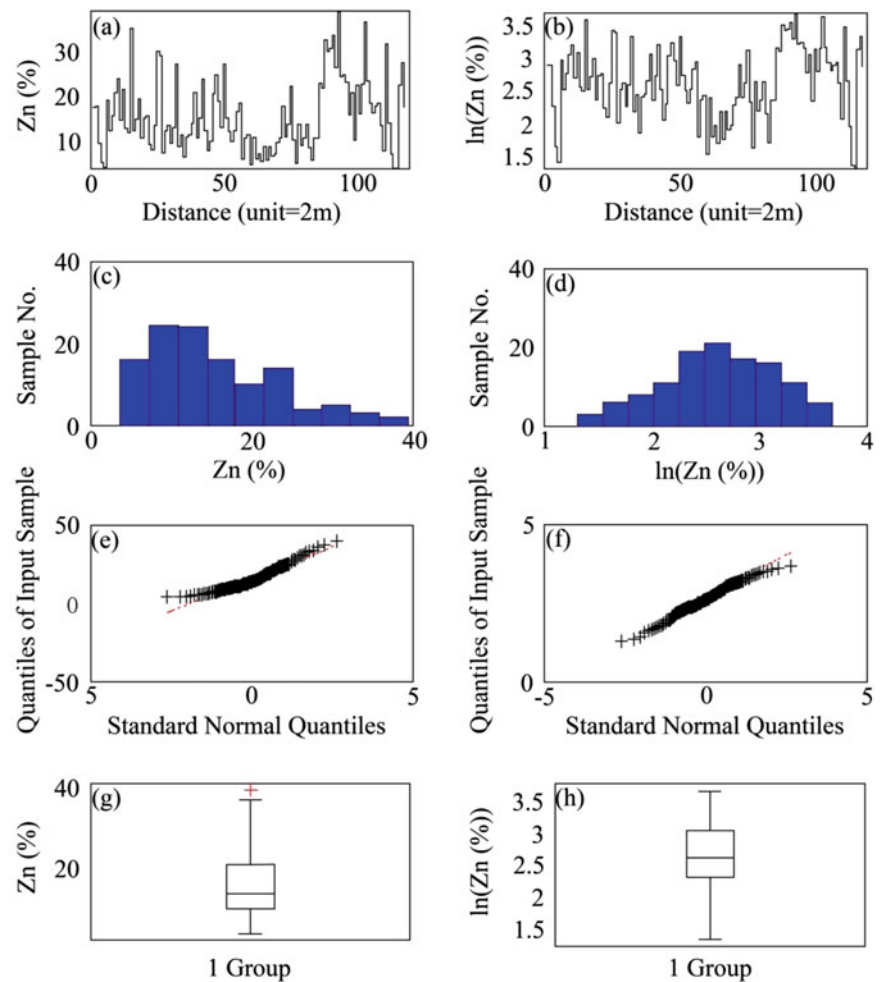
Figure 1 shows the statistical distribution plots of De Wijs Data set as shown in Table 1. Figure 1(a) is the Zn content (%) of 118 Samples, and Figure 1(b) shows the value of natural

De Wijs Model, Table 1 Analysis data of Zn content in 118 samples from Pb-Zn quartz veins (%)

No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Zn	17.7	17.8	9.5	5.2	4.1	19.2	12.4	15.8	20.8	24.1	14.7	21.6	12.8	11.9	35.4	12.3	14.9	19.6	10.6	15.1
No.	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Zn	15.6	9.3	8.1	13.5	30.2	29.1	7.4	12.3	13.6	9.5	13.1	27.4	8.8	11.4	6.4	11	11.4	14.1	20.9	10.6
No.	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
Zn	15.3	24	12.3	7.8	9.9	20.7	25	19.1	13.1	27.4	15.2	12.2	10.1	12.3	16.7	18.6	6	10.6	11.3	4.7
No.	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
Zn	10.9	6	7.2	5.5	8.9	5.8	8.9	6.7	7.2	9.7	10.8	17.9	10.9	13.7	22.3	10.2	5.1	13.9	9	10.6
No.	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
Zn	13.8	8.5	5.6	10.6	10.6	23	21.8	32.8	30.2	30.8	33.7	26.5	39.3	24.5	24.9	23.2	16	20.9	10.3	22.6
No.	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120
Zn	16.2	22.9	36.9	23.5	18.5	16.4	17.9	18.5	13.6	7.9	31.9	14.1	7.1	3.9	3.7	22.5	27.6	17.3		

De Wijs Model,

Fig. 1 Distribution of De Wijs Data set (a) Zn content (%) of 118 Samples; (b) Value of natural logarithm of Zn content; (c) Histogram of Zn content; (d) Histogram of natural logarithm of Zn content; (e) Q-Q plot of Zn content; (f) Q-Q plot of natural logarithm of Zn content; (g) Boxplot diagram of Zn content; and (h) boxplot diagram of natural logarithm of Zn content



logarithm of Zn content. Correspondingly, Figure 1(c) is the Histogram of Zn content; Figure 1(d) is the Histogram of natural logarithm of Zn content; Figure 1(e) is Q-Q plot of Zn content; Figure 1(f) is Q-Q plot of natural logarithm of Zn content; Figure 1(g) is Boxplot diagram of Zn content; Figure 1(h) is boxplot diagram of natural logarithm of Zn content. Obviously, De Wijs showed that the 118 zinc values were approximately log-normally distributed as shown in Figure 1.

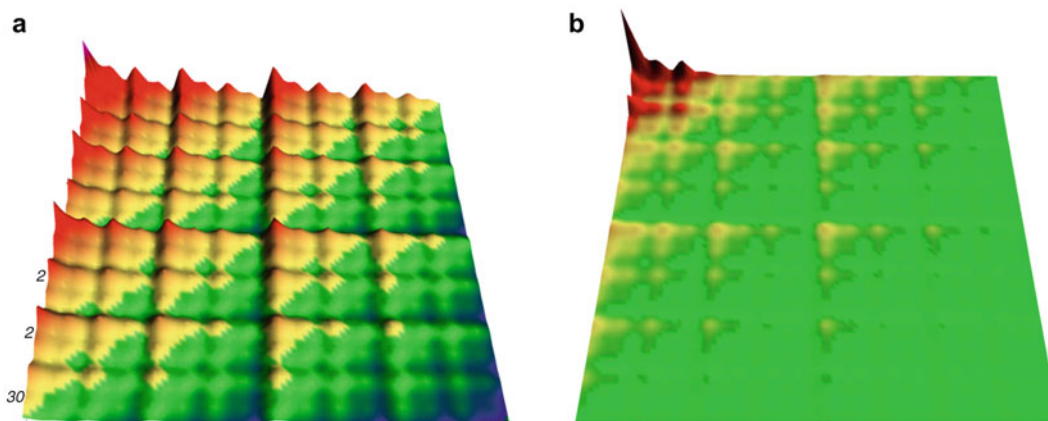
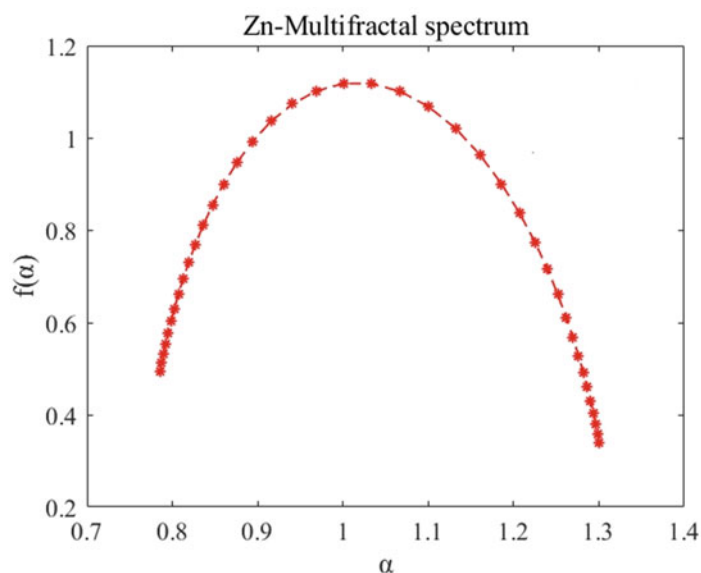
Figure 2 is the multifractal spectrum curve of Zn content distribution which is calculated by the method of moments (Halsey et al. 1986; Evertsz and Mandelbrot 1992). It can be seen clearly that the data set is of continuous multifractal property (Cheng and Agterberg 1996).

Selected Case Study of De Wijs Model

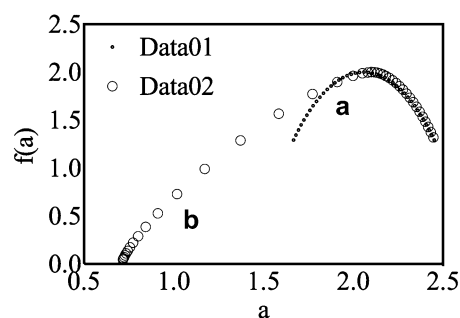
De Wijs model describes the iterative segmentation process of metal element distribution of self-similarity. Assuming that the ore vein with volume V initially has average grade Z_v , it is divided into two blocks on average, with grade values of $(1 + d) Z_v$ and $(1 - d) Z_v$ respectively; $0 \leq d \leq 1$. d , also called dispersion index, is the basic constant of the model,

which has nothing to do with the size of block V . Therefore, the metal content is relatively enriched in one subarea and relatively depleted in the other. After repeating the above two operations for n times, the block volume V is continuously cut out, and the relative enrichment or depletion degree of the metal content in it becomes higher and higher, and the overall distribution of grade becomes extremely uneven. Theoretically, it obeys the logbinomial distribution, and their logarithms have dispersion variance (De Wijs 1951; cf. Matheron 1962, p. 309). The construction of De Wijs model depends on three parameters: the overall average concentration Z_v , the dispersion index d , and the number of segmentation n . The relationship between the extreme value of multifractal singularity index $\alpha_{\min}$ and the parameter d of de Wijs model is proved (Agterberg 2001). The multifractal characteristics of the De Wijs model are studied, and the degree of enrichment d is identified by one multifractal detrending moving average analysis (MFDMA) method (Lai and Wan 2016).

For the De Wijs model in two-dimensional space, the multifractal patterns based on De Wijs multiplicative cascade models can be constructed as follows. Without loss of generality, the initial concentration Z_v is set as 1 in one region. Then the region can be divided horizontally into left and right

De Wijs Model,**Fig. 2** Multifractal spectrum
Curve of Zn content distribution
from De Wijs data**De Wijs Model, Fig. 3** Modeling De Wijs data with or without local superimposition of different dispersion indexes. (a) Data 01: $d = 0.2$; (b) Data 02: $d_1 = 0.2, d_2 = 0.6$

subregions first, then the two subregions are further vertically divided up and down, and thus the subregions are continuously divided into four parts, and the increasingly uneven singularity distribution in the whole space can be obtained. Keeping the dispersion index d constant, the concentration values are now $(1 + d)^2$, $(1 + d)(1 - d)$, $(1 - d)(1 + d)$, and $(1 - d)^2$ after first iteration. According to this iterative rule, the four small subregions can still be divided into four smaller regions. This process can be repeated indefinitely, but in fact, after only a few rounds, the data is dense enough for us to study the spatial distribution patterns of geochemical data. Figure 3(a) is the result of simulated data after five iterations when we choose the dispersion index d of 0.2. Figure 4(a) is the multifractal analysis results of De Wijs modeling data when $d = 0.2$. It can be seen that such kind of construction results based on De Wijs modeling are of typical symmetrical multifractal spectrum curve.

**De Wijs Model, Fig. 4** Multifractal analysis results of De Wijs modeling data with or without local superimposition of different dispersion indexes. (a) Data 01: $d = 0.2$; (b) Data 02: $d_1 = 0.2, d_2 = 0.6$

De Wijs model is also used to simulate the geological process of superimposition of late stages. In the long history

of geological evolution, such as in the process of mineralization, materials are often considered to have different stages of local superposition. A current geochemical field is the final product of a series of geological and geochemical processes in a specific area. In each process, ore-forming materials from different sources may be added to form geochemical anomalies in the geochemical background. In order to simulate this process, De Wijs data set based on the dispersion index d_1 of 0.2 is considered as a typical background field, and the simulation results are locally superimposed on the upper left corner of the model by another De Wijs data set with the dispersion index d_2 of 0.6. The overlay area is arranged in $1/16$ or 32×32 grids of the total area. Figure 2 and Figure 4(b) show the De Wijs dataset and multifractal spectrum curve with local superposition effect, respectively. Obviously, in Figure 2, the later superimposition implies more obvious local singularity and is easier to form strong geochemical anomalies. Figure 4(b) shows that the original multifractal spectrum curve becomes asymmetric and presents a left deviation pattern due to the superposition of late processes, which is considered to be a favorable metallogenic mode (Xie and Bao 2004).

Some new methods have been proposed to make the De Wijs model more universally applied. For example, the generator matrix of De Wijs model is set as a nonsquare matrix, so that self-affine multifractal distribution patterns can be generated. In addition, before each segmentation, the positions of each element in the generator matrix are randomly changed or changed with some constraint conditions, so as to generate random multifractal distribution patterns (Agterberg 2007). By changing the parameter d in the model construction, De Wijs model is also derived from the random variables to generate isotropic geochemical data and anisotropic geochemical data of different spatial distribution patterns (Chen et al. 2007).

Summary and Conclusions

Such kind of data construction process of De Wijs model is one of typical multiplicative cascade processes which captures the basic features of the forming processes of geochemical fields through many rounds of material redistribution, enriched somewhere and depleted elsewhere. And hence, it has been used as a conceptual but idealized model to simulate the multifractal geochemical fields with both isotropic and anisotropic backgrounds and anomalies, which played a vital role in modern geostatistics and mathematical geosciences area.

Bibliography

- Agterberg FP (1974) Geomathematics: mathematical background and geo-science application. Elsevier scientific publishing company, Amsterdam
- Agterberg FP (1981) Cell-value distribution models in spatial pattern analysis. In: Craig RG, Labovitz ML (eds) Future trends in geomathematics. Pion Limited, London
- Agterberg FP (1989) Computer programs for mineral exploration. Science 245:76–81
- Agterberg FP (2001) Multifractal simulation of geochemical map patterns. J China Univ Geosci 12(1):31–39
- Agterberg FP (2005) Application of a three-parameter version of the model of de Wijs in regional geochemistry. In: Cheng QM, Bonham-Carter GF (eds) Proceedings of IAMG 5: GIS and spatial analysis, held in Toronto, Canada, August 21–26, 2005, vol 1. China University of Geosciences Printing House, Wuhan, pp 291–296
- Agterberg FP (2007) New applications of the model of de Wijs in regional geochemistry. Math Geol 39(1):1–25
- Agterberg FP (2012) Sampling and analysis of chemical element concentration distribution in rock units and orebodies. Nonlinear Process Geophys 19:23–44
- Agterberg (2018) Can multifractals be used for mineral resource appraisal? J Geochem Explor 189:54–63
- Arias M, Gumiel P, Sanderson JD et al (2011) A multifractal simulation model for the distribution of VMS deposits in the Spanish segment of the Iberian Pyrite Belt. Comput Geosci 37:1917–1927
- Chen Z, Cheng Q, Xie S (2007) A novel iterative approach for mapping local singularities from geochemical data. Nonlinear Process Geophys 14:317–324
- Cheng Q (1997) Multifractal modeling and lacunarity analysis. Math Geol 29(7):919–932
- Cheng, Q (2000) Interpolation by means of multifractal, kriging and moving average techniques. In: CD Proceedings of GAC/MAC meeting GeoCanada 2000, 29 May–2 June, Calgary. <http://www.gisworld.org/gac-gis/geo2000.htm>
- Cheng Q (2012) Multiplicative cascade processes and information integration for predictive mapping. Nonlinear Process Geophys 19: 57–68
- Cheng (2015a) Multifractal interpolation method for spatial data with singularities. J S Afr I Min Metall 115:235–240
- Cheng (2015b) Fractal density and singularity analysis of heat flow over ocean ridges. Sci Rep 6:19167. <https://doi.org/10.1038/srep19167>
- Cheng Q, Agterberg FP (1996) Comparison between two types of multifractal modeling. Math Geol 28(8):1001–1016
- De Wijs HJ (1951) Statistics of ore distribution: (1) frequency distribution of assay values. Geol Mijnb 13:365–375
- De Wijs HJ (1953) Statistics of ore distribution: (2) theory of binomial distribution applied to sampling and engineering problems. Geol Mijnb 15:12–24
- Diggle PJ, Menezes R, Su T (2010) Geostatistical inference under preferential sampling. Appl Stat 59(2):191–232
- Evertsz CJG, Mandelbrot BB (1992) Multifractal measures (appendix B) [A]. In: Peitgen H-O, Jurgens H, Saupe D (eds) Chaos and fractals [C]. Springer Verlag, New York, pp 922–953
- Ford A, Blenkinsop GT (2009) An expanded de Wijs model for multifractal analysis of mineral production data. Mineral Deposita 44(2): 233–240
- Gao X (2010) Multifractal modeling of 2D geochemical landscape based on Wavelet Leaders. China Mining Magazine 19(6):111–115. [in Chinese with English Abstract]

- Gumiel P, Sanderson JD, Arias JM et al (2010) Analysis of the fractal clustering of ore deposits in the Spanish Iberian Pyrite Belt. *Ore Geol Rev* 38:307–318
- Halsey TC, Jensen MH, Kadanoff LP et al (1986) Fractal measures and their singularities, the characterization of strange sets. *Phys Rev A* 33:1141–1151
- Krige DG (1952) A statistical analysis approach to some basic borehole values in the Orange Free State goldfield. *J Chem Metall Min Soc S Afr* (September):47–64
- Krige DG (1958) Lognormal-de Wijsian geostatistics for ore evaluation, Monograph series. South African Institute of Mining and Metallurgy, Johannesburg
- Lai J, Li W, Xiong X (2016) Multifractality analysis of De Wijs model based on multifractal detrending moving average analysis. *J Hunan Univ Art Sci* 28(2):4–10. [in Chinese with English Abstract]
- Lovejoy S, Schertzer D (2007) Scaling and multifractal fields in the solid earth and topography. *Nonlinear Process Geophys* 14:465–502
- Mandelbrot BB (1975) Stochastic models for the Earth's relief, the shape and the fractal dimension of the coastlines, and the number-area rule for islands. *Proc Natl Acad Sci U S A* 72:3825–3828
- Matheron G (1962) *Traité de Géostatistique Appliquée*. Mémoires Bur Rech: Géol Minières 14:33
- Mondal D (2015) Applying Dynkin's isomorphism: an alternative approach to understand the Markov property of the de Wijs process. *Bernoulli* 21(3):1289–1303. <https://doi.org/10.3150/13-BEJ541>
- Xie S, Bao Z (2004) Fractal and multifractal properties of geochemical fields. *Math Geol* 36(7):847–864

Decision Tree

Rajendra Mohan Panda¹ and B. S. Daya Sagar²

¹Geosystems Research Institute, Mississippi State University, Starkville, MS, USA

²Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, Karnataka, India

Definition

A decision tree is “a regression through classification for data mining and other applications represented with an inverted tree-like structure, where the root at the top is the input and leaves at the bottom are the outcomes or decisions.”

Terminology

Root node: Input for the decision tree splits into internal nodes which ultimately terminate as a decision node. The root node is the parent node of internal nodes.

Internal node: Intermediate nodes (or branches) are child nodes of the root node. Internal nodes split into decision nodes and are parent nodes of decision nodes.

Terminal node: Decision nodes (or leaves) are child nodes of internal nodes. Decision nodes do not split and have no child nodes.

Depth: Number of splits or divisions.

Pruning: Selective removal of internal nodes or sub-nodes.

Splitting: Division of data to nodes or sub-nodes.

Introduction

A decision tree is a machine learning technique for decision-based analysis and interpretation in business, computer science, civil engineering, and ecology (Bel et al. 2009; Debeljak and Džeroski 2011; Krzywinski and Altman 2017). It provides solutions to varieties of regression data mining problems used for decision making and good management practices (Maimon and Rokach 2005; Breiman et al. 2017). A decision tree derives the conclusion of an event through a series of regression and classification. The input for a decision tree is the best predictor and is defined as the root node. The root node splits recursively into decision nodes in the form of branches or leaves based on some user-defined or automatic learning procedures. In principle, a decision tree is a regression through inductive classification. Inductive classification is a computational procedure to extract the relevant information from unknown data through repeated splitting. A decision tree is a common tool for decision making and has many procedural advantages such as (a) simple data preparation, (b) easy interpretation, (c) less vulnerability to outliers, (d) capable of fitting both categorical and numerical variables, and (e) effectiveness in nonlinear setups. A decision tree needs less computational power for model building and low memory catch-ups. Despite several advantages, a decision tree cannot apply to continuous variables of infinite sizes (Rutkowski et al. 2014). The decision tree is often unstable and error-prone. These instabilities in the decision tree are taken care of by bagging and boosting (discussed later).

The decision tree algorithms differ mainly in their formation, e.g., (a) binary or nonbinary and (b) measure of impurity or uncertainty (Rutkowski et al. 2014). Several decision tree algorithms e.g., ASSISTANT, CART (Classification and Regression Tree), C4.5, chi-squared automatic interaction detection (CHAID), ID3 (Iterative Dichotomizer 3), and quick unbiased efficient statistical tree (QUEST), are proposed and some of the most popular methods are discussed. They include ID3 (Quinlan 1986), its extended version C4.5 (Quinlan 1986), and CART (Breiman et al. 1984). ID3 measures the entropy (a measure of uncertainty or randomness), where data split iteratively into many subsets until they belong to the same class. The entropy is calculated using the following formula (where ‘ p ’ is the probability of target class

'i' and 'c' is the number of observations/outcomes/possibilities):

$$\text{Entropy} = - \sum_{i=1}^c p_i * \log_2(p_i)$$

In the advanced version of ID3 (C4.5), the data splitting was based on information gain (difference in entropy values between parent and child nodes), where the data splitting is done by checking the decrease in error rate. Greater weight is given to attributes with large domains and the final class is decided by a majority vote. C4.5 algorithm is based on the gain ratio that selects the best one out of all possible splits. The gain ratio is defined as:

$$\text{Gain ratio} = \frac{\text{Information gain}}{\text{Entropy}}$$

where Information gain = Entropy_{parent node} – Entropy_{child node}.

But this version of C4.5 has the disadvantage of giving importance to rare classes disproportionately. Lewis and Catlett (1994) corrected these discrepancies in C4.5 by the probability threshold of the loss ratio (LR / LR + 1), but not by the majority vote (Lewis and Catlett 1994), where the loss ratio (LR) is a relative cost of false positives or false negatives. The LR = 1 gives equal importance to both errors, the LR > 1 gives more value to the false positive than the false negative, and alternately, an LR < 1 prioritizes false negatives on false positives. The probability threshold of the loss ratio assigns the class values at the decision node by specifying the false positives and false negatives through sensitivity analysis and the maximum normalized information gain.

The CART algorithm is a binary tree based on Gini's impurity index (Breiman et al. 1984). The Gini index measures the deviations between the probability distributions of each split and selects a sample with a minimum loss function. A Gini index value at zero indicates perfect purity and a value at 1 describes the maximum impurity. The Gini index is defined as:

$$\text{Gini Index} = 1 - \sum_{i=1}^c (p_i)^2$$

where 'p' is the probability of target class 'i' and 'c' is the number of observations/outcomes/possibilities.

Based on the type of tree produced, ID3 and C4.5 are nonbinary and CART is binary. ID3 is the fastest and only applicable to categorical variables, whereas C4.5 and CART are efficient in dealing with categorical and numerical algorithms. While ID3 cannot handle missing values, C4.5 and CART can easily handle missing values. On the other hand, CART has an efficient mechanism to eliminate nonsignificant variables, but not C4.5. While the calculation of entropy is slow, the Gini index is faster. Although all three popular

learning mechanisms have some advantages and disadvantages, CART is considered to be simple but most useful in recent research (Krzywinski and Altman 2017).

Stability in a Decision Tree

Decision trees group observations of similar response values, but a small change in observed values or data sample has the propensity to create perturbations in extracting information. These discrepancies in information extraction affect the predictability of a decision tree, and an unusual result might occur without realistic conclusions. To check these instabilities in the predictive performances of a decision tree, many algorithms are used. Bootstrapped aggregation (bagging), boosting or arcing (adaptive resampling and combining), and pruning adjacent nodes of low gain are widely used methods for the maintenance of such stabilities and overfitting (Maimon and Rokach 2005). *Bagging* helps in decreasing the variance in the prediction by generating multiple datasets for training through randomization and repetitive combinations. *Random forest*, a popular model for predictive analysis, is based on the principle of bagging. This ensemble method increases variance in the tree through random subsampling and reduces attributes available at leaf nodes for tree formation to a minimum. On the other hand, *boosting* adjusts the weight of an observation based on the best iterative class. Boosting generates multiple sets of training variants and fixes the classification errors by paying greater attention to the variant classified incorrectly. Boosting has wide applications in popular predictive models such as *gradient boosting models* and *support vector machines*. Both bagging and boosting are weak learners, but when combined, both create a strong learner. The deeper the decision tree, the more complex the model. There are numerous software tools for decision tree analyses and these are categorized based on operational compatibilities, computational speed, abilities in data handling, algorithm structure, and usability (Table 1). Software tools use datasets in .txt and/or .csv file formats for decision tree analysis. The output specifications vary with the tools.

Practical Implications of Decision Tree Analysis

The Eastern Himalayas Example

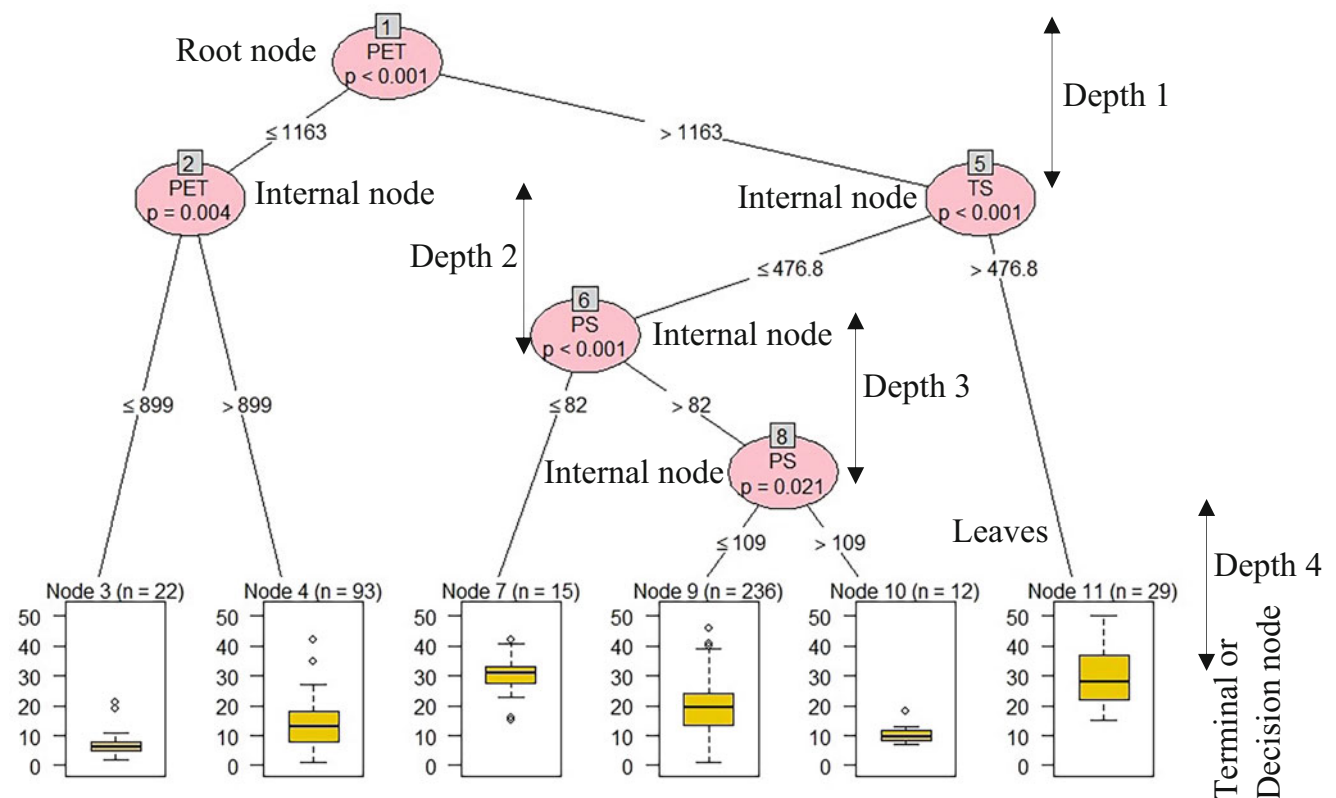
The decision tree analysis was applied to investigate the influences of major climatic predictors on the plant richness (number of species) in the Eastern Himalayas (Panda 2018). The tree described the greater influence of potential evapotranspiration (PET) on plant richness at >1163 ha/yr. Within this limit, an increase in temperature seasonality (>476.8) could further enhance plant richness. A low level of

precipitation seasonality (≤ 82) could further increase plant richness, but with a decrease in temperature seasonality (≤ 6.8). On the other hand, plant richness was found to decrease substantially with the decrease in potential

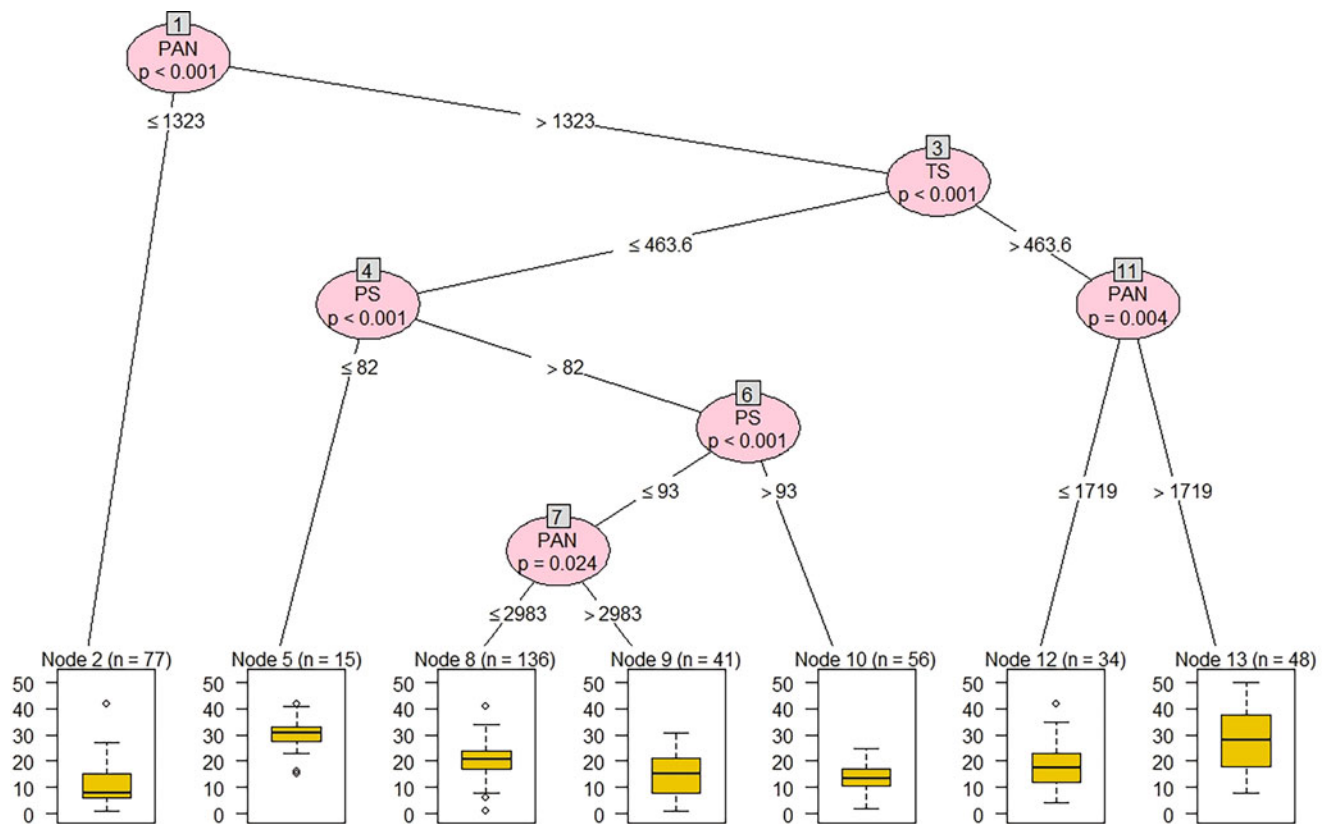
evapotranspiration ≤ 899 ha/yr. The precipitation seasonality was a crucial factor for plant richness at > 109 mm (Fig. 1a). Similarly, mean annual precipitation (PAN, > 1719 mm) led to an increase in plant richness subject to high-temperature

Decision Tree, Table 1 Popular software tools/libraries for decision tree analysis

Software tools	Windows	Linux	Mac	Libraries	Remarks
KNIME	√	√	√	N/A	Powerful and popular with 1500 modules and a wide choice of advanced algorithms
MALLET (Machine Learning Language Toolkit)	√	N/A	N/A	Java	Efficient to convert text to features, with applications in Maximum Entropy, and Naïve Bayes
OC1 (Oblique Classifier 1)	√	√	√	ANSI C	AI and graphical display in astronomy and forensic sciences with options for standard and oblique (multivariate) trees
Orange	√	√	√	Python	Worldwide applications in teaching and training institutes, with support for Tree Learner, Simple Tree Learner, and C45Learner
Rapid Miner	√	√	√	Java	No prior coding knowledge required Easily integrate with Weka
Rattle	√	√	√	R	Applications in statistics, business, teaching, and training
Scikit-learn	√	√	√	Python	Support for ID3, C4.5, C5.0, CART algorithms
SilverDecisions	√	√	√	Java	Software to create decision trees in research and development; image export options for 'json', 'png', and 'svg' formats
Simple Decision Tree	√	√	√	N/A	Excel Add-in
SMILES	√	√	×	C++	Compatible with ensemble models and neural networks
Weka (Waikato Environment for Knowledge Analysis)	√	√	√	Java	Powerful tool with support for ID3, C4.5, Naïve Bayes, k-nearest neighbors
YaDT (Yet another Decision Tree builder)	N/A	N/A	N/A	C++	Support C4.5 algorithm



Decision Tree, Fig. 1a Interactions between precipitation seasonality (PS), potential evapotranspiration (PET), and temperature seasonality (TS) and plant richness; 'n' describes strengths of each node



Decision Tree, Fig. 1b Interactions between precipitation seasonality (PS), annual precipitation (PAN), and temperature seasonality (TS) and plant richness; 'n' describes strengths of each node

seasonality (i.e., >463.6). The temperature seasonality of ≤ 463.6 and precipitation seasonality of ≤ 82 accelerated plant richness of the Eastern Himalayas. Alternately, a heavy mean annual precipitation (i.e., >2983 mm) could decrease plant richness by water-logging and probably due to seed rotting. The precipitation seasonality of >93 could reduce the plant richness further, even at a low level of temperature seasonality, i.e., <463.6 (Fig. 1b). This suggests that seasonal fluctuations in precipitation and temperature have significant control over the plant richness of the Eastern Himalayas.

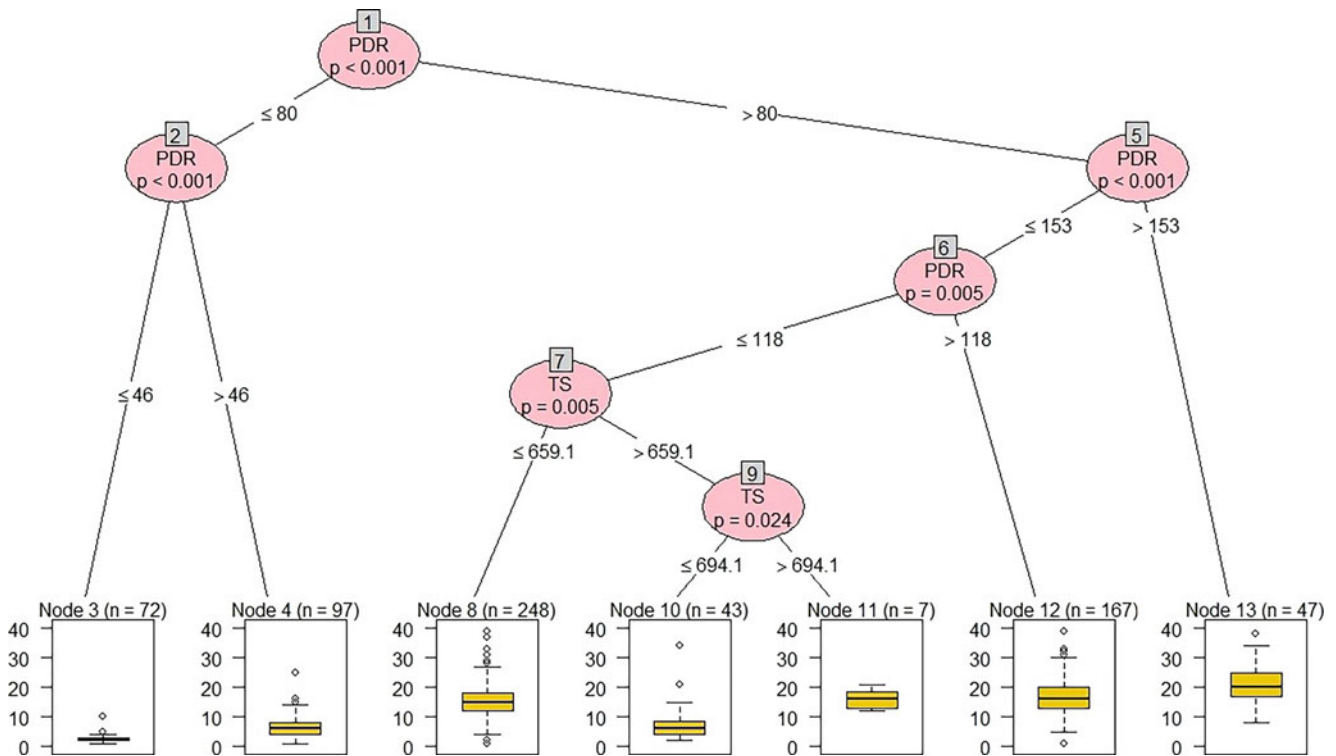
The Western Himalayan Example

A decision tree analysis in the Western Himalayas revealed that precipitation of the driest quarter (PDR) < 80 mm has great significance for plant richness, which could minimize its impact with mean annual precipitation >352 mm. This indicates that the plant richness in the Western Himalayas is dependent on water availability, where greater precipitation in the driest quarter is a limiting factor (Panda 2018). However, the plant's survival continues even at mean annual precipitation <352 mm and PDR ≤ 46 mm. The temperature seasonality (TS) is a crucial factor, and >659 would create thermal stress on the plant richness. Alternately, a TS value of

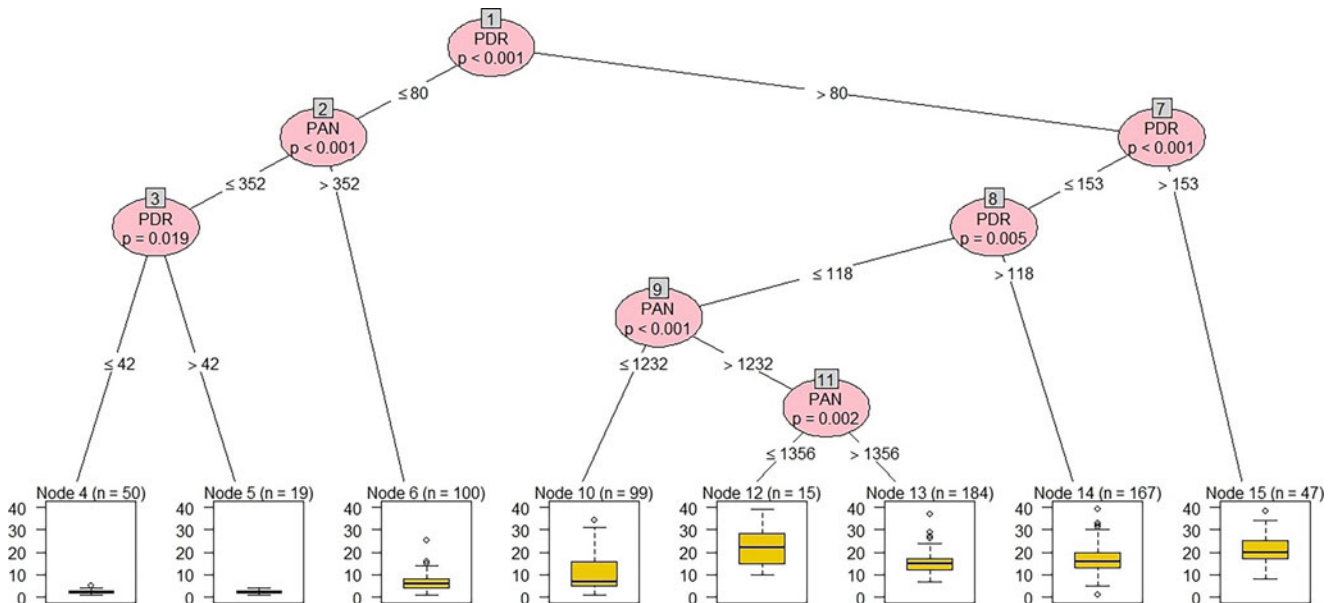
≤ 659 could accelerate the plant richness. This suggests that the fluctuations in seasonal temperatures are crucial in determining the plant richness of the region (Fig. 2a).

A similar decision tree analysis in the Western Himalayas indicating the interactions between variables and plant richness showed a significant correlation between precipitation of the driest quarter (PDR) and plant richness. In general, an increase in precipitation during the driest period (>153 mm) accelerated plant richness. A PDR value ranging between 118 mm and 153 mm was predicted as being crucial for determining plant richness. Precipitation of the driest quarter (≤ 118 mm) influenced plant richness more significantly with a reduced rate of temperature seasonality (i.e., $\leq 659\%$). And for obvious reasons, a high-temperature seasonality strongly regulated plant richness at a low level of precipitation in the driest quarter (Fig. 2a).

The precipitation of the driest quarter between 80 and 118 mm could favor the plant richness in the Western Himalayas subject to the mean annual precipitation of ≤ 1232 mm. While greater precipitation in the driest quarter (118–153 mm) was found to have a strong positive effect on plant richness, its value >153 mm was found to be less limiting for plant richness. A moderate PDR and good mean annual precipitation favored plant richness in the Western



Decision Tree, Fig. 2a Interactions between precipitation of driest quarter (PDR) and temperature seasonality (TS) and plant richness; 'n' describes strengths of each node



Decision Tree, Fig. 2b Interactions between precipitation of driest quarter (PDR) and mean annual precipitation (PAN) and plant richness; 'n' describes strengths of each node

Himalayas. This substantiates the limiting role of water availability in plant richness in dry conditions of the Western Himalayas (Fig. 2b).

Conclusions

A decision tree is a tool for decision making and management in many data mining procedures. This is a machine learning

method that involves both regression and classification principles. The decision tree has many advantages over standard regression or classification owing to its ability to use both categorical and numerical variables and its effectiveness in nonlinear modeling setups. The hierarchical tree structure of a decision tree has a root at its top and leaves or decision nodes at its bottom. The deeper the decision tree, the more complex the model. A decision tree can be nonbinary (e.g., ID3, C4.5) or binary (e.g., CART) and depends on different algorithms for splitting or subsampling. ID3 and C4.5 are based on entropy principles and CART on measures of impurity. Of all of them, CART is the simplest and a very popular algorithm, used in decision tree analysis of recent research. Despite many advantages, the decision tree has problems of instabilities, which can be easily handled by bagging and boosting. Independently, bagging and boosting are weak learners, but when combined, both become strong learners. Some of the good examples of decision tree algorithms involving principles of bagging and boosting for bias-free predictions are random forest, gradient boosting model, and support vector machines, widely implemented in ecological, engineering, scientific, and socioeconomic decision making and planning.

Cross-References

- [Automatic and Programmed Gain Control](#)
- [Cartogram](#)
- [Entropy](#)

References

- Bel L, Allard D, Laurent JM, Cheddadi R, Bar-Hen A (2009) CART algorithm for spatial data: application to environmental and ecological data. *Comput Stat Data Anal* 53:3082–3093
- Breiman L, Friedman JH, Olshen R, Stone CJ (1984) Classification and regression trees. Wadsworth, Belmont
- Breiman L, Friedman JH, Olshen RA, Stone CJ (2017) Classification and regression trees. Routledge
- Debeljak M, Džeroski S (2011) Decision trees in ecological modeling. In: *Modelling complex ecological dynamics*. Springer, Berlin, Heidelberg, pp 197–209
- Krzywinski M, Altman N (2017) Classification and regression trees. *Nat Methods* 14:757–758
- Lewis DD, Catlett J (1994) Heterogeneous uncertainty sampling for supervised learning. In: *Machine learning proceedings*. Morgan Kaufmann, pp 148–156
- Maimon O, Rokach L (eds) (2005) *Data mining and knowledge discovery handbook*. Springer
- Panda RM (2018) Environmental determinants of plant richness in Indian Himalaya. The Ph.D. thesis, submitted to Indian Institute of Technology Kharagpur, pp. 1–218
- Quinlan JR (1986) Induction of decision trees. *Mach Learn* 1:81–106
- Rutkowski L, Jaworski M, Pietruczuk L, Duda P (2014) The CART decision tree for mining data streams. *Inf Sci* 266:1–15

Deep CNN-Based AI for Geosciences

Mehala Balamurali

Australian Centre for Field Robotics, University of Sydney, Sydney, NSW, Australia

Definition

In deep learning, a convolutional neural network (CNN, or ConvNet) is a subset of artificial neural network that is specially used for processing data that has grid-like topology. As the inclusion of “convolutional” in name, these neural networks apply convolution in lieu of standard matrix multiplication in at least one of their layers (Ian Goodfellow 2016).

Introduction

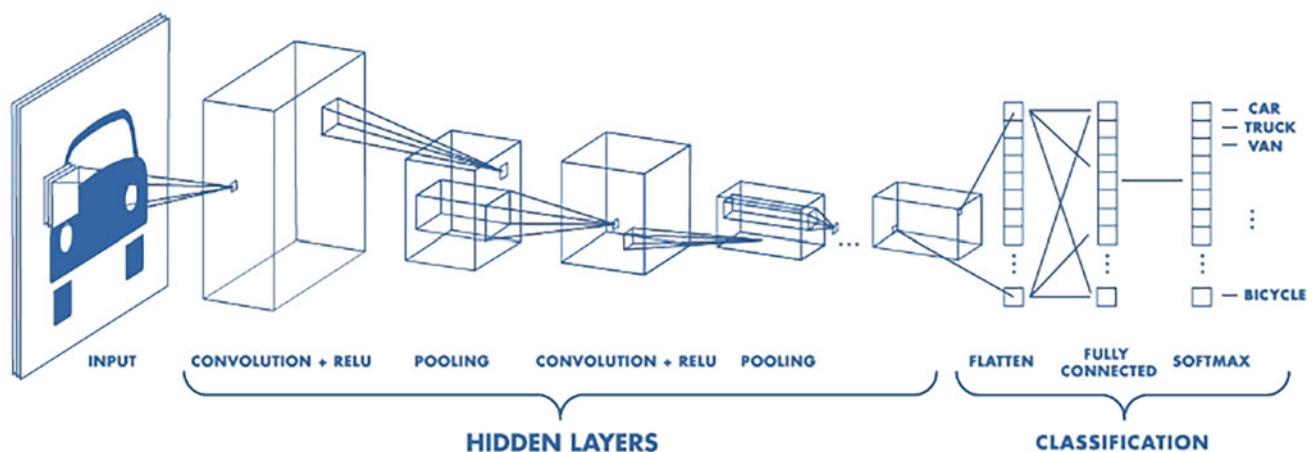
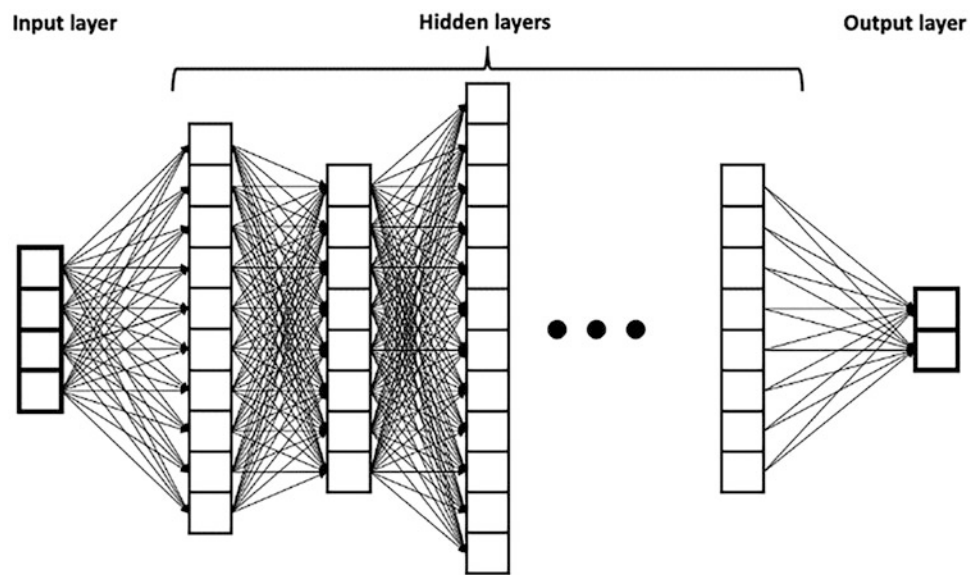
Deep learning along with machine learning are subsets of artificial intelligence. Deep learning algorithms are more state-of-art and unsupervised comparing to machine learning approaches. A deep neural network (DNN) is an artificial neural network (ANN) with multiple layers between the input and output layers (Alzubaidi et al. 2021). Deep learning taught the machine with inputs, and target should be detected and classified after several hidden layer in the output layer (Fig. 1). The deep neural network imitates the neurons, and their connections in our brain on how the information is being received and processed.

There are many deep neural network (DNN) algorithms. Among them **deep convolutional neural networks** are widely used in identifying patterns in images and videos. The gaining popularity in computer vision with deep Learning has been raised and improved with time, mainly the implementation of the convolutional neural network.

CNN can effectively capture the spatial and temporal dependencies of data that has grid-like topology using multiple filters. Unlike other methods where features are hand engineered, in CNN, the features are learned with adequate training.

In recent years, application of convolutional neural networks (CNN) is gaining popularity in the field of Geosciences. CNN would map the input data and make prediction after processing in the layers between. The common

Deep CNN-Based AI for Geosciences, Fig. 1 A deep learning architecture with multiple hidden layers. (Source: Wikimedia c CC)



Deep CNN-Based AI for Geosciences, Fig. 2 A classic CNN structure. (Source: <https://goo.gl/zondfq>)

types of layers can be seen in a classic CNN structure are convolutional layer, pooling layer, and fully connected layer (Fig. 2).

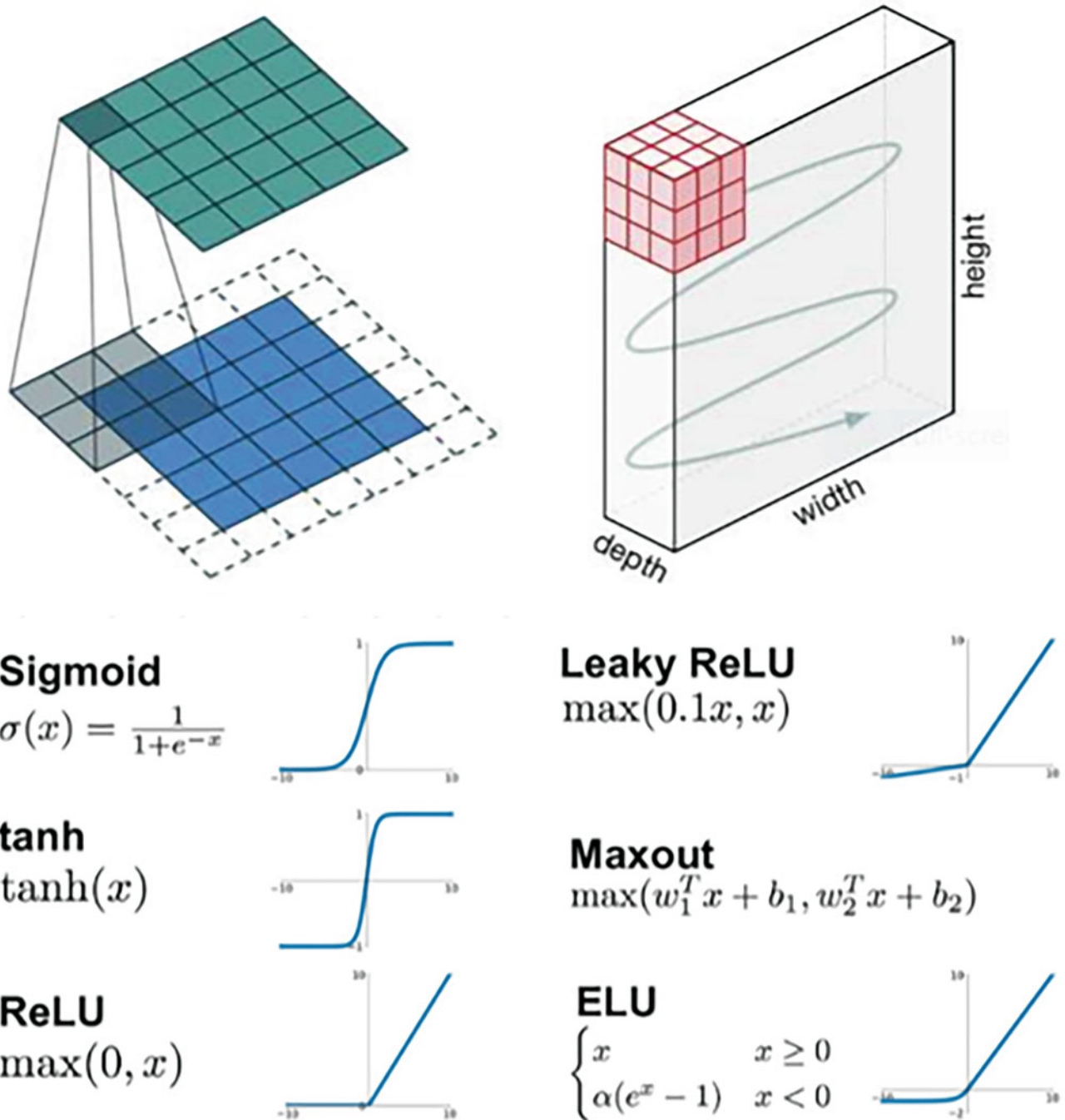
Convolutional Layer

As name suggested, the convolutional layer (Fig. 3) is an important part of the structure of CNN. The process of convolution in this layer extracts the features from the input data. In the image data, the filters known as kernels convolute the weights and heights from the input pixel values and returns the important features in a similar dimension of image or in different dimension.

When the input is RGB images, they are read as pixels as a height x width x depth ($M \times N \times 3$) matrix. The dot product of the corresponding image matrix and a small dimension of

filter matrix with similar depth will return the feature map. The process of sliding filters through the width and the height of the image along the operation of dot multiplication is known as convolution. Different filters retrieve different features, and the outputs are passed to the next layer. With deeper layers, more sophisticated details can be identified.

The parameters of convolution operations are filter size, padding, stride, dilation, and activation function. The left image in the top row of Fig. 3 shows a 3 by 3 filter used with a 5 by 5 image with padding of one that results 5 by 5 image. Padding is used to maintain both output and input image sizes same by padding zeros onto input image in each dimension before the convolution operation. Stride determines how many pixels the filter moves by, in right or down direction. The activation functions used in the hidden layers

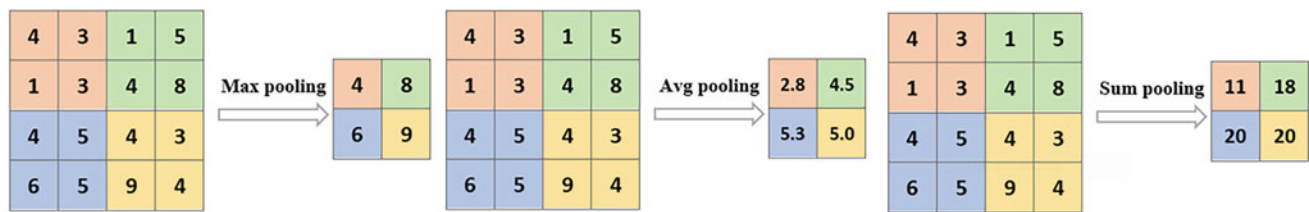


Deep CNN-Based AI for Geosciences, Fig. 3 Convolutional layer and kernel movement example [<https://theano-pymc.readthedocs.io>] at row 1 and Examples of activation functions [Shruti Jadon on [Medium.com](https://medium.com)] at row 2

control the learnability of the model from the training data. Hence the selection of activation functions state model prediction. The bottom row of Fig. 3 shows some example of activation functions that are commonly used in computer vision: sigmoid, hyperbolic tangent, rectified linear unit (ReLU), leaky ReLU, Maxout, and exponential linear units (ELU).

Pooling Layer

The pooling layer performs the down sampling using different type of pooling layers and thus combination of convolutional and pooling layers reduce the number of parameters while maintains the feature information. As seen in Fig. 4, max pooling finds the largest value within the selected matrix, the average pooling calculates the average



Deep CNN-Based AI for Geosciences, Fig. 4 Maximum, average, and sum pooling layers

of the values in the selected region, and the sum pooling layer returns the sum of the values within the selected region.

Fully Connected Layers

In CNN architecture, fully connected layers construct the last few layers where all neurons from previous layer are associated to each activation unit of the next layer. Hence these layers are the next time-consuming layers after convolution layers. The output from the last pooling layer or convolution layer is input into the fully connected layer in the flatten format.

Training a Deep CNN Model

CNN models commonly use backpropagation algorithm where the operation finds the kernels in convolution layers and weights of fully connected layers that minimize the difference between the output prediction and given ground truth labels on a training dataset. The learnable parameters are calculated using loss functions and updated using optimization algorithm. The given data with ground truth labels are split into three sets: training data, validation data, and test data. While the model is trained using training data, validation data is used to observe the model performance during training and fine tune the hyper parameters. Test data are further used to evaluate the model performance on unseen data after the model is trained and finetuned.

Transfer Learning

Training a deep learning model needs enormous, labelled data, but those are rarely available in some domains. Transfer learning is one of the alternative approaches that adopts a CNN model which is previously trained on an abundance of natural labelled images. This method has been gaining popularity in many research fields including geosciences, and the exchange of knowledge transfer between natural images and images from geosciences has been continuously studied.

Transfer learning can be established in two different ways. As in Fig. 5b, a pretrained model can be used as a feature extractor by replacing the fully connected layer by any classifiers such as SVM and random forests. On the other hand, as in Fig. 5c, the pre-trained model can be fine-tuned with

labelled data from geosciences by replacing not only the fully connected layer also with some part of the convolutional layers.

Some Commonly Used CNN Architecture

Some of the commonly used deep convolutional neural network architectures in computer vision tasks (Alzubaidi et al. 2021) are below.

AlexNet (Year)

AlexNet consists of eight layers including five convolutional and three fully connected layers. ReLU nonlinearity is implemented after all the layers and followed by overlapping pooling and then local normalization process. An overfitting problem may be occurred by the large number of parameters because AlexNet has 60 million parameters. In order to reduce this problem, data augmentation, principal component analysis (PCA), and dropout methods are implemented.

VGGNet

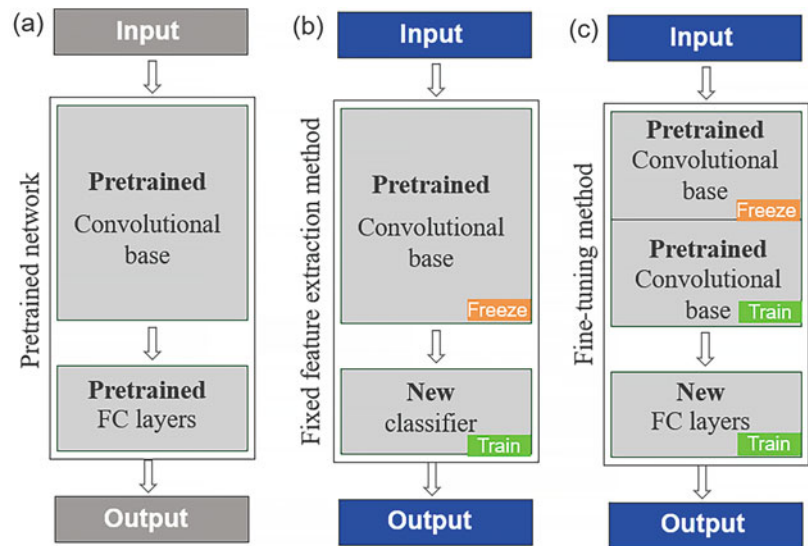
As the winner of the 2013 ILSVRC competition, VGGNet consisted of 16 to 19 layers, which is significantly deeper compared with the architecture of AlexNet for instance. Small three by three filters are implemented in VGGNet, which sets a trend of network with small filter size to aim to increase depth.

Inception Net/GoogLeNet

From the winning of ILSVRC price in 2014, Inception Net has been developed and has up to version 4 now. Inception v1 (GoogLeNet), which was developed by a team in Google, introduces high-accuracy inception blocks with low complexity. Inception blocks implement different size filters in the same convolutional block to capture various spatial features in different operations. This is a classic feature that continues to the following versions of Inception Net. Inception v2 (BN-Inception) is version 1 with addition batch normalization which copes with internal covariant shift situation. Inception v3 is updated with factorization. Finally, Inception v4 is a pure inception version with no residual connections that could be trained with memory optimization and no need for partitioning.

Deep CNN-Based AI for Geosciences, Fig. 5 (a)

Pretrained model is used as feature extractor (b) or (c) fine-tuning technique



ResNet (2015)

ResNet is considered to implement residual learning to avoid the decrease of performance due to the increasing number of stacking layers. In ResNet, skip connections were implemented between shallow and deep layers to prevent the gradient vanishing and degradation of the model. Additionally, skip connections could skip the redundant layers and maintain the ability of computer vision while ensuring the performance and non-degradation of the model after saturation.

Inception ResNet

Combined with Inception Net and ResNet explained above, the first version, Inception-Resnet-v1 was implemented which is faster but lower accuracy than Inception v3. Adding ReLU as pre-activation unit will make the model go deeper. However, an exceeded number of filters might raise model instability and network death. In the following version, Inception ResNetv2 has much faster training speed and accuracy and similar computational cost as Inception v4.

MobileNet

MobileNet is normally used in mobile vision applications on image classification when the environment has limited resources. MobileNet implemented a small number of parameters with high efficiency. MobileNet is also a good option for portable devices as the size of the model is shrunk.

Applications in Geosciences

In recent years deep CNN models have provided more accurate results and are getting popularity in the field of geosciences. Unlike traditional machine learning approaches that are based on a structured data with carefully selected

attributes, deep CNN can process raw unstructured data themselves within layers of the network. However, deep CNN models rely on large, labeled data. Hence, the first step in the model approach, domain experts' knowledge is needed to robustly train these models with accurate ground truth data.

Classification

Deep CNN applications in geosciences have already been seen in many applications in earth and planetary sciences. Among them, deep CNN models have been used as image classification tools for rock images classification and reservoir classification (Evgeny et al. 2020) and interpreting lithology and subsurface from drill cores (De Lima et al. 2019). As the seismic data is usually logged in the time-space domain, transforming this data into gray scale images let the researchers to utilize the advantage of CNN models to studying stratigraphic natures.

Segmentation

Singh et al. (2020) used state of the art deep CNN models (DeepLab, UNet, DenseNet and SegNet) to segment river surface into water and two types of ice with limited labeled and noisy data. Some other studies used classic image segmentation CNN models to segment faults (Yu and Ma 2021).

Detection

In earth and planetary sciences, geologists found detecting and counting craters are key feature to determine the evolution of planets (DeLatte et al. 2019). Scratch CNN models were used to extract the features from image data and then used them to train the classifiers. On the other hand, models such as GoogLeNet, VGGNet, and ResNet that were pre trained on natural images, have been used to automate the crater detection. Faster R-CNN model was used to localize the craters in the planetary images.

Furthermore, pretrained deep CNN models have been used to perceive the number of mineral grains in sediments or sands using the microscope scans (Julien et al. 2019). Detection of geological features such as faults, layers, dips, boundaries, and cracks have been treated similar to object detection in other computer vision application and were automatically detected using deep CNN models.

Other

Digital elevation maps (DEM) are another vital tool for geologists. Deep CNN models have used to map low- and high-resolution digital elevation map (Jiao et al. 2020). Deep CNN models were used to separate the signals and noise in geophysical data (Zhang et al. 2017). While in many geoscience applications, CNN have used for processing 2D and 3D image data, deep CNN models have been widely adopted by research community to test its capability in interpreting frequency and time domain geophysical data in single dimension (Yu and Ma 2021).

Summary

Deep convolutional neural networks (DCNNs) have accomplished remarkable achievement across diverse domains and have been gaining popularity in geoscience applications. While these models have become leading methods in solving many complex tasks such as image classification and object detection, getting to know the key concepts, advantages, and limitations of these models is crucial, in order to fully utilize them in geoscience research such that the models can provide experts with insights that assist them in decision making.

Cross-References

- [Artificial Intelligence in the Earth Sciences](#)
- [Cluster Analysis and Classification](#)
- [Machine Learning](#)
- [Mathematical Geosciences](#)
- [Neural Networks](#)
- [Object-Based Image Analysis](#)
- [Pattern Classification](#)

Bibliography

- Alzubaidi L, Zhang J, Humaidi AJ et al (2021) Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* 8:53. <https://doi.org/10.1186/s40537-021-00444-8>
- DeLatte DM, Crites ST, Guttenberg N, Yairi T (2019) Automated crater detection algorithms from a machine learning perspective in the convolutional neural network era. *Adv Space Res* 64(8): 1615–1628. ISSN 0273–1177. <https://doi.org/10.1016/j.asr.2019.07.017>

- De Lima RP, Bonar A, Coronado DD, Marfurt K, Nicholson C (2019) Deep convolutional neural networks as a geological imageclassification tool. *Sediment Rec* 17, 4–9. (PDF) Application of Deep Learning in Petrographic Coal Images Segmentation. Available from: https://www.researchgate.net/publication/356211009_Application_of_Deep_Learning_in_Petrographic_Coal_Images_Segmentation. Accessed 22 Dec 2021
- Evgeny E. Baraboshkin, Leyla S. Ismailova, Denis M. Orlov, Elena A. Zhukovskaya, Georgy A. Kalmykov, Oleg V. Khotylev, Evgeny Yu. Baraboshkin, Dmitry A. Koroteev, Deep convolutions for in-depth automated rock typing *Comput Geosci* 2020 135 104330., ISSN 0098-3004, <https://doi.org/10.1016/j.cageo.2019.104330>
- Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, p 326
- Jiao D, Wang D, Lv H, Peng Y (2020) Super-resolution reconstruction of a digital elevation model based on a deep residual network. *Open Geosci* 12(1):1369–1382. <https://doi.org/10.1515/geo-2020-0207>
- Julien M, Kevin BL, Paul B (2019) Mineral grains recognition using computer vision and machine learning. *Comput. Geosci* 130:84–93.
- Singh H, Kalke ML, Ray N (2020) River ice segmentation with deep learning. *IEEE Trans Geosci Remote Sensing* 58(11):7570–7579. <https://doi.org/10.1109/TGRS.2020.2981082>
- Yu S, Ma J (2021) Deep learning for geophysics: current and future trends. *Rev Geophys* 59:e2021RG000742. <https://doi.org/10.1029/2021RG000742>
- Zhang K, Zuo W, Chen Y, Meng D, Zhang L (2017) Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Trans Image Process.* 26(7):3142–3155. <https://doi.org/10.1109/TIP.2017.2662206>. Epub 2017 Feb 1. PMID: 28166495.

Deep Learning in Geoscience

Fuchang Gao
Department of Mathematics and Statistical Science,
University of Idaho, Moscow, ID, USA

Synonyms

Deep Neural Networks

Definition

Deep learning is a group of machine learning algorithms that has been widely used in geoscience for learning features in high-dimensional data. Selecting a right model and tailoring it to fit for the unique spatial-spectral-temporal nature of the data is key to the successfulness of the application.

Introduction

Deep learning is a group of machine learning algorithms that uses a multilayer neural network architecture to learn a high-dimensional function f from its values at a set of sample

points. By training on a set of sample data X , the goal is to find a sequence of operators $T_1, T_2, \dots, T_h$ such that $T_h \circ T_{h-1} \circ \dots \circ T_1(\tilde{X})$ best approximates $f(\tilde{X})$ for unseen data $\tilde{X}$ sampled from the same distribution as X . The number h is called the depth of the neural network. It is assumed that each operator T_i is of the form

$$X \mapsto (\sigma \circ L_1(X), \sigma \circ L_2(X) \dots, \sigma \circ L_q(X)),$$

where σ is a nonlinear function, and $L_1, L_2, \dots, L_q$ are linear (affine) operators defined as

$$L_i(X) = w_{i1}x_1 + w_{i2}x_2 + \dots + w_{ip}x_p + b_i, \\ X = (x_1, x_2, \dots, x_p).$$

During neural network training, the parameters $w_{i1}, w_{i2}, \dots, w_{ip}$ and b_i are adjusted to minimize the overall approximation error. Approximation theory in mathematics ensures that any high-dimensional continuous function f in a bounded domain can be approximated this way with $h \geq 2$. However, it has been known that in many cases, larger h leads to better performance. When h is large (e.g., $h \geq 7$), the learning algorithm is called deep learning.

Architecture, or more specifically, where each linear operator is defined, determines the performance of a neural network on specific data types. Among the most popular deep learning models are convolutional neural networks, recurrent neural networks, autoencoders, generative adversarial networks, and graph neural networks, each having many successful applications in geoscience.

Convolutional Neural Networks (CNNs)

CNNs are specific kinds of neural networks designed for better capturing spatial dependence among neighboring variables. For digital images, if we use double indices to label the variable (pixel value) at location (i, j) as x_{ij} , then a linear operator in a 2D CNN is defined as $\sum_{(i,j) \in W} c_{ij}x_{ij} + b$, where

W is a $h \times w$ rectangular window centered at (i, j) . Such a linear operator defines a specific mechanism to extract a local feature around (i, j) . CNNs assume such a mechanism is shared by different locations. Thus, each specific local feature only requires $h \times w + 1$ parameters. The small size of $h \times w$ allows a CNN to pack a sequence of such linear operators in one layer. With many layers, a deep CNN can potentially extract many kinds of features in the data, often enabling it to outperform models based on a few hand-crafted features.

The 2D CNNs can be naturally extended to handle multi-dimensional data cubes. For example, hyperspectral imagery (HSI) data are typically presented in 3D cubes, with two spatial dimensions and one spectral dimension of a typical size of over 200. However, the number of variables in a 3D data cube can be typically over a million. Thus, a naïve use of 3D CNNs may face a computational challenge, which is further coupled by the typically smaller number of labeled data points. Below are two representative approaches to use CNNs on HSI.

The first is the multiscale 3D deep CNN model in Qi and Zhang (2020), which leverages two branches of a 3D CNN to extract spatial and spectral features separately. The first branch consists of six 3D convolutional layers interlaced by batch normalizations and ReLu activations, and skip connections between the first, the third, and the fifth layers. It transforms a $9 \times 9 \times 200$ block into 24 spectral features of size $7 \times 7 \times 1$, to be used in the second branch. The second branch crops out two center patches of size $7 \times 7 \times 200$ and $5 \times 5 \times 200$, respectively. The $7 \times 7 \times 200$ patch is transformed by a five-layer 3D CNN with skip connections to 24 features of size $5 \times 5 \times 1$ before being merged with the saved patch of size $5 \times 5 \times 200$ to form 224 features of size $5 \times 5 \times 1$. Additional five convolutional layers with skip layers and fully connected layers are used to finish the classification.

The outstanding performance of the model is derived not only from the task segregation of the two branches, but also from the depth and the skip connections in each branch. Indeed, good performance of CNNs often requires significant depth, which was unfortunately missing in some earlier CNN models for the same HSI data.

A deep CNN requires a powerful GPU and a long time of training. One approach to avoid extensive training is to integrate some existing knowledge into the models. Recent work (He et al. 2019) exemplifies such an approach.

In that article, the authors trained a 2D CNN on hand-crafted feature extractions from the original HSI. The method is as follows: First, apply the maximum noise fraction (MNF) transformation, a transformation that is like principal component analysis (PCA), to reduce the spectral dimension to 20. Next, for each pixel, choose a sequence of M square neighborhoods of varying sizes, centered at the pixel. For a $T \times T$ square neighborhood centered at x_i , find the covariance matrix of the T^2 spectral vectors of dimension L defined by

$$C_{L \times L} = \frac{1}{T^2 - 1} \sum_{j=1}^{T^2} (x_j - \mu)(x_j - \mu)^T$$

where x_j is the spectral vector at pixel x_j in the square neighborhood of x_i , and μ is the mean of the x_j 's. These multiscale covariance maps (MCMs) are of the same shape $T \times T$,

allowing a standard 2D CNN to complete the classification task. The model was reported to have outperformed some other CNN models, as well as some other machine learning models for the same tasks.

While it is worth discussing if the use of these specific MCMs is the best way for capturing spatial-spectral features of HSI, the novelty of this approach is that it introduces a creative way to jointly integrate spatial and spectral information. Note that in this approach, each covariance matrix is the simple average of $g(\mathbf{x}_j - \boldsymbol{\mu})$ over the neighborhood for a specific function g . Thus, the spatial information captured by these features is limited. Nonetheless, this approach opened a door for many meaningful modifications and improvements. For example, one may design different function g or train averaging weights for better integrating spatial information.

Note that the feature maps created in the paper were symmetric matrices. Therefore, half of the information was redundant. Also note the high correlation between neighboring spectral bands is lost in the MNF-transformed spectral vectors. Thus, there seems to be no good reason to use a 2D CNN on these particular features. However, if this approach is applied to spatial-temporal data without a prior nonmonotonic transformation on the temporal axis, CNNs are likely to be beneficial.

Recurrent Neural Networks (RNNs)

In the neural networks we discussed above, the goal was to learn the unknown function f from its behavior at a sequence of data points $X_1, X_2, \dots, X_N$, which were assumed to be independently sampled from the same sample space. In applications, however, we often encounter cases when the sequence of data points $X_1, X_2, \dots, X_N$ are strongly dependent, and more importantly, the function f depends not only on the current data X_N , but also on the previous data $X_1, X_2, \dots, X_{N-1}$. Indeed, this is a typical situation for temporal data, such as daily air pollution, temperature, precipitation, solar radiation, groundwater hydrographs, or co-seismic landslides. Similar dependencies also appear in spatial-spectral data. RNNs are specifically developed to deal with such dependence between sequential data. In a standard neural network, for each input X_N , the neural network uses the operator L to output $\sigma \circ L(X_N)$ without using any information about $\sigma \circ L(X_{N-1})$. In an RNN, the previous output is kept as history h_{N-1} and used together with the current input X_N to produce the new output. Because h_{N-1} was also produced using both X_{N-1} and h_{N-2} , and so on, an RNN has the potential to capture the dependence of the target function on $X_1, X_2, \dots, X_N$.

Careful control of the information flow from history h_{N-1} is of key importance. Without it, an RNN may suffer from the vanishing gradient problem. Long short-term memory

(LSTM) is a specific architecture that facilitates this need. It has a cell unit, an input gate, an output gate, and a forget gate to regulate the information flow. There are many variations of LSTM, such as bidirectional LSTM, convolutional LSTM, graph LSTM, tree-structured LSTM, and gated LSTM. Another very popular RNN is the gated recurrent unit (GRU), which also has a forget gate but not an output gate. GRUs have fewer parameters than LSTMs, making them perform better on smaller datasets, while LSTM may fit large datasets better. However, LSTM may suffer from overfitting. The following two recent articles exemplify the current uses of LSTM and GRU in geoscience research.

In Xiao et al. (2019), the researchers combined adaptive boosting (AdaBoost) ensemble learning with LSTM to overcome the overfitting challenge of LSTM for predicting sea surface temperature using daily satellite time series data at selected sites. Before feeding to the LSTM and AdaBoost, the authors used polynomial fitting to model the seasonal pattern and use it to deseasonalize the raw data. An LSTM is used on the normalized deseasonalized data. Then, adaptive boosting is used on a sequence of weak decision tree regressors produced by a classification and regression tree (CART) to get a new prediction. The two predictions were then averaged to get the final prediction. The combined model significantly outperformed each individual model, as well as other earlier models.

In Hang et al. (2019), the researchers applied RNNs to HSI classification, treating the spectral axis as the time axis. The authors noted that adjacent spectral bands contain a lot of redundant information, while remote spectral bands contain complementary information. Based on this property, the authors built a model with two interconnected GRUs, with the first one removing redundant information within neighboring spectral bands, and the second one learning complementary information from nonadjacent spectral bands. They also used CNN to capture spatial features, as follows. Starting with the HSI data in a 3D cube of size $m \times n \times k$, a small center cube of size $w \times w \times k$ is taken for each pixel and divided into k matrices of size $w \times w$. A 2D CNN is then built to convert each matrix into a d -dimensional feature vector. Then the sequence of k d -dimensional vectors is split into l subsequences $S_1, S_2, \dots, S_l$. Next, each subsequence S_i is fed to the first GRU, outputting an h -dimensional vector F_i^1 . Then, the sequence of l vectors $F_1^1, F_2^1, \dots, F_l^1$ is fed to the second GRU, outputting a vector F^2 . Finally, $F_1^1, F_2^1, \dots, F_l^1$ and F^2 are fed into a final linear layer to generate $l+1$ output vectors with corresponding loss functions $L_1^1, L_2^1, \dots, L_l^1, L^2$. A weighted average of these outputs with weights $w_1^1, w_2^1, \dots, w_l^1, w^2$ is used for final classification, and the same weights are used to define the

loss function
$$L = \frac{1}{l} \sum_{i=1}^l w_i^1 L_i^1 + w^2 L^2$$
 to regulate the

contributions of different outputs. The skip connectivity technique that feeds a copy of the outputs of the first GRU to the last linear layer is a useful technique also used in the 3D CNN introduced earlier. Using weighted loss to softly regulate contributions is also a popular technique.

Autoencoders (AEs)

An autoencoder (AE) is a special kind of neural network in which the output aims to recover the input. If in the middle there is a bottleneck layer with a smaller number of neurons than the input, then the output of this layer can serve a lower dimensional representation of the input data. Thus, AEs usually have a much small bottleneck layer in the middle, and they are often used for dimensionality reduction. If no activation function is used in the network, then such a lower dimensional representation converges to that from PCA. In this sense, AEs have more capacity. Besides generating a lower dimensional representation, a trained AE can also be used for noise reduction and missing value imputation, as the output of an AE preserves essential features of the input. Low-dimensional representations obtained from AEs often outperform those obtained by PCA, as measured by subsequent machine learning algorithms such as clustering. For instance, Mousavi et al. (2019) demonstrated that an unsupervised clustering based on AE can reach the precision of supervised methods on predicting seismic signals without using the labeled data.

Another desired use of AEs is to generate synthetic data by slightly altering the encoded vector in the bottleneck layer. However, a standard AE is not capable of serving that purpose: By randomly altering the encoded vector, the neural network output is often not meaningful. Variational autoencoders (VAEs) have been developed to regulate the distribution of the encoded vectors of meaningful inputs in the latent space. For each input, a VAE produces two vectors μ and σ in the latent space, then randomly selects an independent normal vector (independent among the coordinates) with mean μ and standard deviation σ , and passes them to decoder (the rest part of the neural network). A penalty function is added to encourage all the vectors μ to stay as close to the origin as possible, while keeping the distance between μ_i and μ_j as large as possible for different classes. With this added feature, VAEs can better generate meaningful data like the input, which makes them extremely useful in synthesizing underrepresented data for training machine learning models, as well as in data assimilation (e.g., Canchumuni et al. 2019).

As generative model, VAEs should be compared with generative adversarial networks (GANs), which have also been popularly used in geoscience. For example, GANs have been used to generate synthetic data for 3D seismic facies classification.

Graph Neural Networks (GNNs)

The CNN and RNN models discussed above are very successful in analyzing spatial-spectral-temporal data when there is an underlying Euclidean structure in spatial-spectral-temporal space, meaning that adjacent variables are highly correlated, while remote variables are independent. There are, however, data that are generated from non-Euclidean domains. For example, in the climate data, remote regions may be highly correlated in climate variables such as sea level, temperature, or precipitation. These teleconnections can be represented as a graph with complex adjacency. Indeed, even in Euclidean space, discrete data points may also be expressed using a tree-like structure. GNNs were first studied in late 1990s, and recently they have been actively investigated. In particular, many variants of convolutional graph neural networks (CGNNs), graph variational autoencoders (GAEs), and recurrent graph neural networks (RGNNs) have been developed. In a convolutional layer in GNNs, each node aggregates feature information from its adjacent nodes. After feature aggregation, an activation is applied to the resulting outputs. By stacking multiple layers, each node receives information from all the nodes in the entire graph. To downsize the network, a pooling layer is used to generate a subgraph. At the end of the convolutional layers, a readout layer sums or averages the features in the final subgraph to generate a vector, which is then connected to the fully connected final output layer (Defferrard et al. 2016). GAEs encode a graph and its node information into a vector in latent space and recover the graph data from the encoded vector through reconstructing the graph adjacency matrix (Kipf and Welling 2016). The construction of RGNNs is more involved. For example, because graphs may have different numbers of nodes and different edge connections, to recurrently update the hidden state for each node, a sum operator is needed, and to ensure convergence, certain restrictions must be put on recurrent functions. Detailed descriptions of the construction of RGNN can be found in Scarselli et al. (2009). Because CGNNs, RGNNs, and GAEs are tailored for data in non-Euclidean domains, they are of great potential in geoscience, as exemplified by recent applications (e.g., Khodayar et al. 2020).

Summary

Deep learning models need to be selected based on data. CNNs were designed for better capturing spatial dependence among neighboring variables. RNNs better fit for temporal data in which previous processed data help the current prediction, and they may also be used for spectral data. AEs are often used for dimensionality reduction. VAEs and GANs are extremely useful in synthesizing underrepresented data.

GNNs have great potential in geoscience for studying data in which the relation among the variables has a network structure.

Acknowledgment This book chapter is based upon work supported by the National Science Foundation under Grant No. 2019609.

Bibliography

- Canchumuni SWA, Emerick AA, Pacheco MAC (2019) Towards a robust parameterization for conditioning facies models using deep variational autoencoders and ensemble smoother. *Comput Geosci* 128:87–102. <https://doi.org/10.1016/j.cageo.2019.04.006>
- Defferrard M, Bresson X, Vandergheynst P (2016) Convolutional neural networks on graphs with fast localized spectral filtering. *Proc NIPS*:3844–3852
- Hang R, Liu Q, Hong D, Ghamisi P (2019) Cascaded recurrent neural networks for hyperspectral image classification. *IEEE Trans Geosci Remote Sens* 57(8):5384–5394. <https://doi.org/10.1109/TGRS.2019.2899129>
- He N, Paoletti ME, Haut JM, Fang L, Li S, Plaza A, Plaza J (2019) Feature extraction with multiscale covariance maps for hyperspectral image classification. *IEEE Trans Geosci Remote Sens* 57(2):755–769. <https://doi.org/10.1109/TGRS.2018.2860464>
- Khodayar M, Mohammadi S, Khodayar ME, Wang J, Liu G (2020) Convolutional graph autoencoder: a generative deep neural network for probabilistic spatio-temporal solar irradiance forecasting. *IEEE Trans Sustain Energy* 11(2):571–583. <https://doi.org/10.1109/TSTE.2019.2897688>
- Kipf TN, Welling M (2016) Variational graph auto-encoders. *NIPS workshop on Bayesian deep learning*
- Mousavi SM, Zhu W, Ellsworth W, Beroza G (2019) Unsupervised clustering of seismic signals using deep convolutional autoencoders. *IEEE Geosci Remote Sens Lett* 16(11):1693–1697. <https://doi.org/10.1109/LGRS.2019.2909218>
- Qi W, Zhang X (2020) A multi-scale 3D convolution neural network for spectral-spatial classification of hyperspectral imagery. *IOP Conf Ser: Earth Environ Sci* 502:012015
- Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G (2009) The graph neural network model. *IEEE Trans Neural Netw* 20(1):61–80
- Xiao C, Chen N, Hu C, Wang K, Gong J, Chen Z (2019) Short and mid-term sea surface temperature prediction using time-series satellite data and LSTM-AdaBoost combination approach. *Remote Sens Environ* 233:111358. <https://doi.org/10.1016/j.rse.2019.111358>

Dendrogram

Rogério G. Negri
São Paulo State University (UNESP), Institute of Science and Technology (ICT), São José dos Campos, São Paulo, Brazil

Definition

A dendrogram is a graph representation able to express a hierarchical structure and the correlation among elements of a dataset.

A dendrogram comprises a tree-based diagram used to depict hierarchical relations between objects. The dissimilarity levels between the objects drive the interpretation of this graph, expressed by the heights of links that generate the clusters.

Figure 1 shows an example of a dendrogram composition regarding the set $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_6\} \subset \mathcal{X}$, whose elements are vectors or tuples defined on a space $\mathcal{X}$. The dissimilarities between the objects are usually computed through a distance measure defined in $\mathcal{X}$ and conveniently chosen according to the application purpose. If $\mathcal{X}$ is a vector space, the Euclidean distance is usually used as the dissimilarity measure. Large values imply larger dissimilarities.

A Clustering Point of View

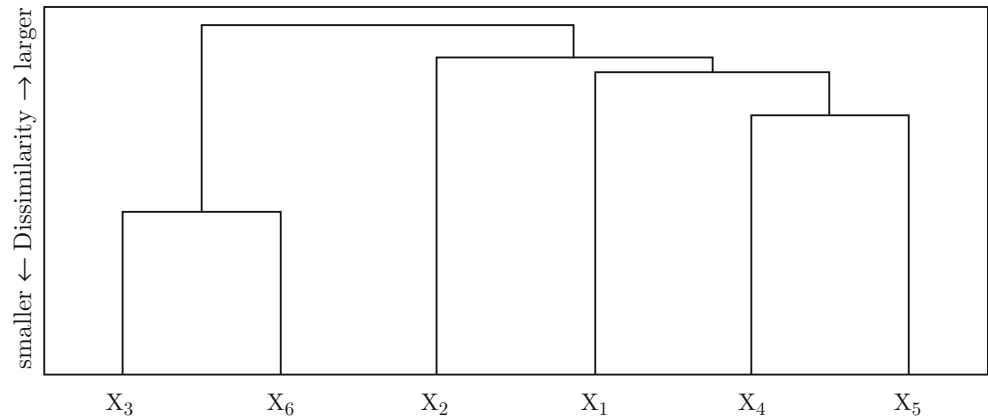
Naturally, a dendrogram provides an explicit hierarchical clustering configuration in both agglomerative and divisive approaches. Specifically, the agglomerative approach builds the hierarchical structure through consecutive clustering over the dataset until a single cluster is achieved (Fig. 1 – from the smaller to the larger dissimilarity value). In reverse, the divisive approach starts from one single cluster with all elements and then is submitted to successive splits until obtaining clusters composed by a single element (Fig. 1 – from the larger to the smaller dissimilarity value).

Different hierarchical clustering algorithms are found in the literature, for example, the Single Link, Complete Link, Average Link, and Ward's method (Theodoridis and Koutroumbas 2008; Webb and Copsey 2011). Independent of the adopted hierarchical clustering algorithm, the value expressed in a dissimilarity matrix is the basis for the clustering process. Moreover, such matrix is updated, whereas the clusters are merged or split.

In due course, $d(\mathbf{x}_i, \mathbf{x}_k)$ denotes the dissimilarity between the elements $\mathbf{x}_i$ and $\mathbf{x}_k$, where $d: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$. While $d(\mathbf{x}_i, \mathbf{x}_k) \rightarrow \infty$ represents a large dissimilarity between the compared elements, a larger similarity is observed if $d(\mathbf{x}_i, \mathbf{x}_k) \rightarrow 0$. Furthermore, $d_i(\mathbf{x}_i)$ may be adopted to denote the dissimilarity between the element $\mathbf{x}_i$ to the cluster $\mathcal{G}_j$ as well as $D(\mathcal{G}_j, \mathcal{G}_\ell)$ the dissimilarity between $\mathcal{G}_j$ and $\mathcal{G}_\ell$. Definitions for $d_i(\cdot)$ and $D(\cdot, \cdot)$ are peculiar of each method. The dissimilarity matrix at Table 1 contains the values $d(\cdot, \cdot)$ that gives place to the dendrogram of Fig. 1 using Ward's method.

It is essential to highlight that results from hierarchical clustering do not partition the original dataset into a pre-defined number of subsets but a structure that allows identifying the clusters by choosing a dissimilarity threshold. Figure 2 depicts a reanalysis of the structure at Fig. 1, where thresholds are inserted to define the clusters configuration. Using the τ_1 threshold, two clusters are defined: $\{\mathbf{x}_3, \mathbf{x}_6\}$

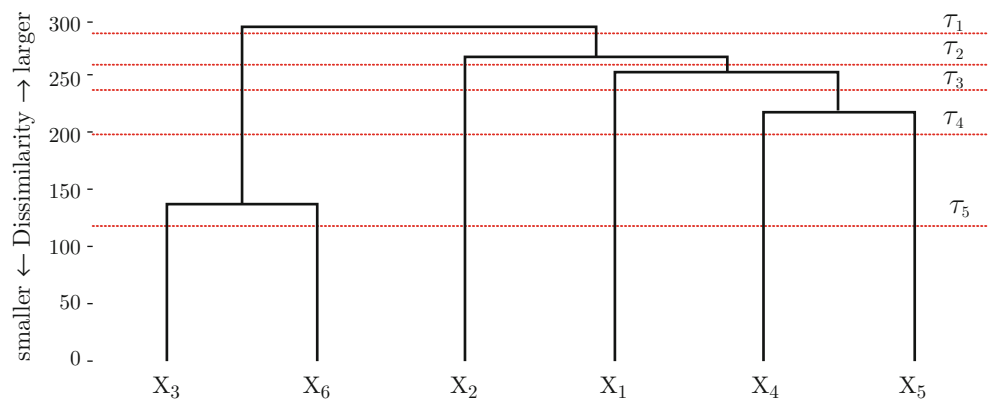
Dendrogram, Fig. 1 Example of dendrogram



Dendrogram, Table 1 Example of dissimilarity values between pairs of elements. The lower triangle portion it is omitted since $d(\cdot, \cdot)$ is symmetric

$d(\cdot, \cdot)$	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0	4	13	24	12	8
x_2		0	10	22	11	10
x_3			0	7	3	9
x_4				0	6	18
x_5					0	8.5
x_6						0

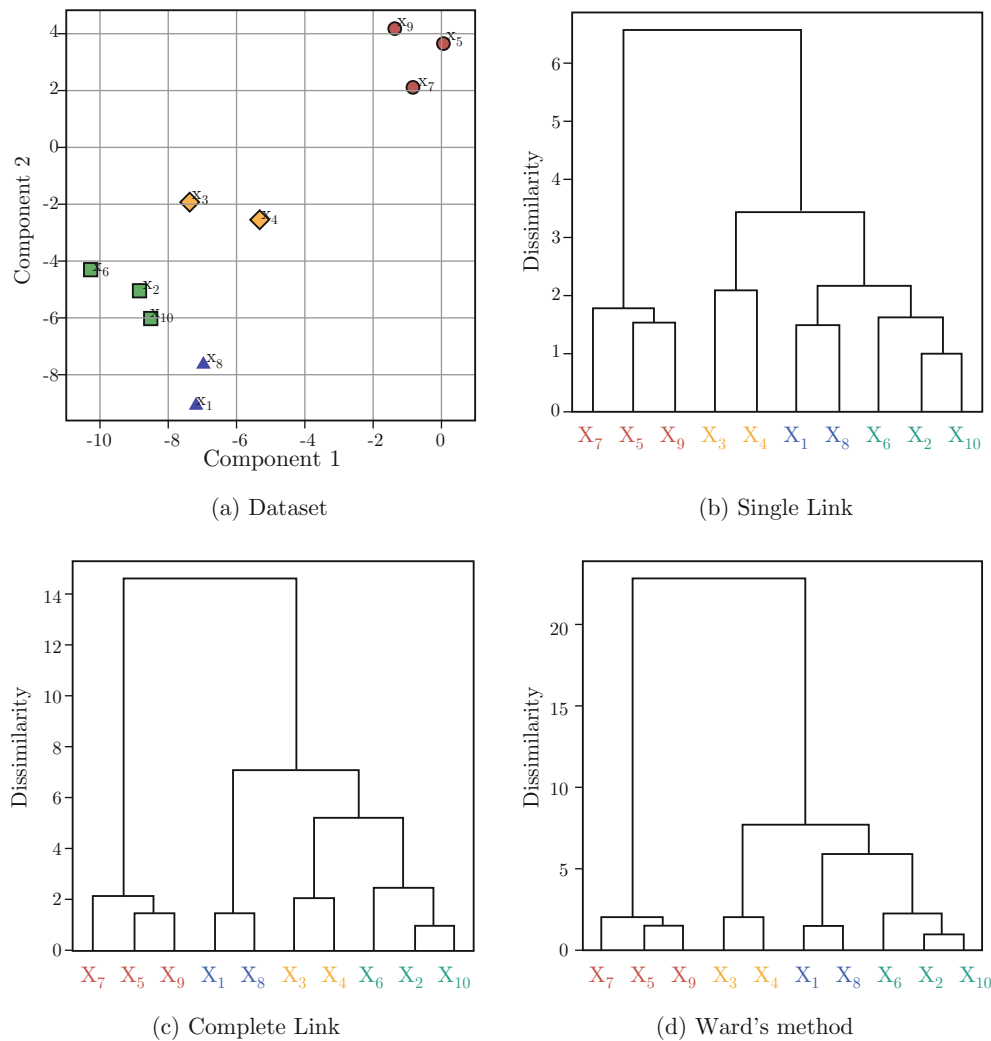
Dendrogram, Fig. 2 Examples of thresholds over a dendrogram



and $\{x_1, x_2, x_4, x_5\}$. Regarding τ_4 , five clusters are defined: $\{x_3, x_6\}$, $\{x_1\}$, $\{x_2\}$, $\{x_4\}$ and $\{x_5\}$.

Figure 3a shows a simulated dataset expressed in terms of two components (i.e., a bidimensional vector space). The different colors and shapes denote an expected pertinence regarding four clusters. The Single Link, Complete Link, and Ward's method, using the Euclidean distance as a dissimilarity measure, are applied to cluster the dataset. The results

are presented in the dendrograms of Fig. 3b–d. It is possible to observe that the dissimilarity values are not common among the methods. Moreover, some structures diverge according to the adopted method, as shown by the Complete Link, regarding the dissimilarity structure between the clusters $\{x_3, x_4\}$ and $\{x_2, x_6, x_{10}\}$, when compared to the other methods. The within-clusters dissimilarity proportion also changes, as demonstrated by the height of the links that establishes the clusters



Dendrogram, Fig. 3 An example of dataset and dendrograms from distinct hierarchical clustering methods

$\{x_3, x_4\}$ and $\{x_1, x_2, x_6, x_8, x_{10}\}$ according to the Single Link and Ward's method.

Applications in Geosciences

Recent studies in geosciences have been adopting dendrograms as a support tool for data analysis. Typically, it appears that after the use of hierarchical clustering processes, Bu et al. (2020) perform a study about hydro-chemical groundwater classification. Dendrograms were used to analyze the collected samples, expressed through diverse chemical constituent concentrations, and validate the results through distinct hierarchical clustering methods. Employing registers of resistivity, porosity, spontaneous potential, and gamma-ray profile, Torghabeh et al. (2014) employ dendrograms to understand the hierarchical relation between reservoirs of

gas shale. Based on the geographical distance from the spores' point of origin to a region in Spain, Askew and Wellman (2020) perform an analysis utilizing dendrograms to verify the dispersion of terrestrial plant, zooplankton, and phytoplankton marine communities, which allowed identifying unusual occurrences in distinct places. Garcia et al. (2015) use dendrogram to demonstrate the effectiveness of a clustering method proposed to identify classes in a shallow water environment based on reflectance information measured by remote sensing.

Summary

Dendrograms are useful to represent the relations achieved in a hierarchical clustering process. Such relations are based on dissimilarity values computed over the data clustering

process, represented by the height of the links that draw this graphic representation.

Cross-References

► [Cluster Analysis and Classification](#)

Bibliography

- Askew AJ, Wellman CH (2020) Reconstructing the terrestrial flora and marine plankton of the middle Devonian of Spain: implications for biotic interchange and palaeogeography. *J Geol Soc* 177(2). <https://doi.org/10.1144/jgs2019-080>
- Bu J, Liu W, Pan Z, Ling K (2020) Comparative study of hydrochemical classification based on different hierarchical cluster analysis methods. *Int J Environ Res Public Health* 17(24). <https://doi.org/10.3390/ijerph17249515>
- Garcia RA, Hedley JD, Tin HC, Fearn PRCS (2015) A method to analyze the potential of optical remote sensing for benthic habitat mapping. *Remote Sens* 7(10):13157–13189. <https://doi.org/10.3390/rs71013157>
- Theodoridis S, Koutroumbas K (2008) *Pattern recognition*, 4th edn. Academic, San Diego
- Torghabeh AK, Rezaee R, Moussavi-Harami R, Pradhan B, Kamali MR, Kadkhodaie-Ilkhchi A (2014) Electrofacies in gas shale from well log data via cluster analysis: a case study of the Perth basin, western Australia. *Central European Journal of Geosciences* 6:393402. <https://doi.org/10.2478/s13533-012-0177-9>
- Webb AR, Copsey KD (2011) *Statistical pattern recognition*, 3rd edn. Wiley. <https://doi.org/10.1002/9781119952954>

Differential Equations

Sergio Zlotnik and Jaume Soler Villanueva
E.T.S. de Ingeniería de Caminos, Universitat Politècnica de Catalunya, Barcelona, Spain

Definition

The characteristic that makes an equation to be called “differential” is that it describes a relationship between an unknown function and its derivatives. That is, a relationship between a quantity and the variations of the quantity with respect to the independent variables.

Introduction

A large number of natural processes are described by differential equations, and therefore, these equations have been

widely studied in the past and at present. Processes dealing with the conservation of a quantity are governed by a differential equation, making them interesting in countless applications in science and engineering. The movement of hot air generating wind, erosion of the bed of a river, temperature in the Earth’s interior, decay of radioactive elements, deformation of cortical and mantle rocks; all these processes and many others are modeled using differential equations. Continuum mechanics is built upon differential equations, and even discontinuous processes like the generation of fractures or multi-phase flows are usually modeled using differential equations.

Unfortunately, differential equations are very difficult mathematical objects, and we know how to solve just a few of the simplest ones. Although, because of their practical use, many methods have been developed to obtain approximate solutions. An approximate solution is a function that strictly does not fulfill the equations, but it is very close to the actual solution. When the difference between the approximation and exact solution is small and under control, we use the approximation rather than the exact solution for practical scientific and engineering purposes.

The solution of a differential equation is a function. When the equation is used to describe a physical process, the solution is usually a function of space and/or time. For example, let us assume we want to study the decay of a radioactive substance. What we are trying to obtain is how much substance is present at a given moment in time. We want to find a function, $Q(t)$, that describes the amount of radioactive substance at time t . The differential equation that describes (“governs”) the process reads

$$\frac{dQ(t)}{dt} = -k Q(t). \quad (1)$$

This equation provides a relationship between the unknown function, $Q(t)$, and its derivative with respect to time $\frac{dQ(t)}{dt}$ (the rate of change of Q with time). Namely, this equation states that the change of Q in time is proportional to Q with a proportionality constant $-k$.

A second example is the temperature distribution in the Earth’s crust. The temperature $T(x, y, z)$ is a function of the spatial coordinates and, in equilibrium, satisfies

$$-k \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} \right) = \rho H, \quad (2)$$

where k is the (constant) thermal conductivity, ρ is the density, and H is a heat source (e.g., the heat generated by the decay of radioactive elements). To simplify the notation, we do not write the dependence on the spatial coordinates of the temperature T in Eq. (2). Both ρ and H can be functions of space,

although Eq. (2) is a simplification of a more general case assuming that k is a scalar and that it does not change in space.

In order to seek the solution of a differential equation, we need to specify the domain in which the equation must be satisfied. That is, the range where the independent variables can move. An example of a domain in Eq. (1) can be any value of t from t_0 to t_{end} . In Eq. (2), the domain is a region in space and is usually denoted by $\Omega \in \mathbb{R}^3$. The spatial domain might take complex shapes, for example, the volume of the crust, from the surface to the Moho, between some latitude and longitude. This volume has a curved surface and might have a very complex base. When the solution depends on both, space and time, then the domain is composed of both the space and time parts.

The solutions of Eqs. (1) and (2) on their corresponding domains are not unique. There is a family of solutions that fulfill each one of the equations. The problem will be mathematically well posed only when the solution is unique. The uniqueness of the solution is obtained when the problem is closed with appropriate boundary conditions. Boundary conditions provide information of the solution at the boundary of the domain, determining which is the actual solution within the family described by the equation. An example of boundary condition for Eq. (1) is $Q(t_0) = 10$, meaning that, at the initial time t_0 the quantity of substance is known and takes the value 10. Boundary conditions for Eq. (2) could be of different kinds, for example, (i) assuming that the temperature at the surface is known and/or (ii) the heat flux at the base of the domain is known. These boundary conditions are reasonable from the physical point of view. Not all the boundary conditions close a problem properly; it is possible that boundary conditions are contradictory or insufficient. Therefore, attention must be paid when specifying them. In practice, when using differential equations to model some physical phenomena, it is usually more challenging to determine the proper boundary conditions, and domain and material properties, than actually solve the equations.

For the complete problem governed by Eq. (1), including the domain and boundary conditions reads, find $Q(t)$ such that

$$\begin{cases} \frac{dQ(t)}{dt} = -k Q(t) & \text{for } t \in [t_0, t_{\text{end}}], \\ Q(t_0) = 10 \end{cases} \quad (3)$$

and similarly, for Eq. (2) the problem reads, find $T(x, y, z)$ such that

$$\begin{cases} -k \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} \right) = \rho H & \text{in } \Omega \\ T = T_D & \text{on } \Gamma_D \\ -k \nabla T \cdot \mathbf{n} = f_N & \text{on } \Gamma_N. \end{cases} \quad (4)$$

In Eq. (4), the boundary is denoted by Γ and is partitioned into two nonoverlapping parts: Γ_D and Γ_N . The former is where the solution is known, and the latter is where the heat flow normal to the boundary is known. The vector $\mathbf{n}$ indicates the direction normal to the boundary.

Systems of differential equations, where several of many equations are coupled to each other, are usually found in practice. The book of Zill (2013) provides useful examples, but there is a vast literature available on the topic.

Classification

Differential equations are a large family of equations with different characteristics, and therefore, there are many different ways to classify them. Equation (1) is called an ordinary differential equation (ODE) because all its derivatives are with respect to one variable (in this case, t). When the differential equation has derivatives with respect to several variables, as in Eq. (2), it is called partial differential equation (PDE). This classification is important as the methods used to solve each case are different. Other classifications that are relevant to choosing the solution method follow.

The boundary conditions that close an ODE define two types of problems: in an initial value problem, we start from a known point and seek the solution advancing from that initial point. This is the case of problem (3), and it is the usual case for problems depending on time. Methods to solve initial value problems are called marching methods or time-integration methods. Note that in problem (3) the time domain could be closed (e.g., $t \in [t_0, t_{\text{end}}]$) or could be open (e.g., $t > t_0$). A marching scheme can continue advancing in time while the solution is of interest.

If boundary conditions are specified at different boundary points, e.g., the solution is known at t_0 and t_{end} , then it is called a boundary value problem. The solution in this case describes an equilibrium between the known boundary points, and, in this case, the domain must be closed.

The order of a differential equation is the order of its highest derivative. Equation (1) is order 1, and Eq. (2) is order 2. ODEs of order n can be easily converted in a system of n ODEs of order 1. This is convenient as it is simpler to build methods to solve first-order ordinary equations.

PDEs used in practice to describe natural processes are usually first or second order (fourth-order equations are found in some particular applications but are not as usual).

PDEs of order 2 are classified depending on the kind of physical problem they describe: propagation problems involve space and time domains. Usually, these are initial value problems in time and boundary value problems in

space, and therefore, the complete space solution must be known as initial boundary condition. This particular type of boundary condition is usually called *initial condition*.

Equilibrium problems are steady-state problems (do not involve time) in a closed-space domain. They describe equilibrium solutions. The methods used to solve them are called relaxation methods.

Eigenproblems are special equilibrium problems that depend on space and one parameter. The solution exists only for some special values of that parameter (i.e., eigenvalues). The eigenvalues are not known a priori and must be determined together with the equilibrium configuration of the system.

History

The history of differential equations goes back to the beginning of infinitesimal calculus, when it was realized that most laws of physics actually reduce to differential equations, either ordinary or partial.

In the eighteenth century, differential equations provided an accurate model of the solar system. The Newtonian gravitational interaction between n planets and the sun is a system of $3(n + 1)$ second-order ODEs, equivalent to a system of $6(n + 1)$ first-order equations. Astronomical applications were developed by Euler, Gauss, Lagrange, Laplace, Jacobi, Poincaré, and many others (e.g., see Kline 1990).

PDEs were introduced to model the gravitational Newtonian potential (Laplace equation), motion of fluids, and behavior of elastic bodies. Since the nineteenth century, extraordinary advances have been made in the scientific and technological applications of differential equations. In 1862, James Clerk Maxwell published his electromagnetic field theory based on a set of partial equations, and Einstein's field equations of general relativity are a system of PDE (e.g., see Gray 2021). The solution of very complex ODE and PDE is crucial in modern engineering disciplines such as generation and transmission of electric power, aerodynamics of aircraft, and calculus of structures in civil engineering, just to mention a few.

Solution Methods

Only a few very simple differential equations can be solved by a closed formula involving elementary functions. In this situation, two different routes are possible. First, to somehow simplify the equation and find analytical (i.e., given by mathematical expressions) approximate solutions. This is

the traditional approach taken in astronomy with the theory of small perturbations. Second, to resort to numerical approximations, which give directly the numerical solution but no analytical expression. Numerical methods used just elementary arithmetic operations but had a major difficulty: the number of operations could easily be counted in billions.

A great number of approximate numerical methods have been (and are still being) developed. The name “approximate” must be understood in the sense that the accuracy of the numerical solution can be as high as needed, but the smaller the intended error, the higher the computational effort is. However, the digital computer has made it possible to carry out billions of operations in seconds.

The majority of the methods are based on a “discretization” of the equation, which means approximating the derivatives in the equation by finite quotients of increments. A number of numerical methods for ODEs were developed during the nineteenth century, mostly for astronomical applications. At the end of that century and early twentieth, the Runge–Kutta methods were introduced and are currently the methods of choice except for some very specific problems, for example, computing the orbits of the planets (Sidelmann, 1992). Methods for PDEs have seen a similar evolution, with the only drawback that they usually need a much more extensive computational effort. Around the mid-twentieth century, the finite element methods were introduced (e.g., see Gupta and Meek 1996 for a historical review) and, together with the availability of digital computers, have become the most widely used methods in science and engineering.

There are many introductory and advanced books on numerical methods for differential equations. Some examples worth recommendation are the books of Zienkiewicz et al. (2013) for finite element methods, Hoffman (1992) for finite difference methods, and Leveque (1992) for finite volume methods.

Notation

The notation used to write and explain differential equation is diverse and varies depending on the book, authors, and application field. Each different notation emphasizes a different aspect of the equation or becomes convenient in a specific situation, but all express the same ideas. Here, we present some usual notations to allow the reader to translate them into the form that she/he feels more familiar with.

Scalar numbers are a, b, x, y, i, j (usually i and j are integers and other letters are real valued). Vectorial quantities are denoted as $\mathbf{v}, \bar{\mathbf{v}}$ or in index notation v_i . Tensor

quantities are $\mathbf{K}$, $K_{i,j}$, or $\underline{K}$. Derivatives in ODEs are denoted as $\frac{dQ}{dt}$ and in PDEs as $\frac{\partial u}{\partial x}$ or $\partial_x u$ and, more rarely, (but usual in the context of finite elements) $u_{,x}$. Derivatives with respect to time are usual and have a special (shorter) notation: $\dot{u}$. The most usual operators are gradient and divergence. Gradients are noted as ∇u , $\text{grad } u$, and, using index notation $\frac{\partial u}{\partial x_i}$, $\partial_i u$ or $\partial u_{,i}$, while divergence as $\nabla \cdot \mathbf{u}$, $\text{div } \mathbf{u}$, or in index notation $\partial_i u_i$. One special operator, the Laplacian, is commonly written as $\nabla^2 u$ or Δu .

Conclusion

Differential equations are widely used to model physical phenomena. They describe processes in which the variables are continuous (continuum mechanics is built upon PDEs), but they are sometimes used to model noncontinuous processes (e.g., fractures). Numerical methods are the most common procedure to seek a solution. These numerical solutions are approximate but useful in practice when errors are under control. Caution should be taken in this regard as the decisions taken during the numerical solution procedure might undermine the results and produce inaccurate and non-physical answers.

Cross-References

- [Eigenvalues and Eigenvectors](#)
- [Ordinary Differential Equations](#)
- [Partial Differential Equations](#)

Bibliography

- Gray J (2021) Change and variations. A history of differential equations to 1900. Springer
- Gupta K, Meek J (1996) A brief history of the beginning of the finite element method. *Comput Methods Appl Mech Eng* 39: 3761–3774
- Hoffman J (1992) Numerical methods for engineers and scientists. McGraw-Hill
- Kline M (1990) Mathematical thought from ancient to modern times. Oxford University Press
- Leveque R (1992) Numerical methods for conservation laws. Lectures in Mathematics, ETH Zürich
- Sidelmann PK (1992) The computation of solar system ephemerides at the JPL using a variable stepsize and variable order multistep method. In: Explanatory supplement to the astronomical almanac. University Science Books
- Zienkiewicz OC, Taylor RL, Zhu J (2013) The finite element method: its basis and fundamentals, 7th edn. Butterworth-Heinemann
- Zill DG (2013) A first course in differential equations with modeling applications, 10th edn. Brooks/Cole, Cengage Learning

Digital Elevation Model

Pradeep Srivastava¹ and Sanjay Singh²

¹Earth Sciences, Indian Institute of Technology, Gandhinagar, Gujarat, India

²Signal and Image Processing Group, Space Applications Centre (ISRO), Ahmedabad, GJ, India

Definition and Scope

Digital elevation model (DEM) is a digital, three-dimensional representation of the terrain under consideration. Elevations are defined with respect to a reference surface known as Datum.

Datum is a mathematical representation of the shape and size of the earth. It identifies the origin and orientations of the reference axes for map projection and the reference surface for height measurements. Different datum exists for different regions of the globe, to minimize the error between the geoid (equipotential surface of the earth) and the ellipsoid (mathematical representation for that region). WGS 84 is a universal datum, used as reference ellipsoid by most of the data providers, which is based on Earth Gravitational Model (EGM 2008). The scope of DEM generation in this chapter is to determine elevation above the ellipsoid.

Digital elevation model (DEM) and digital surface model (DSM) are closely related themes used in Earth Sciences to characterize the geometric properties of the Earth's surface. The emphasis in DEM is on determining the elevation of a given point on the Earth's surface above a reference surface (geoid, ellipsoid) and study the behavior of data set of elevation values over a region. In DSM the characteristics of the two (or fractal) dimensional surface of the Earth is studied. In this Encyclopedia we have attempted to treat the two topics as complementary. Accordingly, mathematics of determining the elevation from noncontact (including remotely sensed) observations is given in this chapter on DEM. The mathematics related to characterization of Earth's surface as geometrical or topological artifact is given in the chapter titled digital surface model (DSM).

Introduction

The non contact observations (from satellite, areal, UAV, or terrestrial observations) form the data sets for determining the elevation of the ground point above a given reference surface (Florinsky 2016). The discipline catering to stereo observations and determination of three-dimensional coordinates from these observations is called (Aerial, Satellite, Structure from Motion) Photogrammetry.

Under ideal conditions the elevation h can be computed simply by knowing the stereoscopic parallax as given in Eq. 1.

$$h = H - B * f / p \quad (1)$$

where B is base length, H is height of the platform, f is focal length of the camera and p is the absolute parallax.

The relation of height accuracy versus B/H is given in Eq. 2.

$$\Delta h = \Delta s / \left(\frac{B}{H} \right) \quad (2)$$

Here Δs is the error in identification of pixel and Δh is the error in height.

Extracting Elevation/3D Coordinates from Stereo Imagery: A Generic Photogrammetric Data Adjustment Approach

A photogrammetric approach to determine heights involves (a) determination of position and orientation of the platforms (airborne, satellite, UAV/Drones) followed by (b) computation of heights using the precise knowledge of the orientation. The reconstruction of stereo model depends upon the platform and sensor (frame or line scanner) used to acquire stereo imagery. Bundle adjustment is a technique, which tries to reconstruct the imaging geometry for all bundles of rays emanating from the ground to the sensor at one go. This technique includes modeling of all orientation parameters, viz., orbit, attitude, and camera geometry, and adjusts them simultaneously using precise imaging geometry in the presence of a set of limited GPS observed highly accurate control points and initial knowledge of all parameters, resulting in improvement of interior, exterior orientation parameters, and generating ground coordinates of the conjugate points simultaneously (Gopala Krishna and Lehner 1994; Grodecki et al. 2004; Dave 2013). The possible parameters that can be adjusted during block adjustment are as follows:

Interior orientation parameters

- Focal length f
- Principle point coordinates (x, y)

Exterior orientation parameters

- Orbit parameters (Position (X, Y, Z) , Velocity Vector $(Xdot, Ydot, Zdot)$ as a function of time (for satellite-based platforms).

- Attitude parameters (Roll, Pitch and Yaw). They vary as a function of time for satellite-based platforms.

Ground Coordinates

- Ground coordinates of the conjugate points (X, Y, Z)

The relationship between the ground feature and corresponding point along with acquisition location or perspective center is specified by collinearity equations. The observation equations are derived by linearizing the collinearity condition equations with respect to each of the parameters. Two observation equations are obtained for each ground point in one image.

The combined form of observation equations, treating orientation parameters and ground coordinates as unknown, can be written as Eq. (3).

$$\begin{aligned} & \left(\frac{\partial E1}{\partial \alpha_0} \right)^0 \Delta \alpha_0 + \left(\frac{\partial E1}{\partial \alpha_1} \right)^0 \Delta \alpha_1 + \left(\frac{\partial E1}{\partial \beta_0} \right)^0 \Delta \beta_0 + \left(\frac{\partial E1}{\partial \beta_1} \right)^0 \Delta \beta_1 \\ & + \left(\frac{\partial E1}{\partial \gamma_0} \right)^0 \Delta \gamma_0 + \left(\frac{\partial E1}{\partial \gamma_1} \right)^0 \Delta \gamma_1 + \left(\frac{\partial E1}{\partial X} \right)^0 \Delta X \\ & + \left(\frac{\partial E1}{\partial Y} \right)^0 \Delta Y + \left(\frac{\partial E1}{\partial Z} \right)^0 \Delta Z + V_{L_j} = -E1^0 \\ & \left(\frac{\partial E2}{\partial \alpha_0} \right)^0 \Delta \alpha_0 + \left(\frac{\partial E2}{\partial \alpha_1} \right)^0 \Delta \alpha_1 + \left(\frac{\partial E2}{\partial \beta_0} \right)^0 \Delta \beta_0 \\ & + \left(\frac{\partial E2}{\partial \beta_1} \right)^0 \Delta \beta_1 + \left(\frac{\partial E2}{\partial \gamma_0} \right)^0 \Delta \gamma_0 + \left(\frac{\partial E2}{\partial \gamma_1} \right)^0 \Delta \gamma_1 \\ & + \left(\frac{\partial E2}{\partial X} \right)^0 \Delta X + \left(\frac{\partial E2}{\partial Y} \right)^0 \Delta Y + \left(\frac{\partial E2}{\partial Z} \right)^0 \Delta Z + V_{S_j} = -E2^0 \end{aligned} \quad (3)$$

where V_{L_j} and V_{S_j} are residuals along line and pixel directions to account for random errors, $(\alpha_0, \alpha_1, \dots, \beta_0, \beta_1, \dots, \gamma_0, \gamma_1, \dots)$ are the unknown orientation parameters, (X, Y, Z) are ground coordinates to be determined.

Each point in a given image will give two equations (4).

$$\begin{matrix} V_i & + & B_{I_{ij}} & \Delta I_i & + & B_{O_{ij}} & \Delta O_j & = & \epsilon_{ij} \\ (2,1) & & (2,6) & (6,1) & & (2,3) & (3,1) & & (2,1) \end{matrix} \quad (4)$$

V_i denotes the corrections needed to account for the random error. The subscript I is used to denote corrections to the orientation parameters and subscript O is used to denote corrections to the ground coordinates.

Assuming there are n points which are imaged in m images, the system of equations using Eq. (4) can be written as

$$\begin{aligned}
& \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ \vdots \\ V_n \end{bmatrix} + \begin{bmatrix} B_{I_1} \\ B_{I_2} \\ B_{I_3} \\ \vdots \\ B_{I_n} \end{bmatrix} \Delta_I \\
& + \begin{bmatrix} B_{O_1} & & & \\ & B_{O_2} & & \\ & & B_{O_3} & \\ & & & \ddots \\ & & & & \ddots \end{bmatrix} \begin{bmatrix} \Delta_{O_1} \\ \Delta_{O_2} \\ \Delta_{O_3} \\ \vdots \\ \Delta_{O_n} \end{bmatrix} \\
& = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}
\end{aligned}$$

i.e.,

$$\begin{aligned}
& \begin{matrix} V & B_I & \Delta_I & B_O & \Delta_O \\ (2mn, 1) & (2mn, 6m) & (6m, 1) & (2mn, 3n) & (3n, 1) \end{matrix} \\
& + \\
& \begin{matrix} \epsilon \\ (2mn, 1) \end{matrix}
\end{aligned} \quad (5)$$

Similarly, observation equations for orientation parameters (Eq. (6)) and ground coordinates (Eq. (7)) can be stacked as

$$\begin{matrix} V_I & -\Delta_I & = C_I \\ (6m, 1) & (6m, 1) & (6m, 1) \end{matrix} \quad (6)$$

$$\begin{matrix} V_O & -\Delta_O & = C_O \\ (3n, 1) & (3n, 1) & (3n, 1) \end{matrix} \quad (7)$$

Thus, combining Eqs. (5), (6), and (7) gives a complete mathematical model of the photogrammetric block adjustment problem.

$$V + B_I \Delta_I + B_O \Delta_O = \epsilon$$

$$V_I - \Delta_I = C_I$$

$$V_O - \Delta_O = C_O \quad (8)$$

Expressing this equation as a single matrix equation, we have

$$\begin{aligned}
& \begin{bmatrix} V \\ V_I \\ V_O \end{bmatrix} + \begin{bmatrix} B_I & B_O \\ -1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \Delta_I \\ \Delta_O \end{bmatrix} = \begin{bmatrix} \epsilon \\ C_I \\ C_O \end{bmatrix} \\
& \bar{V} + \bar{B} \Delta = \bar{C} \quad (9)
\end{aligned}$$

Let $\bar{W}$ be a weight matrix.

$$\bar{W} = \begin{bmatrix} W & & \\ & W_I & \\ & & W_O \end{bmatrix}$$

where W is $(2mn, 2mn)$ matrix for weight of collinearity condition. W_I is $(6m, 6m)$ matrix for weight of orientation parameters and W_O is $(3n, 3n)$ matrix for weight of the object-space coordinates

A least square solution of the model in above equation results in the following normal equation:

$$(\bar{B}^T \bar{W} \bar{B}) \Delta = \bar{B}^T \bar{W} \bar{C} \quad (10)$$

The set of normal equation may be expressed as

$$N \Delta = K \quad (11)$$

The following interesting observations can be made from the Eq. (11) (Slama 1980).

- The N matrix is symmetrical about the principal diagonal.
- The upper left-hand corner of the N matrix consists of (6×6) – submatrices along the principal diagonal. Each matrix is corresponding to one particular image, and all the other elements outside of these submatrices are zero.
- The lower right-hand corner of the N matrix consists of (3×3) – submatrices along the principal diagonal. Each matrix is corresponding to particular point, and all the other elements outside of these submatrices are zero.
- The contribution of each set of the Block Adjustment observation equations can also be identified within the normal equation.

Equation (11) represents normal equation. It contains the set of equations which is represented in the form of matrix as follows:

$$\begin{matrix} N & \Delta & = K \\ (6m + 3n, 6m + 3n) & (6m + 3n, 1) & (6m + 3n, 1) \end{matrix} \quad (12)$$

The **N** matrix is a sparse and is not well conditioned.

These two issues are to be accounted in the implementation for solving normal equations. The direct solution of this set of equations will involve solving $(6m + 3n)$ simultaneous equations. Even for a small block of photographs, the size of the normal equation can become very large making the solution impractical. Instead of solving large matrix in a single shot, the above set of equations are further reduced in smaller sizes and solve by substitution method. The unknowns are obtained or refined iteratively till the convergence is achieved.

Satellite Photogrammetry for Line-scanner Imagery

This section deals with DEM generation using stereo images acquired by optical linear array sensors on-board satellite platforms like SPOT-5, Cartosat-1 which are dedicated satellites acquiring stereo imagery (Srivastava et al. 2007). The photogrammetric models are used to mathematically describe the relationship between ground coordinates and image coordinates (Srivastava et al. 2020; Srinivasan et al. 2013; Dave 2013). The two popular modeling approaches are: (i) Rigorous Sensor Model (physical model called RSM) and (ii) Rational Polynomial Model (RPM). Either of the models is used to reconstruct a precise relationship between 2D images to 3D ground coordinates using orbit and attitude data, available as part of ancillary data. The core of the modeling is the establishment of the precise viewing direction or look vector in the presence of distortions arising from camera tilt, spacecraft motion and orientation, earth rotation and curvature in addition to surface relief through series of coordinate transformations specially chosen for each mission with proper definition and conventions.

Important points to be considered in line scan imagery obtained from satellite borne optical sensors (Srivastava and Gopala Krishna 2020):

- **Orbital Constraints:** The position of the imaging camera mounted on a satellite follows its orbit so that the camera positions corresponding to imaged lines are related by an orbital constraint. As is well known from orbital mechanics, the orbital path of a free flying satellite around the Earth is an ellipse in an inertial coordinate system. In an Earth-centered Earth-fixed coordinate system

(a coordinate system used for providing ground position), it is a complex curve in space.

- **Attitude behavior within a scene:** Attitude is the orientation of the imaging camera with respect to the ground coordinate system. This is a straightforward concept when working with aerial frame photographs. There is only one set of angles (roll, pitch, yaw, or quaternions) which represent the attitude of imaging geometry. Usually the value is initiated to zeros and determined as part of photogrammetric adjustment. In this case, the value is changing with every imaging line, and the relation between body coordinate system and the ECEF ground coordinate system is complex, even some transformations are not orthogonal. The attitude, as understood by and as measured by the onboard Attitude Orbit Control System (AOCS), is the orientation of the body coordinate system with respect to the orbital coordinate system and not with respect to the ground coordinate system as understood in photogrammetry. The remaining relation needs to be worked out from the data on orbital position of the satellite at the time.

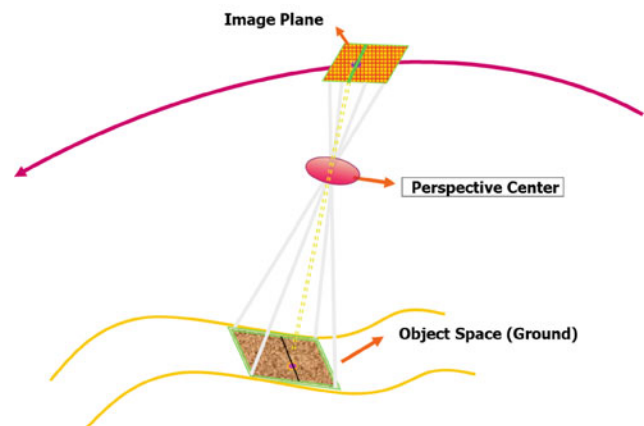
This section deals with the photogrammetric models and their applications in deriving orientation parameters and Digital Elevation Models.

Rigorous Sensor Model (RSM)

The RSM is the physical model which is based on the collinearity condition which states that the perspective center, image point and the corresponding object space point all lie in same line (Fig. 1).

Mathematically it can be stated as

$$\begin{bmatrix} f \\ -x \\ -y \end{bmatrix} = KM \begin{bmatrix} X_A - X_S \\ Y_A - Y_S \\ Z_A - Z_S \end{bmatrix} \quad (13)$$



Digital Elevation Model, Fig. 1 Imaging sensor geometry

where $\begin{bmatrix} f \\ -x \\ -y \end{bmatrix}$ is the focal plane coordinates, $\begin{bmatrix} X_A \\ Y_A \\ Z_A \end{bmatrix}$ are the ground coordinates and $\begin{bmatrix} X_S \\ Y_S \\ Z_S \end{bmatrix}$ are the coordinates of the perspective centre. M1, M2, and M3 are the first, second, and third rows of the rotation matrix M.

Coordinate Systems

This section discusses the various coordinate systems required to establish the transformations. The following coordinate systems are used to carry out the required transformation in the imaging sensor model (Gopala et al. 2010; Srinivasan et al. 2013):

- (i) Earth-centered Inertial coordinate system (ECI)
- (ii) Earth-centered Earth Fixed system (ECEF)
- (iii) Local orbital coordinate system
- (iv) Spacecraft body coordinate system
- (v) Image or Focal Plane coordinate system (image space).

Usually both mappings, that is, relating image to ground coordinates and vice versa, are required for data processing towards DEM generation. Details on how the above systems are related to one another are discussed in the following sections.

Estimation of Image Coordinates from Ground Coordinates

The sequence of transformations for computing the image coordinates using ground coordinates as inputs is as follows:

- Latitude, Longitude, and Height (ϕ, λ, h) of a ground point is converted to the Cartesian ECEF coordinate system.
- Transformation from ECEF to ECI coordinates uses sideral angles.
- ECI coordinates are transformed to orbital frame of reference using state vectors (X, Y, Z – position and $\dot{X}, \dot{Y}, \dot{Z}$ – velocity).
- Orbital system coordinates to body frame coordinate system using the attitude.
- The body frame coordinate system to Sensor (imaging payload or camera) is carried out using the spacecraft mounting angles and alignment angles.
- Sensor frame to image plane or Focal plane coordinate system is performed using the payload geometry of the sensor under consideration and then is converted to image coordinate system (time or scanline, pixel).

Since the ‘time’ is the unknown parameter in estimation of image coordinates, direct solution is not possible but estimated iteratively using collinearity condition between two known times.

Transformation matrix **M** is a product of three rotation matrices:

M = **RL** * **RA** * **RO**, where

RL is look angle rotation matrix transformation from spacecraft body coordinate system to sensor image coordinate system.

RA is attitude rotation matrix transformation from orbital coordinate system to spacecraft body coordinate system and it is given as a product of three rotations Rpitch, Rroll, and Ryaw.

RA = Ryaw * Rroll * Rpitch

RA matrix is filled by using orbit to body attitude in True of Date (TOD) reference system, and **RO** is orbit rotation matrix transformation from ECI coordinate system to orbital coordinate system.

Estimation of Ground Coordinates from Image Coordinates

This process is the reverse of the ground to image computation process mentioned in the above section. This is directly using the collinearity equations. As such, this is the solution of the look vector intersecting the earth surface (ellipsoid) using the above sequence of transformation from sensor frame to geocentric and then computing the ground latitude, longitude. Precise height of the object is found out using stereo imaging geometry. The above image to ground relation is considered as “Rigorous Sensor Model” of an image and is used for transformation between 2D image space and the 3D object space. It includes the physical parameters about the camera, such as focal length, principal point location, pixel size, lens distortions, and orientation parameters of the image such as position and attitude.

Rational Polynomial Model (RPM) (Grodecki 2001; Tao and Ha 2001; Sanjay et al. 2008)

In this approach, the complex imaging geometry involving orbit, attitude, camera geometry, etc. is modeled through Rational Polynomial Coefficient (RPC). Each scanline number can be expressed as a function of ground coordinates in terms of ratio of cubic polynomials. The constant term in the denominator is taken as 1. So for a given scanline number, we have

$$scan = \frac{P_1(X, Y, Z)}{P_2(X, Y, Z)} \quad (14)$$

Where,

$$\begin{aligned} P_1(X, Y, Z) &= \sum_{i,j,k=0}^{i+j+k \leq 3} a_{ijk} X^i Y^j Z^k, \quad P_2(X, Y, Z) \\ &= \sum_{i,j,k=0}^{i+j+k \leq 3} b_{ijk} X^i Y^j Z^k, \quad b_{000} = 1 \end{aligned}$$

Similarly each pixel number can also be expressed as a function of ground coordinates as ratio of cubic polynomials, that is,

$$pixel = \frac{P_3(X, Y, Z)}{P_4(X, Y, Z)} \quad (15)$$

Where,

$$\begin{aligned} P_3(X, Y, Z) &= \sum_{i,j,k=0}^{i+j+k \leq 3} c_{ijk} X^i Y^j Z^k, \quad P_4(X, Y, Z) \\ &= \sum_{i,j,k=0}^{i+j+k \leq 3} d_{ijk} X^i Y^j Z^k, \quad d_{000} = 1 \end{aligned}$$

Here (X, Y, Z) are the normalized object space coordinates, that is, normalized latitude, longitude, and height, respectively. *scan* and *pixel* are the normalized scanline number and pixel number between $(-1, +1)$.

Some Applications of Photogrammetric Adjustment Are:

Reconstruction of Attitude Behavior

Derived collinearity equations from Eq. (13) can be expressed as

$$E1 = (xM1 + fM2) [\overline{XA} - \overline{XS}] \quad (16)$$

$$E2 = (yM1 + fM3) [\overline{XA} - \overline{XS}] \quad (17)$$

where $\overline{XA} = \begin{bmatrix} X_A \\ Y_A \\ Z_A \end{bmatrix}$, $\overline{XS} = \begin{bmatrix} X_S \\ Y_S \\ Z_S \end{bmatrix}$ and $M1, M2$, and $M3$ are the first, second and third rows of the rotation matrix M .

Rotation matrix M is the function of orbit as well as attitude, as explained in the previous section. The attitude refinement model is normally assumed to be

$$roll = originalroll + \alpha_0 + \alpha_1 * time + \alpha_2 * time^2 + \dots$$

$$pitch = originalpitch + \beta_0 + \beta_1 * time + \beta_2 * time^2 + \dots$$

$$yaw = originalyaw + \gamma_0 + \gamma_1 * time + \gamma_2 * time^2 + \dots$$

where, $(\alpha_0, \alpha_1, \dots, \beta_0, \beta_1, \dots, \gamma_0, \gamma_1, \dots)$ are the time varying error model coefficients for roll, pitch, and yaw, respectively. First the Eqs. (16) and (17) are linearized by Taylor's series expansion (Eqs (18) and (19)) which will contain differentials of refinement parameters chosen for modelling.

$$\begin{aligned} &\left(\frac{\partial E1}{\partial \alpha_0}\right)^0 \Delta \alpha_0 + \left(\frac{\partial E1}{\partial \alpha_1}\right)^0 \Delta \alpha_1 + \left(\frac{\partial E1}{\partial \beta_0}\right)^0 \Delta \beta_0 \\ &+ \left(\frac{\partial E1}{\partial \beta_1}\right)^0 \Delta \beta_1 + \left(\frac{\partial E1}{\partial \gamma_0}\right)^0 \Delta \gamma_0 + \left(\frac{\partial E1}{\partial \gamma_1}\right)^0 \Delta \gamma_1 = -E1^0 \end{aligned} \quad (18)$$

$$\begin{aligned} &\left(\frac{\partial E2}{\partial \alpha_0}\right)^0 \Delta \alpha_0 + \left(\frac{\partial E2}{\partial \alpha_1}\right)^0 \Delta \alpha_1 + \left(\frac{\partial E2}{\partial \beta_0}\right)^0 \Delta \beta_0 \\ &+ \left(\frac{\partial E2}{\partial \beta_1}\right)^0 \Delta \beta_1 + \left(\frac{\partial E2}{\partial \gamma_0}\right)^0 \Delta \gamma_0 + \left(\frac{\partial E2}{\partial \gamma_1}\right)^0 \Delta \gamma_1 = -E2^0 \end{aligned} \quad (19)$$

The equation (9) which describes the general formulation of block adjustment can be reduced to include only the refinement parameters for attitude, resulting in small set of equations for a single image.

$$[V] + [B_I][\Delta_I] = [e] \quad (20)$$

Using the above equation, the refinement parameters for attitude are computed as least squares solution. The parameters are computed iteratively and hence the corrections to the initial approximate values are updated at each iteration.

Reconstruction of Orbital Parameters

In orbit-based space resection, the orbital parameters inclination, longitude of ascending node, and true anomaly are chosen for modeling. The collinearity equations (16) and (17) are linearized by Taylor series expansion with respect to orbital elements. The orbit model is chosen as

$$\begin{aligned} \text{inclination} &= \text{originalinclination} + i_0 + i_1 * \text{time} \\ &+ i_2 * \text{time}^2 + \dots \end{aligned}$$

$$\begin{aligned} \text{longitudeofascendingnode} &= \text{originallongitudeofascendingnode} \\ &+ \Omega_0 + \Omega_1 * \text{time} + \Omega_2 * \text{time}^2 \\ &+ \dots \end{aligned}$$

$$\begin{aligned} \text{trueanomaly} &= \text{originaltrueanomaly} + F_0 + F_1 * \text{time} + F_2 \\ &* \text{time}^2 + \dots \end{aligned}$$

where, $(i_0, i_1, \dots, \Omega_0, \Omega_1, \dots, F_0, F_1, \dots)$ are the time varying error model coefficients for inclination, longitude of ascending node, and true anomaly, respectively.

The error model for orbit is assumed as a time varying components. The parameters (deltas) are computed by least square method from the following equations:

$$\begin{aligned} & \left(\frac{\partial E1}{\partial i_0} \right)^0 \Delta i_0 + \left(\frac{\partial E1}{\partial i_1} \right)^0 \Delta i_1 + \left(\frac{\partial E1}{\partial \Omega_0} \right)^0 \Delta \Omega_0 \\ & + \left(\frac{\partial E1}{\partial \Omega_1} \right)^0 \Delta \Omega_1 + \left(\frac{\partial E1}{\partial F_0} \right)^0 \Delta F_0 + \left(\frac{\partial E1}{\partial F_1} \right)^0 \Delta F_1 = -E1^0 \end{aligned} \quad (21)$$

$$\begin{aligned} & \left(\frac{\partial E2}{\partial i_0} \right)^0 \Delta i_0 + \left(\frac{\partial E2}{\partial i_1} \right)^0 \Delta i_1 + \left(\frac{\partial E2}{\partial \Omega_0} \right)^0 \Delta \Omega_0 \\ & + \left(\frac{\partial E2}{\partial \Omega_1} \right)^0 \Delta \Omega_1 + \left(\frac{\partial E2}{\partial F_0} \right)^0 \Delta F_0 + \left(\frac{\partial E2}{\partial F_1} \right)^0 \Delta F_1 = -E2^0 \end{aligned} \quad (22)$$

The equation (9) can be reduced to include only the refinement parameters for orbit, resulting in

$$[V] + [B_I][\Delta_I] = [\epsilon] \quad (23)$$

Using the above system of equation, the refinement parameters for orbit can then be computed by using least squares. The parameters are computed iteratively and hence the corrections to the initial approximate values are updated at each iteration.

Corrections to Viewing Geometry with RPCs

Correction parameters to scan and pixel directions are determined from ground control (Groddecki and Dial 2003) to express the object space to image space relationship.

The RPC-based space resection model equation for each GCP can be written as follows:

$$\text{Scan} = \Delta g + g(\phi, \lambda, h) \text{Pixel} = \Delta h + h(\phi, \lambda, h) \quad (24)$$

where,

Scan and Pixel are measured line and sample coordinates of the GCP with object space coordinates (ϕ, λ, h) , and Δg and Δh are the adjustable functions which model the difference between the measured and the nominal line and sample coordinates of GCP.

g and h are the denormalized RPC models for image i for given line and sample,

$$\begin{aligned} g(\phi, \lambda, h) &= L(\phi, \lambda, h) \cdot \text{SCAN_SCALE} \\ &+ \text{SCAN_OFF}h(\phi, \lambda, h) \\ &= S(\phi, \lambda, h) \cdot \text{PIXEL_SCALE} + \text{PIXEL_OFF} \end{aligned} \quad (25)$$

Δg and Δh are the adjustable functions

$$\Delta g = a_0 + a_S \cdot \text{Pixel} + a_L \cdot \text{Scan} + a_{SL} \cdot \text{Pixel} \cdot \text{Scan} + \dots$$

$$\Delta h = b_0 + b_S \cdot \text{Pixel} + b_L \cdot \text{Scan} + a_{SL} \cdot \text{Pixel} \cdot \text{Scan} + \dots$$

where, $a_0, a_S, a_L, \dots$ and $b_0, b_S, b_L, \dots$ are the image adjustment parameters for scan and pixel direction, respectively. Truncated polynomial model to linear model is as shown below

$$\Delta g = a_0 + a_S \cdot \text{Pixel} + a_L \cdot \text{Scan}$$

$$\Delta h = b_0 + b_S \cdot \text{Pixel} + b_L \cdot \text{Scan} \quad (26)$$

The adjustment parameters a_0, a_S, a_L and b_0, b_S, b_L are determined during the RPC-based space resection process.

3D Point Coordinates Computation

DEM with RSM The conjugate points (image point pair) are used in this process to generate 3D coordinates. The line connecting the perspective centers, object point, and corresponding image point from individual cameras are made to intersect in this process by space intersection models to derive the 3D ground coordinates.

Initial ground coordinates are computed by the intersection of rays from two cameras. The initial ground coordinates are iteratively refined by least squares method. For one conjugate point out of doublet or triplet of overlapping images, the system of equation would be

$$\begin{aligned} & \left(\frac{\partial E1}{\partial X} \right)^0 \Delta X + \left(\frac{\partial E1}{\partial Y} \right)^0 \Delta Y + \left(\frac{\partial E1}{\partial Z} \right)^0 \Delta Z \\ & = -E1^0 \quad \left(\frac{\partial E2}{\partial X} \right)^0 \Delta X + \left(\frac{\partial E2}{\partial Y} \right)^0 \Delta Y \\ & \quad + \left(\frac{\partial E2}{\partial Z} \right)^0 \Delta Z \\ & = -E2^0 \end{aligned} \quad (27)$$

where (X, Y, Z) are the ground coordinates of the conjugate point.

One ground point corresponding to the same feature in two overlapping images will give four equations which will be solved iteratively to obtain the final ground coordinates.

Equation (9) can be reduced to include only the refinement parameters for ground coordinates, resulting in

$$[V] + [B_O][\Delta_O] = [\epsilon] \quad (28)$$

DEM with RPM Separate sets of RPCs, corresponding to separate images, model the mappings between the ground coordinates and the image coordinates. The rational functions w.r.t scan and pixel direction can be given as

$$\begin{aligned}\text{Scan} &= p(\emptyset, \lambda, h) \\ \text{Pixel} &= r(\emptyset, \lambda, h)\end{aligned}\quad (29)$$

Where,

$$\begin{aligned}p(\emptyset, \lambda, h) &= \frac{N_{\text{scan}}(X, Y, Z)}{D_{\text{scan}}(X, Y, Z)} * \text{scan_scale} \\ &\quad + \text{scan_offset } r(\emptyset, \lambda, h) \\ &= \frac{N_{\text{pixel}}(X, Y, Z)}{D_{\text{pixel}}(X, Y, Z)} * \text{pixel_scale} + \text{pixel_offset}\end{aligned}\quad (30)$$

Here $(\emptyset, \lambda, h)$ are object-space coordinates of a point under consideration and (X, Y, Z) are the normalized ground coordinates of the same point which can be obtained from $(\emptyset, \lambda, h)$

$N_{\text{scan}}(X, Y, Z)$, $D_{\text{scan}}(X, Y, Z)$ are the numerator and denominator of the rational polynomial in scan direction, respectively

$N_{\text{pixel}}(X, Y, Z)$, $D_{\text{pixel}}(X, Y, Z)$ are the numerator and denominator of the rational polynomial in pixel direction, respectively

Corresponding linearized equations can be written as

$$\begin{aligned}\text{scan} &= p(\emptyset_0, \lambda_0, h_0) + \left[\frac{\partial p}{\partial z^T} \right]_{z=z_0} dz \\ \text{pixel} &= r(\emptyset_0, \lambda_0, h_0) + \left[\frac{\partial r}{\partial z^T} \right]_{z=z_0} dz\end{aligned}\quad (31)$$

Where,

$$z = [\emptyset \quad \lambda \quad h]^T$$

z_0 is the value of z at initial values of $(\emptyset, \lambda, h)$

In the case of two overlapping images, Eq. (31) can be written as

$$\begin{aligned}\text{Scan}_1 &= p_1(\emptyset, \lambda, h) \\ \text{Pixel}_1 &= r_1(\emptyset, \lambda, h) \\ \text{Scan}_2 &= p_2(\emptyset, \lambda, h) \\ \text{Pixel}_2 &= r_2(\emptyset, \lambda, h)\end{aligned}\quad (32)$$

Hence,

$$Adz = B\quad (33)$$

where,

$$A = \begin{bmatrix} \frac{\partial p_1}{\partial z^T} & \frac{\partial r_1}{\partial z^T} & \frac{\partial p_2}{\partial z^T} & \frac{\partial r_2}{\partial z^T} \end{bmatrix}_{z=z_0}^T$$

$$dz = [d\emptyset \quad d\lambda \quad dh]^T \text{ and}$$

$$B = \begin{bmatrix} \text{Scan}_1 \\ \text{Pixel}_1 \\ \text{Scan}_2 \\ \text{Pixel}_2 \end{bmatrix} - \begin{bmatrix} p_1(\emptyset_0, \lambda_0, h_0) \\ r_1(\emptyset_0, \lambda_0, h_0) \\ p_2(\emptyset_0, \lambda_0, h_0) \\ r_2(\emptyset_0, \lambda_0, h_0) \end{bmatrix}$$

Equation (33) is solved by least squares. It is solved iteratively by adding delta values to the initial values of the ground coordinates.

Photogrammetric Adjustment of Aerial Stereo Photo Blocks

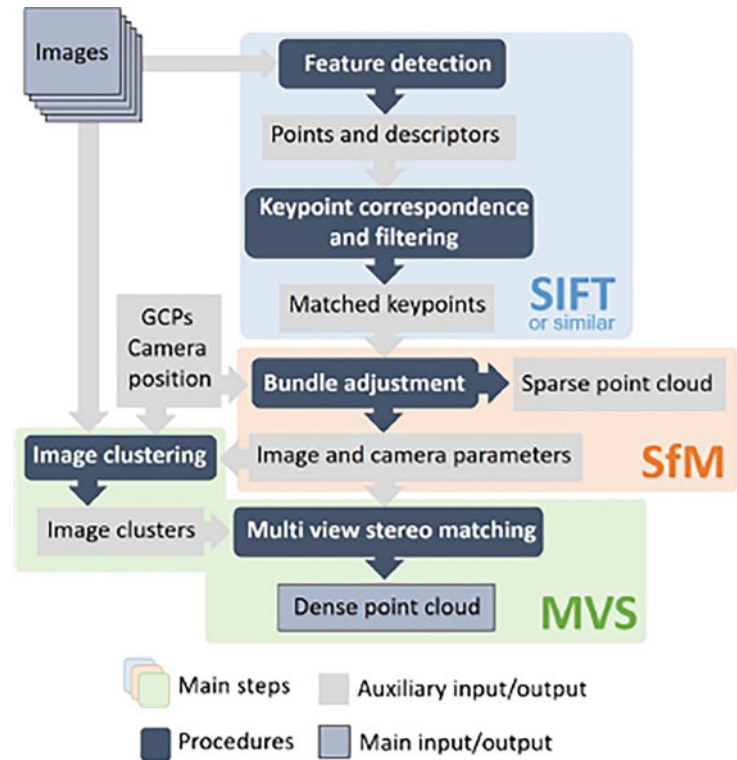
The images acquired with the aerial platforms are captured by frame cameras (analog or digital) and have large overlaps (of the order of 60%). Aerial campaigns are planned in such a way that sequential frame images have large overlaps as well as sidelaps (of the order of 40%). Each frame can be expressed as a bundle in the collinearity model. The independent models are adjusted using the block adjustment technique as described earlier in section “[Extracting Elevation/3D Coordinates from Stereo Imagery: A Generic Photogrammetric Data Adjustment Approach](#).” Control points/Tie points in the overlap and sidelap regions ensure the consistency between the adjacent images. Many other parameters can be included during the formulation of the original problem. Convergence criterion for each of the parameters is adjusted while forming the weight based on the need for its refinement. DEM is obtained as a product of photogrammetric adjustment of the blocks of images.

DEMs from Unmanned Aerial Vehicles (UAV) Using Structure from Motion (SfM) Approach

SfM is a photogrammetric technique where a set of two-dimensional images sequences are used to create three-dimensional structures (Iglhaut et al. 2019; Westoby et al. 2012). These images are typically acquired by the digital frame cameras on-board UAVs and are acquired as the platform moves with large overlaps between the images. The main advantage of this technique is that it can work with both calibrated and uncalibrated cameras. Both pose/orientation and 3D coordinates can be generated simultaneously. Typical workflow is as shown below (Fig. 2).

It starts with finding out common points between all the images. Since, it uses images with large overlaps, lot of

Digital Elevation Model,
Fig. 2 SfM workflow (Iglhaut
 et al. 2019)



common features are available. The task of feature detection is usually achieved by employing SIFT (described in section “[Scale Invariant Feature Transform \(SIFT\)](#)”). Once the match points are available, relative pose between them can be estimated using Essential matrix (calibrated camera) and Fundamental matrix (uncalibrated camera).

Essential Matrix: This is based on coplanarity condition which relates same feature being imaged with two different views as expressed in Eq. (34).

$$x'^T E x'' = 0 \quad (34)$$

where x' and x'' are the image points in two different views.

$$E = RS$$

R is a rotation matrix and S is skew symmetric matrix,

$$S = \begin{bmatrix} 0 & -T_z & T_y \\ T_z & 0 & -T_x \\ T_y & T_x & 0 \end{bmatrix}$$

The essential matrix is of rank 2 and its two non-singular values are equal. Set of homogeneous Equations (38) can be solved using 5-point algorithm. Using the essential matrix, relative pose can be determined by doing its singular value decomposition. The pose is used to compute the 3D coordinates. Similar process can be used to determine 3D

coordinates for uncalibrated camera by determining Fundamental matrix. The model developed on sparse point cloud between consecutive images, obtained using SIFT, is first refined using the ground control points. Dense points are then computed in overlapping images by employing multi-view stereo image matching. Refined model parameters obtained previously are used for the dense point clouds to get accurate ground coordinates. These 3D point clouds can be interpolated to generate DEMs at required intervals by using any of the suitable DEM interpolation techniques described in section “[DEM Interpolation](#).”

Image Matching

Image matching is one of the most important steps in DEM generation. In this process conjugate points are usually generated automatically between pair of overlapping images acquired at different times. The density and accuracy of the conjugate points dictates the accuracy of resultant DEM. Major difficulties encountered in image matching are due to variable time of imaging and scale changes due to difference in imaging geometry. Two robust techniques are Hierarchical Image Matching (HIM) technique and matching using Scale Invariant Feature Transform (SIFT).

Hierarchical Image Matching (HIM)

HIM is a popular technique used for generating dense conjugate points (Kannan and Singh 2011). It consists of Condensation, Interest Point identification, Local Mapping and Image Matching. Condensation generates the sub-sampled images from the full resolution sets, creating pyramid layers of the images with varying resolution. Interest Point Operator identifies the high contrast points on the sub-sampled reference images for every layer. At the lower most level of the pyramid that is, most coarser resolution, image matching is done directly by extracting a window area of 16×16 from the reference image around the interest points and correlation is performed on a search area window 32×32 of the other image around the probable location. The image matching results of the previous level are used to map the probable location of interest points in reference image to their probable counterpart in the other image by means of Local Mapping. Thereafter Image matching is done on the locally mapped points. The steps involved in image matching are described below.

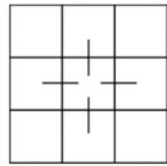
Image Condensation

Stereo images are acquired with different viewing geometry which includes different imaging times and view angles. This results in change in position of the same feature in both the images. This apparent displacement due to change in the position of feature is known as Parallax. Depending on viewing geometry and the height variations in the terrain, parallax may be observed in both x and y or in other words, across the track and along the track imaging directions. Undulations together with the stereo-imaging geometry of the satellite imaging system result in either the x-parallax or y-parallax

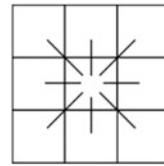
or both. During image condensation, the parallax magnitude is reduced by subsampling the images, usually by skipping the alternate pixels in across and along directions to have uniformity. This image condensation process results in formation of image pyramids at various subsampling levels. The position of the feature on the image is due to viewing geometry and the terrain undulation. The image condensation here attempts to reduce this displacement by means of skipping alternate pixel in both the directions for uniformity and generating a pyramid of subsampled images. Thus an image pyramid is generated by sub-sampling the image data. Each pyramid layer is referred to as Level. The point to be noted here is that though in each Level image is subsampled by skipping the alternate pixels and image breaks are created, HIM is capable of matching it accurately without significant loss of accuracy.

Interest Point Generation

Interest points are the points whose positions are well defined and can be determined accurately. Usually they are high contrast pixels, having high gray count variance with respect to their neighborhood pixels. Corners are a special case of interest point. Interest points are generated on one of the image only, in HIM which is treated as the reference image. One of the method for generation of interest point can be in the first step, screening of potential interest points by selecting the point whose gradients are above the predefined threshold with respect to the few neighborhood pixels and then in the second step filtering the screened candidates in first step using summation of gradients (IV) as weight for all points within the predefined window.



Comparison of gradients



Computation of Interest value gradients (IV)

(35)

where IV is the interest value, dg_i is the difference in the gray count of the interest point with respect to the i^{th} neighborhood pixel.

Local Mapping

The local mapping is used to estimate the location on the other image corresponding to every interest point on the reference image. A bivariate first- order polynomial is used for mapping

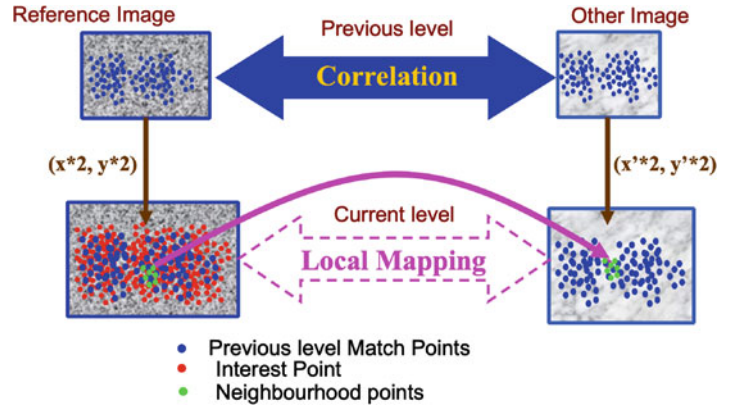
the interest points in the reference image to locations on other image.

$$X_r \rightarrow X_o = a_0 + a_1 * x + a_2 * y + a_3 * x' + a_4 * y'$$

$$Y_r \rightarrow Y_o = b_0 + b_1 * x + b_2 * y + b_3 * x' + b_4 * y' \quad (36)$$

where (X_r, Y_r) are the coordinates of the interest point in reference image, (X_o, Y_o) are the corresponding computed

Digital Elevation Model,
Fig. 3 Image matching process
 (Kannan and Singh 2011)



coordinates for the other image, (x, y) & (x', y') are the match point coordinates for the reference and other images respectively. In order to solve the above equations having 10 unknowns, the top 10 neighbourhood points are taken for computation. The process is also explained in figure below (Fig. 3).

Image Matching

Cross correlation technique is used to match between the reference and search window. A window Area (WA) is extracted from the reference image keeping (x, y) as the centre while, a search area (SA) window which is larger than WA for example of 32×32 pixels is extracted from other image keeping (x', y') as the centre. The positions of SA and WAs are taken from the local mapping file. The cross correlation coefficient between two windows search area s , and window area w of size $N \times N$ is defined as below:

$$R = \frac{\sum_{x=1}^N \sum_{y=1}^N s(x, y) w(x, y)}{\left[\sum_{x=1}^N \sum_{y=1}^N s(x, y)^2 \sum_{x=1}^N \sum_{y=1}^N w(x, y)^2 \right]^{\frac{1}{2}}} \quad (37)$$

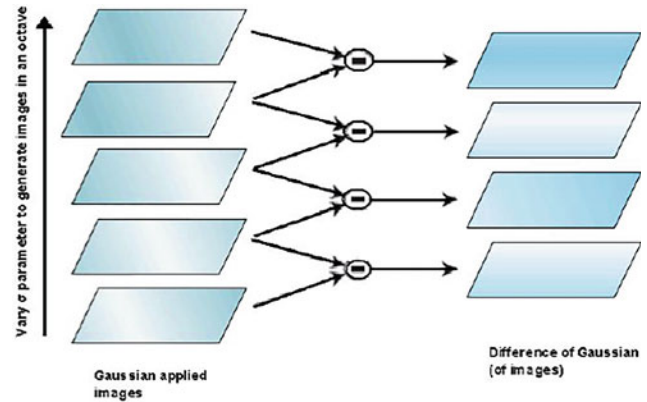
Where $s(x, y)$ and $w(x, y)$ are pixel values at location (x, y) of s and w respectively. The correlation coefficient varies between 0 and 1 and the closer the value of R is to 1, the more similar the two windows are likely to be.

Scale Invariant Feature Transform (SIFT)

Scale invariant feature transform (SIFT) was introduced by David Lowe (2004) as a robust means for feature detection in an image. In this approach, identified features are invariant to image scaling and rotation. The features identified using SIFT are highly distinctive, reducing false matches due to occlusions and noise. SIFT has been used to extract features robustly for refining orientation parameters of the satellite (Neethinathan et al. 2013). The steps involved in feature generation are as follows:

Scale-Space Extrema Detection

In this step, candidate interest points are identified which are invariant to scale: stable across different scales. Entire image



Digital Elevation Model, Fig. 4 Computation of difference of Gaussians (DOG) of images for an octave

is searched for these features over different scales. It is implemented using Difference of Gaussian (DOG). DOG is obtained by subtracting the adjacent Gaussian images in the octave as shown in Fig. 4.

Lowe proposed to use scale-space extrema in DOG function convolved with the image, $D(x, y, \sigma)$, as the feature candidates. They are computed from the difference of two nearby scales separated by a constant multiplicative factor k :

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \quad (38)$$

where $*$ is the convolution operator, $I(x, y)$ is input image, and

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \quad (39)$$

Keypoint Localization

In this step, the exact locations of the potential key points are determined by interpolation based on the stability measure. Brown and Lowe developed a method which uses Taylor's series expansion of the scale space function $D(x, y, \sigma)$ or $D(X)$ for fitting to get the interpolated location

of the maximum. The Taylor series expansion around the point X can be written as:

$$D(X) = D + \frac{\partial D^T}{\partial X} X + \frac{1}{2} X^T \frac{\partial^2 D}{\partial X^2 X} \quad (40)$$

The location of the extremum is determined by taking the derivative of this function with respect to X and setting it to zero. Apart from low contrast points, edges are also eliminated in this process, who have strong response in difference of Gaussian.

Orientation Assignment

In this process, based on local image gradients, orientations are assigned to each keypoint location. The keypoint descriptor is represented relative to this orientation. For assigning orientation to each key point, most prominent gradient orientation is identified using histogram. These steps provides invariance to orientation. The magnitude and orientation is computed by Eq. (41).

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2}$$

$$\theta(x, y) = \tan^{-1}((L(x, y+1) - L(x, y-1)) / (L(x+1, y) - L(x-1, y))) \quad (41)$$

Keypoint Descriptor

In order to generate a distinctive feature descriptor from the keypoint and the orientation information, histograms are computed around the 16×16 region around the keypoint in a block of 4×4 . The histograms computed for these 16 blocks using 8 orientation bins are then stacked and normalized to form a 128-dimensional vector which serves as the feature descriptor for that keypoint.

DEM Interpolation

Conjugate points generated as part of image matching are irregularly spaced points. Hence elevation values computed for these conjugate points will form the irregular DEM. These values need to be interpolated at equally spaced intervals to make it amenable for various end user applications. Based on the accuracy requirements, either local or global fitting techniques can be applied for DEM interpolation. Few possible DEM interpolation algorithms are given below.

Weighted Average

This algorithm is based on the principle that within a neighborhood of a point with unknown z value, closer points gets

more weightage compared to the farther points. Therefore, the weights are assigned to all neighborhood points as an inverse Euclidian distance function. So, z -value at an unknown point can be interpolated using equations

$$z = \sum_{j=1}^n w_j z_j \quad (42)$$

where $\sum w_j = 1$ and $w_j = d(i, j)^{-2} / \sum d(i, j)^{-2}$

$d(i, j)$ is Euclidean distance between the point in question i and neighborhood point j , n is the number of neighborhood points.

The neighbor can be defined simply as the closest points in a circle defined around the point under consideration. Voronoi diagrams can aid in defining the neighborhood.

Polynomial Fitting

In this method, a polynomial is used to model a surface within a neighborhood region of a point for which z -value is to be interpolated. Normally, first, second, or third degree polynomials are used to generate a locally valid surface. The general form of a third degree polynomial is:

$$z = a_1 + a_2x + a_3y + a_4x^2 + a_5xy + a_6y^2 + a_7x^3 + a_8x^2y + a_9xy^2 + a_{10}y^3 \quad (43)$$

This polynomial can represent a surface with two extrema. For estimating z -value, minimum ten neighboring points are required. The discussion regarding the choice of neighbors is similar to that in previous section.

Finite Element Method

This method is also known as B-Spline interpolation. The characteristics of this interpolation method are that this method provides series of patches resulting in a surface that has continuous first and second derivatives. It ensures continuity in (a) z -value, as the surface has no cliffs; (b) slope, as slopes do not change abruptly; and (c) curvature, as a minimum curvature is achieved. For further details, refer Inoque (1986).

Kriging

Kriging is a statistical technique of estimation based on weighted average method. But the shape and size of the neighborhood window is determined from semi-variogram, which is constructed from the sample points (DeMers 1997). For every pair of points, its distance d and squared difference in z -values is calculated and a variogram is drawn with x -axis as d and y -axis as squared difference of z -values. As d increases, the squared difference also increases but it stabilizes after d attains a certain threshold value. Once the variogram has been developed, it is used to estimate distance weights for interpolation. Weights are

selected in such a way that the estimates are unbiased and the variation between repeated estimates is minimum.

Kriging uses the idea of regionalized variable, which varies from place to place with some apparent continuity but cannot be modeled with a single smooth mathematical equation. Kriging is an exact interpolator in the sense that the values interpolated will coincide with the data points at those locations.

Gerchberg Algorithm

This algorithm considers a regular grid, with holes (where elevations are unknown) as input, and computes the elevation for the entire grid by an iterative error-reduction algorithm based on constraints specified in spatial and frequency (Fourier) domains. Here, a regular grid is created with required output grid size, for the area specified by the user in terms of corner coordinates in easting and northing. Subsequently, the iterative error reduction algorithm is executed. Spatial constraint specifies that the resulting interpolated surface passes through all input points after each iteration. The frequency domain constraints dictate the frequencies present in the interpolated band-limited surface. By this approach, the total number of frequency components is made to be equal or less than the total number of input points (Chamy et al. 1994).

DEM Formats

DEM are available in following formats

- (a) Regular grid format where each point represents the planimetric coordinates and elevation over that. The grid intervals are chosen based on the resolution of input data, which does not degrade its accuracy during the process of interpolation.
- (b) Raster form in GeoTiff format where each pixel intensity value represents the elevation. Usually provided in UTM projection and WGS84 datum.
- (c) Triangulated Irregular Network (TIN): TIN overcomes the data redundancy in raster form for uniform region. The irregular points are stored in the form of triangles where no triangle consists of points from the circumcircle formed by these points. The irregularly spaced elevation values form the vertices of the triangle. Hence, TIN is a vector representation of the surface by a set of non-overlapping triangles using irregularly spaced elevation values. It is based on the principle of Delaunay triangulation (Richbourg and Stone 1997).
- (d) Contour Map: It is a map generated using contour lines where each line joins the points at equal height value. Each contour line is separated by a constant difference known as contour interval.

Summary

This chapter presents the mathematical techniques, required to model and determine the digital elevation model (DEM) or the three-dimensional coordinates of points on earth's surface (and hence the elevation above a reference ellipsoid). A generic statistical photogrammetric adjustment approach where all parameters, viz. camera model, platform position and orientation, image point pairs, and ground control points, are assigned appropriate weights and adjusted simultaneously is described. Details of coordinate transformations in case of satellite photogrammetry are presented. The Scale invariant feature transform (SIFT) is introduced as a support algorithm for generating DEM from digital frame images with very large overlap acquired by Unmanned Aerial Vehicles. Brief description of image matching algorithms used for finding homologous points in digital imagery is given.

The image processing techniques developed in Computer Science are finding more and more inroads in getting precise and dense DEM for the planet Earth. On the other hand, the techniques for determination of position and orientation of imaging camera, well developed in the discipline of Photogrammetry, are being used in simultaneous localization and mapping (SLAM) of autonomous devices.

Bibliography

- Chamy MPT, Gopala Krishna B, Nandakumar R (1994) Implementation of an iterative error reduction algorithm for terrain interpolation. *Indian Cartographer* XIV: 135–141
- Dave V (2013) Block adjustment of Cartosat-1 full pass images using Rational Polynomial Model based approach. Post graduate M.Tech thesis under guidance of Sanjay Singh, SAC (ISRO)
- DeMers MN (1997) *Fundamentals of geographic information systems*. Wiley, New York, pp 255–286
- Florinsky IV (2016) *Digital terrain analysis in soil science and geology – digital elevation models*, 2nd edn. pp 77–108, Academic Press
- Gopala Krishna B, Lehner M (1994) Mathematical formulation and design for photogrammetric bundle adjustment software for optical sensors. Report prepared during first author's stay at DLR, Germany (Private Communication)
- Gopala Krishna B, Singh SK, Srinivasan TP (2010) Photogrammetric modelling and data processing: Chandrayaan-1 perspective (for Linear CCD Imaging Sensors), GEOMATICS
- Grodecki J (2001) IKONOS stereo feature extraction – RPC approach. In: *Proceedings of ASPRS 2001 conference*, St. Louis, Missouri
- Grodecki J, Dial G (2003) Block adjustment of high-resolution satellite images described by Rational Polynomials. *Photogramm Eng Remote Sens* 69(1):59–68
- Iglhaut J, Cabo C, Puliti S, Piermattei L, O'Connor J, Rosette J (2019) Structure from motion photogrammetry in forestry: a review. *Curr Forestry Rep* 5:155–168
- Inoque H (1986) A least squares smooth fitting for irregularly spaced data: finite element approach using the cubic B-spline basis. *Geophysics* 51:2051–2062

- Kannan KV, Singh SK (2011) Hierarchical image matching technique for DEM generation from Satellite Stereo Imagery. In: ICISC conference, ADIT, V V Nagar
- Lowe DG (2004) Distinctive image features from scale-invariant key points. *Int J Comput Vision* 60(2):91–110
- Neethinathan P, Singh S, Trivedi S, Srinivasan TP, Gopala Krishna B (2013) Use of SIFT feature extraction and image matching for automatic orthoimage generation for Cartosat-2, XXXIII INCA International Congress, Jodhpur
- Richbourg RF, Stone T (1997) Triangulated irregular network (TIN) representation quality as a function of source data resolution and polygon budget constraints. In: *Proceedings SPIE 3072, Integrating Photogrammetric Techniques with Scene Analysis and Machine Vision III*
- Sanjay KS, Naidu SD, Srinivasan TP, Gopala Krishna B, Srivastava PK (2008) Rational Polynomial Modeling for Cartosat-1 data. *ISPRS Beijing congress proceedings*, July 3–11 2008, vol XXXVII, Part B1, TC-1, pp 885–888
- Slama CC (1980) *Manual of photogrammetry*, 4th edn. The American Society of Photogrammetry
- Srinivasan TP, Singh SK, Suresh K, Alurkar MS, Gopala Krishna B (2013) Planetary models for Chandrayaan-1 data processing. In: 4th national conference on engineering applications of mathematics (NCEAM, 6–7th June 2013), Pune
- Srivastava PK, Krishna BG (2020) Developments in *digital mapping* from Indian Space Borne Line Scanner Imagery: IRS-1C to Cartosats. *J Indian Soc Remote Sens*
- Srivastava PK, Srinivasan TP, Gupta A, Singh SK, Nain JS, Amitabh, Prakash S, Kartikeyan B, Gopala Krishna B (2007) Recent advances in Cartosat-1 data processing, *ISPRS Hannover workshop 2007*, Hannover, Germany
- Tao CV, Ha YA (2001) Comprehensive study of the rational function model for photogrammetric processing. *Photogramm Eng Remote Sens* 67(12):1347–1357
- Westoby MJ, Brasington J, Glasser NF, Hambrey MJ, Reynolds JM (2012) “Structure from motion” photogrammetry: a low-cost, effective tool for geoscience applications. *Geomorphology*, 179:300–314

Digital Filters

Gabor Korvin

Earth Science Department, King Fahd University of
Petroleum and Minerals, Dhahran, Kingdom of Saudi Arabia

Definition

In geoscientific signal processing, a *digital filter* is a system (usually a *time-invariant, linear operator* realized by a computer program) that performs mathematical operations on a *sampled, discrete-time* signal to reduce certain aspects and/or enhance certain other aspects of that signal. The term “*digital*” is meant in contrast with the other major type of filters, the *analog filter*, which is used during signal-recording and is typically realized by an electronic circuit operating on *continuous-time analog signals*.

Introduction

Because of space limitations, we can only discuss *linear, single-channel, convolutional, time-invariant filters*. For other kinds of filters, see, for example, Keating and Pinet (2011) (*nonlinear filters*); Meskó (1966) (*multichannel filters*); the Appendix and Shanks (1967) (*for recursive filters*); Scheuer and Oldenburg 1988 (*time-varying filters*). For the new-comer to digital filtering, the classic texts (Robinson and Treitel 1964, 1980) are still highly recommended.

Single-Channel Convolutional Filtering in Time Domain, the z-Transform

This kind of filtering is realized by *convolution*, which is a special case of *cross-correlation*. The convolution of $x(t)$ with $y(t)$, denoted by $x(t) * y(t)$ is the non-normalized *cross-correlation function* between $x(t)$ and the *time-reversed version* of $y(t)$: $x(t) * y(t) = \int_{-\infty}^{+\infty} x(t - \tau)y(\tau)d\tau$ (Eq. 1). Using *sampled data* $x_n = x(n \cdot \Delta t)$ as *input*, $f_k = f(k \cdot \Delta t)$, $k = 0, 1, \dots, M$ as *filter*, and $y_n = y(n \cdot \Delta t)$ as *output*, and approximating the integral with a finite summation, digital filtering is executed as $y_n = \sum_{k=0}^M x_{n-k}f_k$ (Eq. 2).

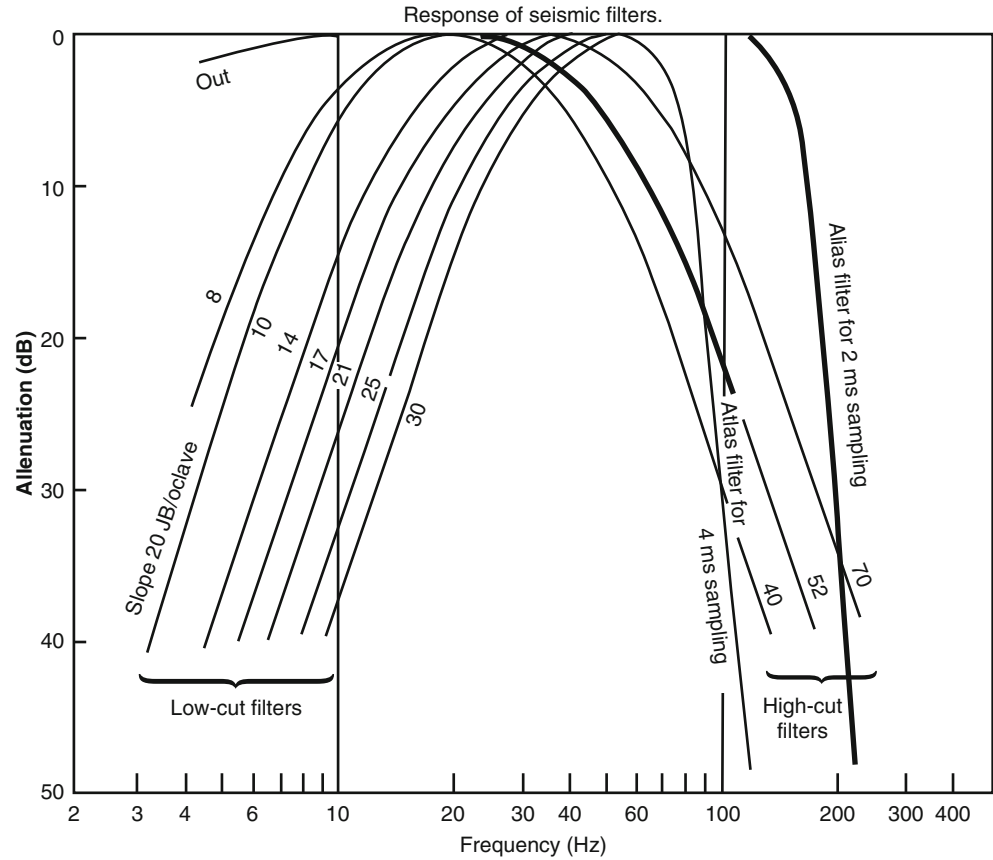
The *transfer characteristics* of filters (both digital and analog filters) are plotted in a particular form: The frequency is in logarithmic scale, and the attenuation of a given frequency is in linear scale, but in logarithmic (dB = decibel = $20 \cdot \log_{10} \text{Amplitude}_{\text{input}} / \text{Amplitude}_{\text{output}}$) units, as shown in Fig. 1. The slopes of the low- resp. high-cuts are specified in *dB/octave* where “octave” is the frequency range between f and $2f$ (where $f \rightarrow 0$ on the low-cut side, resp. $f \rightarrow \infty$ on the high-cut side). The “alias filter” (or “anti-alias filter”) is a special *analog high-cut filter* cutting out frequencies higher than $f_N (= f_{\text{Nyquist}}) = 1/(2 \cdot \Delta t)$ before sampling at equidistant discrete time instants $n \cdot \Delta t$.

Computationally there are many ways to execute filtering in time domain. Consider the simple example *Data* = {4, 2, 1, −1, 3, 5}; *Filter* = {2, 3, −2}; *Result* = ? We can directly use (a) the *convolution algorithm* (Eq.2); (b) the “moving paper-slip technique” (local *scalar product* of the data values with the filter values written on the moving slip, and note that the filter is written on the moving slip in reverse order).

$$\begin{array}{ccccccccc} & 4 & 2 & 1 & -1 & 3 & 5 & & \\ -2 & 3 & 2 & & & & & & \\ & 8 & & & & & & & \end{array}, \begin{array}{ccccccccc} & 4 & 2 & 1 & -1 & 3 & 5 & & \\ -2 & 3 & 2 & & & & & & \\ & 8 & 16 & & & & & & \end{array}, \dots;$$

(c) the “delay line technique: $2 \times \{4, 2, 1, -1, 3, 5, 0, 0, \dots\} + 3 \times \{0, 4, 2, 1, -1, 3, 5, \dots\} - 2 \times \{0, 0, 4, 2, 1, -1, 3, 5, 0, 0, \dots\} = \{8, 16, \dots\}$; but the most interesting and instructive way of filtering is (d) to form polynomials from the filter and the input data, and multiply them

Digital Filters, Fig. 1 Transfer functions for a set of *Ormsby-type bandpass filters* (Ormsby 1961), applied in *Seismic Data Processing*. These filters are specified by their four corner frequencies corresponding to the $-6, 0, -6$ dB points, respectively. (Recall that “ -6 dB” means signal attenuation by a factor 0.5.)



as polynomials of the variable z : $(4 + 2z + z^2 - z^3 + 3z^4 + 5z^5) \cdot (2 + 3z - 2z^2) = 8 + 16z + \dots$.

The transform $Z\{a_n\}: \{a_0, a_1, \dots, a_n\} \rightarrow a_0 + a_1z + \dots + a_nz^n$ (Eq.3) transforming sequences to polynomials is called *z-transform*. Its most important property relates to digital convolution: If $Z\{a_n\} = A(z)$; $Z\{b_n\} = B(z)$, then $Z[\{a_n\} * \{b_n\}] = A(z) \cdot B(z)$ (Eq. 4).

Filtering in Fourier Domain

In digital filtering, the use of the Fourier method is based on the *Convolution Theorem*. Denote the *Fourier transform* of $x(t)$ as function of frequency f by $F[x(t); f] = \int_{-\infty}^{\infty} x(t)e^{-i2\pi ft} dt = X(f)$ (Eq.5) where $i = (\text{in engineering applications sometimes denoted by } j) \sqrt{-1}$. Then (if all Fourier transforms exist) the *Convolution Theorem* states that, for all frequencies, $F[x(t) * y(t)] = X(f) \cdot Y(f)$ (Eq. 6). A particularly fast DFT (*Discrete Fourier Transform*, or *Digital Fourier Transform*) algorithm is the *Fast Fourier Transform* (FFT, discovered by Cooley and Tukey 1965). The FFT is only applicable if the data and the filter have the same length, and the number of data is a high integer power of 2. In practice, the FFT is computed in the following standard

steps: (a) Pad out the data with zeros to obtain a series of length $N^* > N$ where N^* is a power of 2; (b) *taper* (“truncate”) the series of data at its beginning and end; (c) do the same with the series of filter coefficients; (d) use FFT to get the Fourier Transforms of the padded and tapered data and filter. Filtering is then done by multiplying the two Fourier Transforms, and then inverse Fourier Transforming back to time domain. “Tapering” or “truncating” of the data is a critical part of the procedure, and many ways are available in most data-processing software packages to do it. Conventional truncation “windows” are (a) the *rectangular window* or *boxcar function*.

$$W(L) = \begin{cases} 1 & |L| \leq L_m \\ 0 & \text{otherwise} \end{cases} : \quad \text{(b) the Bartlett window}$$

$$W(L) = \begin{cases} 1 - \frac{|L|}{L_m} & |L| \leq L_m \\ 0 & \text{otherwise} \end{cases} : \quad \text{(c) the Hannning window}$$

$$W(L) = \begin{cases} 0.5 + 0.5 \cos \frac{\pi L}{L_m} & |L| \leq L_m \\ 0 & \text{otherwise} \end{cases} : \quad \text{(d) the Hamming window}$$

$$W(L) = \begin{cases} 0.54 + 0.46 \cos \frac{\pi L}{L_m} & |L| \leq L_m \\ 0 & \text{otherwise} \end{cases} : \quad \text{and (e) the Parzen Window}$$

(note that it has a polynomial equation, not trigonometric!)

$$W(L) = \begin{cases} 1 - 6\left(\frac{L}{L_m}\right)^2 + 6\left(\frac{L}{L_m}\right)^3 & |L| \leq 0.5L_m \\ 2\left[1 - \left(\frac{L}{L_m}\right)^3\right]^2 & 0.5L_m \leq |L| \leq L_m \\ 0 & |L| > L_m \end{cases} \quad \text{(Eqs. 7–11).}$$

Transfer Function of Digital Filters

If F is a time-invariant linear system (“filter”), its *transfer function* (TF) for the circular frequency $\omega = 2\pi f$ is defined as $T(\omega) = \frac{\text{output}}{\text{input}} = \frac{F\{e^{j\omega t}\}}{e^{j\omega t}}$ (Eq. 12).

We shall apply this rule, making use of the linearity of F , to compute the transfer function of the digital filter $F = a_0 + a_1 z + \dots + a_n z^n$. We find that the transfer function of the filter can be obtained by substituting $z = e^{-j\omega \cdot \Delta t}$ (where Δt is the sampling interval) into the z -transform of the filter; that is, we have $T(\omega) = a_0 + a_1 e^{-j\omega \cdot \Delta t} + \dots + a_n e^{-j\omega \cdot \Delta t}$ (Eq. 13). The TF (13) is complex-valued, its absolute value is $|T(\omega)| = \sqrt{\sum_{k=0}^n \sum_{l=0}^n \cos[(k-l)\omega \cdot \Delta t]}$ (Eq. 14, the expression below the square root sign is positive semi-definite!).

We usually design the frequency filters in the *Fourier Domain*, get the *weight function* (or *response function*) of the filter by *inverse Fourier Transform*, and finally sample and truncate the result for practical use.

The main problem encountered in digital filter design is caused by the *Gibbs Phenomenon* (an unavoidable oscillatory overshoot of the Fourier series of a function occurring at jumps and non-differentiable discontinuities). Because of this effect, the transfer function of the *sampled and truncated* filter response function will be not exactly equal to the desired and carefully designed theoretical transfer function. The minimization of this error is a main task of modern digital filter design (Meyerhoff 1968).

A filter with very good properties, widely used in *Seismic Data Processing*, is the *Ormsby Filter* (Ormsby 1961). Its transfer function has a trapezoidal shape. In the simplest case, the Ormsby LP (*Low Pass*) filter’s transfer function is given mathematically as

$$H_1(\omega) = \begin{cases} 1 & \text{for } |\omega| \leq \Omega_1 \\ 0 & \text{for } |\omega| > \Omega_2 \\ (\omega + \Omega_2)/\Delta\Omega & \text{for } -\Omega_2 \leq \omega < -\Omega_1 \\ (\Omega_2 - \omega)/\Delta\Omega & \text{for } \Omega_1 \leq \omega \leq \Omega_2 \end{cases}$$

(Eq. 15a, $\Delta\Omega = \Omega_2 - \Omega_1$).

Inverse Fourier Transform and some simple mathematics yield this filter’s response in time-domain:

$$h_1(t) = \frac{1}{\pi t} \sin \frac{(\Omega_1 + \Omega_2)t}{2} \times \frac{1}{\pi t \cdot \Delta\Omega} \sin \frac{t \cdot \Delta\Omega}{2} = \frac{\cos \Omega_1 t - \cos \Omega_2 t}{2\pi^2 t^2 \cdot \Delta\Omega}$$

(Eq. 15b, for $t = 0$ the expression should be computed by l’Hôpital’s rule). As seen in Fig. 1, Ormsby BP (*bandpass*) filters are specified by four numbers, namely the four “corner frequencies” f_1, f_2, f_3, f_4 corresponding to the $-6, 0, 0, -6$ dB points, respectively. (Recall that -6 dB corresponds to an amplitude attenuation by a factor 0.5.)

Time Varying Filtering (TVF, see Scheuer and Oldenburg 1988). Different BP filters emphasize different phenomena and patterns on the observed record. In many cases we need to apply different BP filters to different parts (sections) of the record. First we make a *filter analysis* to see which is the best

filter for a given section. We cannot introduce *abrupt changes* between two consecutive filters, because this would create an artificial, geologically meaningless, false “boundary.” Rather, we prefer to make slow, *gradual changes* by applying a *ramp* of reasonable length centered at the time instants of filter change.

Single-Channel Optimal Filters

Basic equation of optimal filtering: Suppose we have a stationary time-series $x = \{\dots, x_{n-1}, x_n, x_{n+1}, \dots\}$ and we would like to transform it as *closely as possible* (in the $\text{RMS} = \text{Root Mean Square error norm}$) into some *desired* time series $d = \{\dots, d_{n-1}, d_n, d_{n+1}, \dots\}$ by means of *convolving it with a filter* $= \{\sigma_0, \sigma_1, \dots, \sigma_M\}$. The problem of finding the *optimal filter* was solved by Norbert Wiener in the 1940s for *smoothing, interpolating, and predicting stationary time series* (Levinson 1947). If the input $x = \{\dots, x_{n-1}, x_n, x_{n+1}, \dots\}$ and the desired output $d = \{\dots, d_{n-1}, d_n, d_{n+1}, \dots\}$ are *stationary random sequences*, then the coefficients $\{\sigma_0, \sigma_1, \dots, \sigma_M\}$ of the optimal filter can be found from the linear system of equations:

$$\begin{pmatrix} R(0) & R(1) & R(2) & \dots & R(M) \\ R(1) & R(0) & R(1) & \dots & R(M-1) \\ R(2) & R(1) & R(0) & \dots & R(M-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(M) & R(M-1) & R(M-2) & \dots & R(0) \end{pmatrix} \begin{pmatrix} \sigma_0 \\ \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_M \end{pmatrix} = \begin{pmatrix} R_{dx}(0) \\ R_{dx}(1) \\ R_{dx}(2) \\ \vdots \\ R_{dx}(M) \end{pmatrix}$$

(Eq. 16, all arguments of the correlation functions are in units of the sampling time Δt), where $R(k) = R_{xx}(k)$ is the *ACF* (*Autocorrelation Function*) of $\{x_n\}$, $R_{dx}(k)$ is the *cross-correlation function* between the desired output $\{d_n\}$ and the input $\{x_n\}$ at lag k . This is a very large system of equations (in general the matrix has some 50×50 elements or more). Because of the special symmetries of the *Autocorrelation Matrix* (it is a so-called *diagonal-constant “Toeplitz matrix”*) the system of equations can be solved very efficiently by using Levinson’s algorithm (Levinson 1947), but it should be kept in mind that the system is strongly *ill-conditioned*.

A frequently occurring case when we use optimal filtering is the *Prediction of Stationary Time Series* (e.g., the so-called *predictive deconvolution* in *Seismic Data Processing*, see Robinson and Treitel 1980). Suppose we observe a random time series $x(t)$ and want to optimally *predict* (in the sense of minimal *Root Mean Square error*) its future value a time τ ahead, that is $x(t + \tau)$. If a convolutional filter of length $M + 1$ is used for this purpose, that is when we predict for a time τ ahead in the form $x(t + \tau) \approx \sum_{k=0}^M \sigma_k x(t - k)$ (Eq. 17), the optimum prediction filter coefficients should satisfy the equation:

$$\begin{pmatrix} R(0) & R(1) & R(2) & \cdots & R(M) \\ R(1) & R(0) & R(1) & \cdots & R(M-1) \\ R(2) & R(1) & R(0) & \cdots & R(M-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(M) & R(M-1) & R(M-2) & \cdots & R(0) \end{pmatrix} \begin{pmatrix} \sigma_0 \\ \sigma_1 \\ \rho_2 \\ \vdots \\ \sigma_M \end{pmatrix} = \begin{pmatrix} R(k) \\ R(k+1) \\ R(k+2) \\ \vdots \\ R(k+M) \end{pmatrix} \quad (\text{Eq. 18}).$$

Appendix: Recursive Filtering (Shanks, 1967)

In *convolutional filtering*, the output $\{y_n\}$ only depends on the input $\{x_n\}$, $y_n = \sum_{k=0}^M x_{n-k} f_k$ (Eq. 2). In a more general filtering scheme, the output might also depend on some of its own previous values. For example if y satisfies a *difference equation* of the form

$\sum_{k=0}^N a_k y_{n-k} = \sum_{k=0}^M b_k x_{n-k}$ (Eq. 19) where $a_0 \neq 0$, this equation can be changed to the equation of *recursive filtering* (Shanks 1967) $y_n = \sum_{k=1}^M \alpha_k y_{n-k} - \sum_{k=0}^N \beta_k x_{n-k}$ (Eq. 20). The *transfer function* of such a recursive system is described by

$$R(\omega) = \frac{\sum_{k=0}^M \beta_k \exp[-ik\omega \cdot \Delta t]}{1 + \sum_{k=1}^N \alpha_k \exp[-ik\omega \cdot \Delta t]} \quad (\text{Eq. 21}).$$

This form of the transfer function is especially suited for the *zero- and pole design* of digital filters to achieve favorable filtering properties.

Concluding Remarks on the Use of Digital Filtering

We briefly discussed digital filters used in Geosciences, concentrating on *single-channel convolutional filters*. Finally, we want to add a few words of warning concerning their use. It has been proved in *Information Theory* that the application of any kind of linear systems, such as a digital filter, to an observed signal inevitably causes *loss of information* about the signal, that is it would lead to a less precise *geologic interpretation*. If a filtered record is used as input to some *inverse problem* (inverse problems are notoriously *ill-conditioned*, and sensitive to small changes in their input), one might even “detect” spurious geologic anomalies. Anybody who uses filtered data for geologic interpretation should carefully check the reality of the results obtained against the original, unfiltered data.

Cross-References

- Autocorrelation
- Computational Geoscience
- Fast Fourier Transform
- Geocomputing
- Geomathematics

- Geoscience Signal Extraction
- Least Mean Squares
- Signal Analysis
- Signal Processing in Geosciences
- Smoothing Filter
- Spectral Analysis
- Time Series Analysis
- Time Series Analysis in the Geosciences
- Z-transform

References

- Cooley JW, Tukey JW (1965) An algorithm for the machine calculation of complex Fourier series. *Math Comput* 19:297–301
- Keating P, Pinet N (2011) Use of non-linear filtering for the regional-residual separation of potential field data. *J Appl Geophys* 73(4): 315–322
- Levinson N (1947) The Wiener RMS error criterion in filter design and prediction. *J Math Phys* 25:261–278
- Mesko CA (1966) Two-dimensional filtering and the second derivative method. *Geophysics* 31(3):606–617
- Meyerhoff HJ (1968) Realization of sharp cut-off frequency characteristics on digital computers, part 1. *Geophys Prospect* 16:208–219
- Ormsby JFA (1961) Design of Numerical Filters with applications to missile data processing. *J Assoc Comput Mach* 8:440–466
- Robinson EA, Treitel S (1964) Principles of digital filtering. *Geophysics* 29(3):395–404
- Robinson EA, Treitel S (1980) *Geophysical signal analysis*. Prentice-Hall, New Jersey
- Scheuer TE, Oldenburg DW (1988) Aspects of time-variant filtering. *Geophysics* 53(11):1399–1409
- Shanks JL (1967) Recursion filters for digital processing. *Geophysics* 32(1):33–51

Digital Geological Mapping

Chengbin Wang¹ and Xiaogang Ma²

¹State Key Laboratory of Geological Processes and Mineral Resources & School of Earth Resources, China University of Geosciences, Wuhan, China

²Department of Computer Science, University of Idaho, Moscow, ID, USA

Definition

Digital geological mapping is to use the portable digital equipment to collect first-hand geological features information by imitating the process of field geological mapping of geologists. The collected information is further processed to make digital geological maps and databases for sharing.

The History of Digital Geological Mapping

Initiatives of digital geological mapping started in the 1980s. Some geological codes and principles were designed to standardize field data collection for producing digital geological map and database (Wright and Stewart 1990). Along with the fast development of portable equipment, GPS and GIS, computer scientists and geologists tried to develop the software to support the field data collection and replace the traditional paper-based geological mapping method. Fieldlog, a software system of digital geological mapping developed by the Geological Survey of Canada, was released in 1991 (Brodaric 1992, 2004). Subsequently, a series of digital geological mapping software systems were developed and tested in the geological mapping in the Europe, Australian, China, and North America, for instance, ArcPad-ArcGIS system (Pavlis et al. 2010), GSMCAD FieldPad (Williams et al. 1996), MapIT (Blewett and Hazell 1997; Brown and Sprinkel 2008; De Donatis 2007), GeoMapper (PenMap) (Brimhall and Vanegas 2001), FieldBook (Briner et al. 1999), Natmap (Broome et al. 1993), SIGMA system (Jordan and Napier 2016), GeoSurvey (Liu et al. 2001), FieldMove (Lundmark et al. 2020; Senger and Nordmo 2020), StraboSpot (Walker et al. 2019), and DGSS-RGMAP (Li 2011; Li et al. 2008a).

Digital Geological Mapping System

The realization of an efficient digital geological mapping system depends on the development of hardware and software. The digital geological mapping system can be divided into three modules: field data acquisition module, field data cleaning module, and mapping and database module. The field data acquisition module is a portable equipment (e.g., personal digital assistant (PDA), tablet PC, and smart phone) with data acquisition software, camera, GPS, and digital compass. Geologists use the field data acquisition module to collect as much geological feature information as possible, such as attitude, geological boundary, contact relationship of geological bodies, photo, geological sketch, and text description.

The functions of field data cleaning module are to collate geological data collected in the field. Due to the complexity of the field environment, the first-hand geometry and attributes of geological features are semifinished products with irregular shapes and missing or incorrect attribute information. These shortcomings need to be addressed using the field data cleaning module. This module is a special GIS platform for indoor use, which can be an extension tool of secondary development based on the existing GIS platform or an independent GIS tool. The role of mapping and database module is to integrate the cleaned field data to produce the geological map and geological dataset for sharing in a GIS environment.

In the three modules, the biggest challenge is to design a data model in the field data acquisition module for easy operation by geologists in the field and further processing indoor to produce the geological map and database. It is necessary to reach a balance between the complex field environment, limited power supply and screen size, and easy post-processing requirements. Keep it simple stupid is a critical rule for the design of the data model. The point and polyline are preference for data input in the field. Recently, China Geological Survey provided a solution of PRB data framework to imitate the process of data collection in the geological mapping (Li et al. 2008b). As a whole, the PRB model constitutes a data framework for geological mapping. Other information such as sample, attitude, fossil, photo, and sketch can be added nearby and attached to the corresponding PRB element using its ID number.

The Process of Digital Geological Mapping

The workflow of digital geological mapping can be divided into the following steps:

1. Prepare the reference base map of remote sensing imagery, topographic map, and small-scale geological map for geological data collection in the field.
2. Design a geological survey route based on the base map.
3. Geological data collection in the field using portable equipment (e.g., smartphone, PDA, tablet PC, iPad).
4. Cleaning and modification of first-hand geological data indoors.
5. Integrate geological survey data and build a raw data database.
6. Draw the geological map and built a geological database for sharing.

Summary

In traditional geological mapping, a toolbox, including pencil, compass, camera, hammer, magnifying glass, notebook, GPS receiver, and sketch paper is used to collect field geological data. The small size portable equipment replaces the roles of paper, pencil, compass, camera, notebook, and GPS receiver and provides an easy and integrated way to collect geological feature information. Field data collection is not the end of digital geological mapping. The digital geological mapping system also revolutionizes the mapping workflow and can be used to produce digital geological maps and databases for sharing and further use. The development of big data, artificial intelligence, and cloud computation also promotes the transformation from digital geological mapping to smart geological survey, which in turn will lead to many research topics of interest in mathematical geoscience and geoinformatics.

Cross-References

► Database Management System

Bibliography

- Blewett R, Hazell M (1997) User's guide to MapPad and AGSO FIELDPAD for the Apple Newton Palmtop Computer. Australian Geological Survey Organisation
- Brimhall GH, Vanegas A (2001) Removing science workflow barriers to adoption of digital geological mapping by using the GeoMapper Universal Program and visual user interface. US Geological Survey Open File Report:01-223
- Briner AP, Kronenberg H, Mazurek M, Horn H, Engi M, Peters T (1999) FieldBook and GeoDatabase: tools for field data acquisition and analysis. *Comput Geosci* 25(10):1101–1111
- Brodaric B (1992) GSC FIELDLOG v2.83: Geological Survey of Canada, internal publication, p 95
- Brodaric B (2004) The design of GSC FieldLog: ontology-based software for computer aided geological field mapping. *Comput Geosci* 30(1):5–20
- Broome J, Brodaric B, Viljoen D, Baril D (1993) The NATMAP digital geoscience data-management system. *Comput Geosci* 19(10):1501–1516
- Brown KD, Sprinkel DA (2008) Geologic field mapping using a rugged tablet computer. In: Digital mapping techniques '07—workshop proceedings: US Geological Survey open-file report. *Citeseer*, pp 53–58
- De Donatis M (2007) Earth and environmental sciences: field classes with GIS/GPS and tablet PC. In: First international workshop on pen-based learning technologies (PLT 2007). *IEEE*, pp 1–6
- Jordan CJ, Napier B (2016) Developing digital fieldwork technologies at the British Geological Survey. *Geol Soc Lond, Spec Publ* 436(1):219–229
- Li C (2011) Guide of digital geological survey system. Geological Publishing House, Beijing. [In Chinese]
- Li C, Yang D, Li F, Liu C, Liu X, Yu Q, Lv X (2008a) Basic framework of the digital geological survey system and application of its key technology. *Geol Bull China* 27(7):923–944. [In Chinese with English abstract]
- Li C, Zhang K, Yu Q, Yang D, Zhu Y, Ge M (2008b) Theoretical framework of PRB granularity of digital geological mapping. *Geol Bull China* 27(7):945–955. [In Chinese with English abstract]
- Liu G, Wang X, Zhao W, Wu C, Ouyang J (2001) The development of computer-aided regional geological mapping system and field practice teaching reform for base class. *Chin Geol Educ* 03:32–35. [in Chinese]
- Lundmark AM, Augland LE, Jørgensen SV (2020) Digital fieldwork with fieldmove-how do digital tools influence geoscience students' learning experience in the field? *J Geogr High Educ* 44:427–440. <https://doi.org/10.1080/03098265.2020.1712685>
- NACSN (1983) North American stratigraphic code. *Am Assoc Pet Geol Bull* 67:841–875
- Pavlis TL, Langford R, Hurtado J, Serpa L (2010) Computer-based data acquisition and visualization systems in field geology: results from 12 years of experimentation and future potential. *Geosphere* 6(3):275–294
- Senger K, Nordmo I (2020) Using digital field notebooks in geoscientific learning in polar environments. *J Geosci Educ* 69:166–177
- Walker JD, Tikoff B, Newman J, Clark R, Ash J, Good J, Bunse EG, Möller A, Kahn M, Williams RT (2019) StraboSpot data system for structural geology. *Geosphere* 15(2):533–547
- Williams VS, Selner GI, Taylor RB (1996) GSMCAD a new computer program that combines the functions of the GSMAP and GSMEDIT

programs and is compatible with Microsoft Windows and ARC/INFO. US Geological Survey

Wright BE, Stewart DB (1990) Digitization of a geologic map for the Quebec-Maine-Gulf of Maine global geoscience transect, vol 1041–1045. US Government Printing Office

Digital Twins of the Earth

Stefano Nativi and Max Craglia
European Commission –Joint Research Centre (JRC), Ispra, Italy

Synonyms

Earth Digital Twins

Other terms sometimes used with a similar meaning are: City Twins; Digital Earth; Local Digital Twins

Definition

A Digital Twin of the Earth is the digital replica of an Earth system component, structure, process, or phenomenon obtained by merging digital modelling (notably, learning-based models) and real-world observational continuity – that is, natural and societal sensing data streams.

A Digital Twin of the Earth continuously learns and updates itself and must be seen as a living digital simulation model that modifies and changes itself as its physical counterpart changes.

Essential Concepts and Components

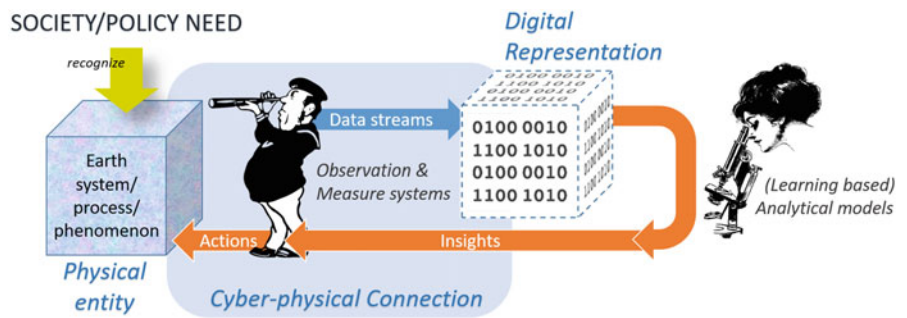
Digital Twin (DT) paradigm and systems have been around for decades in the industrial manufacturing processes. However, it is only with the recent advent of transformative digital technologies and the datafication of society that the DTs paradigm has been more and more applied in the geoscience domain. Empowered by the Internet of Things and AI technologies, the DT model provides the most advanced pattern to generate that intelligence required by decision makers, through the development of cyber-physical systems (Jacoby and Usländer 2020).

In geoscience, DTs of the Earth aim to effectively simulate and predict the behavior of key natural and societal systems/

Disclaimer: The views expressed are purely those of the authors and may not in any circumstances be regarded as stating an official position of the European Commission.

Digital Twins of the Earth,

Fig. 1 The cyber-physical model characterizing a Digital Twin of the Earth



processes/phenomena by leveraging the huge amount of data generated and shared daily on the network, along with the advancement in AI technology. For scientific experts and policy makers, DTs of the Earth are extremely advanced instruments providing a digital replica of important natural and social phenomena and processes (characterizing our planet), the understanding of which is essential to realize a “smart” and more sustainable society. They can be therefore important instruments contributing to implement the UN Sustainable Development Goals (SDGs) agenda and the European Green Deal transition.

By connecting the physical and the virtual worlds, data is transmitted seamlessly from one to another allowing the virtual entity to exist simultaneously and interact with the physical entity. On the other hand, the DT concept permits to decouple the digital replica from its physical entity, making it easier to change it and run fast and safe simulations. Utilizing advanced data-driven analytical procedures, it is possible to generate those insights that could not be carried out using the traditional observation models.

As depicted in Fig. 1, a DT of the Earth includes three main components: the *Physical Entity*, its *Digital Representation*, and the *Connection* channel between the two (Nativi et al. 2020). The main interaction features characterizing a DT are (Jones et al. 2020; ISO TC184 2020): interoperability, information model, data exchange, administration, synchronization, and push mode (i.e., publish and subscribe model). Therefore, the cyber-physical model characterizing a DT of the Earth is cyclical and commonly distinguishes the following set of phases and related activities:

1. Policy operationalization phase: to recognize the Earth entity (i.e., system/process/phenomenon) important to address a given societal or policy need
2. Observation and datafication phase: to continuously observe the acknowledged Earth entity, generating one or more data streams
3. Modelling phase: to utilize the data stream(s) for training and/or ingesting the digital representation (model) of the observed physical entity

4. Analytical phase: to utilize the digital model for generating simulations and/or predictions
5. Actionable intelligence phase: to interpret the generated insights and act-back on the observed Earth entity

DT Versus Digital Model

It is useful to distinguish DTs from the “digital model” concept that is often used with an equal meaning. The two concepts implement different levels of integration between the physical and digital entities (Kritzinger et al. 2018). In particular, a digital model does not implement any form of automated data exchange between the physical and the digital entities (e.g., in Fig. 1, the “Data stream” connection channel is off-line).

Some Significant Research and Innovation Applications

According to a recent survey conducted by the Joint Research Centre (JRC) of the European Commission for the EU Destination Earth initiative (Nativi et al. 2020), the DT of the Earth paradigm has been utilized in several projects and initiatives. The survey collected information from online publications, EU research and innovation frameworks, international initiatives, and a set of the elicited experts – for example, GEO (the intergovernmental Group on Earth Observation) (<https://earthobservations.org/>) experts. Table 1 is generated from this survey, reporting some significant research and innovation applications (categorized by area) that build on the concept of Digital Twins of the Earth. Naturally, the list of projects and initiatives is far from being exhaustive, its aim is to provide an idea of the application domains diversity and the lively activity concerning Digital Twins of the Earth. As shown, the earth system modelling and smart cities are the two most prominent domains.

Digital Twins of the Earth, Table 1 Some relevant applications of the DT of the Earth concept

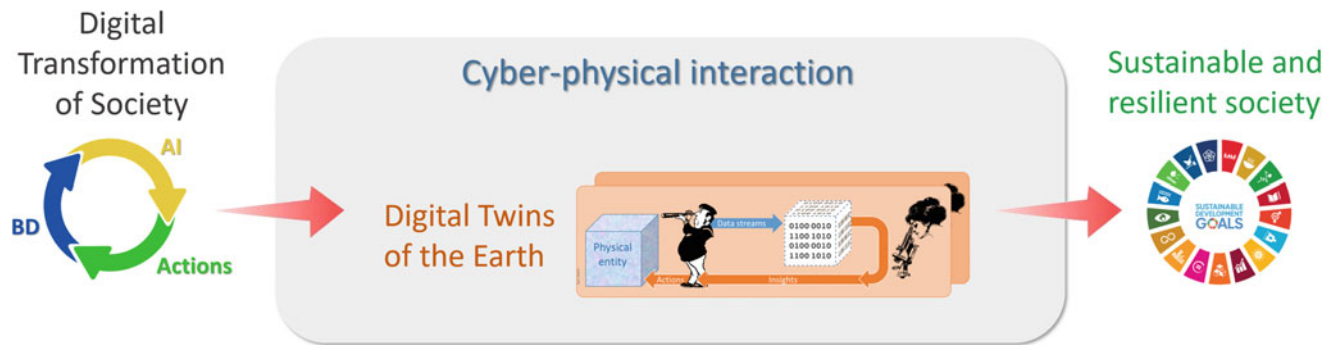
Main research and innovation areas	Examples of significant activities
Earth System Modeling/Digital Earth	EU H2020 CRESCENDO (Progressing on the next generation of European Earth System models) project DATA TERRA (French project) Advanced Earth System Modelling Capacity project (ESM) (German project) Digital Earth: Towards SMART Monitoring and Integrated Data Exploration of the Earth System - Implementing the Data Science Paradigm (German initiative) CMIP Phase 6 (International initiative) OneGeology 4.0 (International initiative) Deep Time Digital Earth (DDE) international initiative Space Climate Observatory (SCO) International Initiatives NSW Digital Twin (Australian initiative) USA NSF EarthCube initiative Descartes Labs: Digital Twin of planet Earth (USA project) Destination Earth initiative of the European Commission International Society for Digital Earth (ISDE)
3D Imaging	AI4GEO (French project) 3DExperienceCity (French and Singapore project) CO3D (French project)
Water and Drought monitoring	SWOT downstream program (French initiative) Drought Watch (ESA funded project)
Forest monitoring	Forest Inventory Program in Finland (Finland and Swedish project)
Ecosystems monitoring	ECOTWIN: Towards the definition of digital twins of specific ecosystems (Italian initiative) Nosvillesvertes (French initiative)
Maritime simulation and monitoring	Kongsberg Marine data streams capacity (SeaDataNet, EMODnet, and Jerico-S3 projects) (EU funded initiative)
Polar region monitoring	EU H2020 ExtremeEarth project
Extreme natural phenomena monitoring	Modular Observation Solutions for Earth Systems (MOSES) (German project) EU H2020 ExtremeEarth project Destination Earth initiative of the European Commission
Food Security monitoring	EU H2020 ExtremeEarth project
Pollution monitoring	Knowledge Hub to analyze and simulate organic pollution (Italian initiative)
Smart cities/City Twins	EU H2020 DUET (Digital Urban European Twins) project City of Zurich (Swiss project) 3DExperienCity (Virtual Singapore and Virtual Rennes) Helsinki Digital Twin (Finland project) Digital Twin Cadaster – Victoria (Australian project) 3DExperienceCity (French project) PortForward: the Digital Twin of the port of Rotterdam (Dutch project) Open Mobility Foundation (OMF) (USA initiative) Digital Built Britain (United Kingdom initiative) Cambridge initiative (United Kingdom initiative) Many other smart city projects all around the world (e.g., Antwerp, Newcastle, Boston, Pasadena, Portland, Dubai, Jaipur, Yingtan, Amaravati, Waterfront Toronto)
Smart Transportation/Infrastructures	ZeroGravity UrbanAI (Finland project) PortForward: the Digital Twin of the port of Rotterdam (Dutch project) Open Mobility Foundation (OMF) (USA initiative) Cambridge initiative (United Kingdom initiative) EU H2020 LEAD (Digital Twins for low emission last mile logistics) project
Smart buildings	EU H2020 SPHERE project
Smart Energy	EU H2020 SPHERE project Kongsberg (Norway project)
Climate Change adaptation strategies in urban areas	LIFE-IP AdaptInGR (EU and Greek funded project)

Current Investigations, Controversies, and Gaps in Current Knowledge

The DT of the Earth model represents a key instrument for developing the Digital Earth vision (ISDE Council 2012). For this reason, the two terms quite often are used

interchangeably. In addition, it constitutes a significant archetype and reference framework for introducing a new scientific discipline (a subdiscipline of Data science) that was introduced by Guo et al. (2020) and called Big Earth Data science.

To embrace the philosophy characterizing the DTs of the Earth, it is important to adopt a truly distributed and digital



Digital Twins of the Earth, Fig. 2 Digital Twins as a key instrument to leverage digital transformation of society for its sustainability and resilience

approach (see the “Digital Ecosystem”) in which data and models are accessible online and are interoperable in a computing and storage environment that provides scalability for the different applications – for example, via edge, containerization, and cloud computing. For the modelling and analytical software, that requires to move progressively from a tool-oriented approach (where access to models is provided via a software tool requiring human interaction) to a service-oriented approach (where machine-to-machine automatically interact through network services and APIs), and ultimately resource-oriented approach (where the code implementing the model is moved and run where the processed big data and the required computing scalability are located) (Nativi et al. 2021). Tao and Qi (2019) argue that the Geosciences academia and research community need to investigate further the following issues:

- Improved modeling techniques -further focusing on data-driven analytical techniques, in addition to the more traditional process-driven ones
- (Big) data streams optimization and their interoperability with modelling platforms – for example, by unifying data and model standards
- Design and implementation of innovative processes to implement and manage DTs of the Earth systems – that is, the DT of the Earth reference framework engineering

According to the scale (and type) of the Earth entity replicated digitally, a given DT can be defined as global, national, regional, or local. Smart cities (aka urban DTs) represent a popular example of local DT (Deng et al. 2021).

In 2019, ISO JTC1 (Information Technology) created an Advisory Group to recognize the standardization landscape, the possible existing gaps, and eventually recommend the creation of Systems Integration entity in the form of a new Subcommittee on Digital Twins.

Summary or Conclusions

The digital transformation of society underpins the development of a cyber-physical world that could prove essential in making human development sustainable and thus save the planet. Digital Twins of the Earth are instrumental in connecting the physical and virtual realms and in generating the knowledge required to minimize the impact of our society on the Earth – as depicted in Fig. 2. The DT of Earth systems represent Earth entities (components, processes, or phenomena) to capture their behavior over time, allow to run simulations, and generate actionable intelligence. For these reasons, there are already many research and innovation projects worldwide as well as initiatives applying the DT of Earth paradigm. The DTs of the Earth paradigm differ from the more traditional digital models because of the full synchronization of the digital representation with the Earth entity. Therefore, a DT of the Earth is a living digital simulation model that modifies and changes itself as its Earth counterpart changes. This important characteristic raises several challenges to the scientific and engineering communities.

Cross-References

► [Geosciences Digital Ecosystems](#)

Bibliography

- Deng T, Zhang K, Zuo-Jun S (2021) A systematic review of a digital twin city: a new pattern of urban governance toward smart cities. *J Manage Sci Eng*. <https://doi.org/10.1016/j.jmse.2021.03.003>
- Guo H, Nativi S, Liang D, Craglia M, Wang L, Schade S, Corban C, He G, Pesaresi M, Li J, Shirazi Z, Liu J, Annoni A (2020) Big Earth Data science: an information framework for a sustainable planet. *Int J Digital Earth* 14:88–105
- ISDE Council (2012) What is digital earth. International Society for Digital Earth. Available at <http://www.digitalearth-isde.org/list-90-1.html>. Accessed 27 May 2021
- ISO TC184 (2020) ISO/DIS 23247 Automation systems and integration — Digital Twin framework for manufacturing. ISO, Geneva

- Jacoby M, Usländer T (2020) Digital twin and internet of things – current standards landscape. *Appl Sci* 10:6519
- Jones D, Snider C, Nassehi A, Yon J, Hicks B (2020) Characterising the Digital Twin: a systematic literature review. *CIRP J Manuf Sci Technol* 29:36–52
- Kritzinger W, Karner M, Traar G, Sihn W (2018) Digital Twin in manufacturing: a categorical literature review and classification. *IFAC-PapersOnLine* 51:1016–1022
- Nativi S, Delipetrev B, Craglia M (2020) Destination Earth: survey on “Digital Twins” technologies and activities, in the Green Deal area. European Commission, Luxembourg
- Nativi S, Mazzetti P, Craglia M (2021) Digital Ecosystems for developing Digital Twins of the Earth: the Destination Earth case. *Remote Sens* 13:2119–2144
- Tao F, Qi Q (2019) Make more Digital Twins. *Nature* 573:490–491

Dimensionless Measures

Ekta Baranwal

Department of Civil Engineering, Jamia Millia Islamia, New Delhi, India

Synonyms

Bare quantities; Pure quantities; Scalar quantities

* Words, ‘quantities’ and ‘metrics’ are the synonyms of ‘measures.’

Definition

Dimensionless measures are those quantities that do not have any physical dimension or are considered to have only one dimension. When any measure is derived from being independent of concurrent variations in the fundamental quantities chosen to derive it, then the derived quantity is dimensionless. These measures are part of “dimensional analysis” used in various fields like science, mathematics, engineering, and economics.

Introduction

Since the nature of the objects and features we see on the earth is three-dimensional. There is a general and universal acceptance that the geospatial data should be acquired and analyzed through methods that aim to examine and explore the features on the globe in three dimensions. But the question arises here “is it justified for doing so?”. Though, in a Penrose conference in 2006, i.e., “Unlocking 3-D Earth Systems—Harnessing New Digital Technologies to Revolutionize Multiscale Geologic Models Durham University, Durham, UK, 17–21 September 2006,” participants discussed understanding the

dimension, uncertainty magnitudes, and managing various types of geoscience datasets. Inspired by this conference, some of the participants produced a better understanding of dimensionality for geospatial or Geoscience data illustrating it as “Describing the dimensionality of geospatial data in the earth sciences—Recommendations for nomenclature” (Jones et al. 2008). According to them, “the dimensionality of an object is independent of whether the object’s location in space is known” (Jones et al. 2008). Therefore, dimensionality can be defined precisely in many ways in mathematics, but there are also varied definitions in science and engineering. For a clear understanding, we must know that mathematically, according to Euclidean geometry, the topological dimension of a single point is zero, for line or curve one, for plane or surface two, and for the volume is three. Hence, there is no dependency of the object’s dimension upon its spatial location through this aspect, i.e., representation of any object with its known attribute values in a coordinate system does not fluctuate its Euclidean dimension.

But in general scientific practice, the dimensionality of an object is considered the synonym of coordinate axes used to represent the position of that object in space. This approach is also widely accepted in earth science to define the dimensionality of geoscience data, for example, for developing maps and photographs of the earth’s surface using geoidal, cartesian, and cylindrical coordinate systems, each having three coordinate axes. Also, this practice is beneficial while transforming the representation of an object’s coordinate system into another coordinate system. Since a coordinate system is utterly capricious and random, the transformation of coordinate systems does not affect the dimension of the object while mapping its position in various coordinate systems. So, in brief, portraying the geoscience data in a particular coordinate system with three parameters or axes represents its dimensionality. This viewpoint leads towards the utilization of dimensional analysis for geospatial or geoscience data, i.e., dimensional analysis can be appropriate in geoscience and related fields.

Dimensional analysis is a mathematical operation and a part of mathematical modeling used to study and understand the nature of objects. It provides the relationship between two or more physical quantities through dimensions and measurement units. Many methods are used to obtain this relation for various applications in *physics, chemistry, mathematics, statistics, engineering, and economics*. Unit conversion and determining the preciseness of the equations are the primary and most common utilization of dimensional analysis. In mathematics, this unit conversion is done through a conversion factor considering no change in the relative amount. Performing the conversion factor requires building a ratio or fraction that equals value one. This developed ratio or fraction with value one is one of the quantities from dimensionless measures. Therefore, dimensionless quantities or measures are an essential part of dimensional analysis. There are several

methods used in Geoscience for analyzing geospatial data through landscape metrics, geospatial metrics, spatial statistics, and temporal and spatiotemporal metrics (Bhatti et al. 2018) like fractal dimension for determining the number of patches; various indices to know about the patch density, richness, shape, diversity, variation, distance, connectivity; integration, correlation, regression, interpolation, and many more mathematical and statistical operations are used in geoscience.

In geoscience, remote sensing images or raster data are widely used to get spatial information through various analyses. Of course, these analyses are based on several mathematical operations. Since the images and raster data form a matrix, their dimensional analysis and mapping are done through the mathematical operations of the dimension matrix. Dimension matrix, in this case, represents the spatial entities in images with unitless pixels. Details of the dimension matrix and dimensionless quantities can be referred to (Tiihonen 2016). Therefore, this geoscience dataset provides many possible dimensional and dimensionless metrics. Among the several measures used in geoscience, some dimensionless measures play a significant role by providing the base to geoscience data and analyzing it like the coefficient of determination, coefficient of variation, correlation, Euler's number, f-number (scale), pixel, Poisson's ratio, radian measure, Rayleigh number, refractive index, relative density, relative permeability, relative permittivity, transmittance, accuracy assessment, Kappa coefficient, spectral index, and goodness of fit. Given the above explanation of ratio, which is a dimensionless quantity used in many forms in geoscience, for example, scale is a ratio and a dimensionless metric that is a widespread, basic, but essential term used in geoscience either we talk about geoinformatics, geospatial information system (GIS), remote sensing, surveying, or photogrammetry. In the below sections, we will see the exploration of few dimensionless quantities widely used in Geoscience.

Dimensionless Measures in Geoscience

In geoscience, numerous metrics are used that belong to dimensionless quantities. For example, refractive index, Rayleigh number, radian measure, relative density, relative permeability, relative permittivity, transmittance, and Poisson's ratio are utilized in remote sensing fundamentals because the principles of remote sensing lie on the footprint of physics and optics. Most of these metrics are part of how various sensors transform electromagnetic radiation from features into signals and images. After that, pixel and scale are the dimensionless quantities used for images in remote sensing and photogrammetry. In GIS, scale is used for both vector and raster data, while the pixel is the metric for raster data. Pixels and scale, both measures, are the core in geoscience because experts can

identify and analyze the earth's surface features through these quantities. Apart from this, several indices were developed to study and extract the information from geospatial datasets, whether it is about land use land cover (LULC), spatiotemporal assessment of an object and area, dynamic changes in biophysical components and fauna and flora, or the assessment and management of various needs and services we are using in our daily lives. For making any decision, assessed and analyzed data of remote sensing and GIS are also validated with ground truth data or highly accurate detailed data. This validation is performed through the accuracy assessment methods, where again some dimensionless metrics are used in the process. Below are some examples to study and understand the role of some dimensionless measures in Geoscience that can also be found in other articles given in the cross-reference section.

Example 1: Pixel and Pixel Value

Photographs and images are two-dimensional representations of our three-dimensional world developed through an array of discrete picture elements. These picture elements are known as pixels which is the smallest and dimensionless measure of an image. Several pixels are arranged in rows and columns to form a photograph. Remote sensing images are the product of electromagnetic radiance of the features of the earth's surface. Therefore, each pixel in these images carries the radiance value, also called the brightness value (BV) (Fig. 1). Since images are digital, the pixel value is also known as a digital number (DN). Pixels in the image and its DN, both measures of a photograph, are dimensionless quantities. Pixel provides the minor information in an image that is considered the single unit element and an image sample. The DN in a pixel ranges from 0 to $2^n - 1$ according to the image's radiometric resolution (e.g., 8-bit, 16-bit, or 32-bit), which is also unitless or dimensionless and used in a variation of color tones in monochromatic and multispectral images. These pixel values and the color of pixels are used for all the quantitative and qualitative calculations in image processing to retrieve numerous information. Of course, the number of pixels varies in an image based on its spatial resolution and the scale on that image and the details of the information it contains.

Example 2: Scale

Scale is one of the essential elements in geoscience datasets, whether it is a photograph in remote sensing and photogrammetry or one kind of map in GIS. Scale is a ratio to represent the coverage area of the ground on an image or map. In other words, the scale represents the globe's features into a two-dimensional image or map or a simulated three-dimensional model, and therefore it is one of the characteristics of image and maps. Scale can be defined differently in various fields, and it is the base behind information details of covered ground area. Details about the features and objects present on a map



Row	0	1	2	3	4	5	6
0	10150	10227	10227	10324	10324	10319	10319
1	10150	10227	10227	10324	10324	10319	10319
2	10190	10209	10209	10280	10280	10382	10382
3	10190	10209	10209	10280	10280	10382	10382
4	10185	10194	10194	10290	10290	10349	10349
5	10185	10194	10194	10290	10290	10349	10349
6	10212	10229	10229	10312	10312	10342	10342
7	10212	10229	10229	10312	10312	10342	10342
8	10213	10271	10271	10249	10249	10375	10375
9	10213	10271	10271	10249	10249	10375	10375
10	10227	10229	10229	10279	10279	10364	10364
11	10227	10229	10229	10279	10279	10364	10364
12	10253	10236	10236	10305	10305	10300	10300
13	10253	10236	10236	10305	10305	10300	10300
14	10314	10323	10323	10343	10343	10269	10269
15	10314	10323	10323	10343	10343	10269	10269
16	10330	10359	10359	10308	10308	10321	10321
17	10330	10359	10359	10308	10308	10321	10321
18	10334	10344	10344	10292	10292	10241	10241
19	10334	10344	10344	10292	10292	10241	10241
20	10255	10309	10309	10261	10261	10268	10268
21	10255	10309	10309	10261	10261	10268	10268
22	10214	10381	10381	10368	10368	10295	10295

Dimensionless Measures, Fig. 1 16-bit Landsat OLI/TIRS image and some of its pixel values

are directly related to the scale of that map. “Scale is the ratio between the distances measured on the map and the corresponding distances measured on the ground” (Lo and Yeung 2007). Scale can be represented in fraction, statement, or linear graph and generally categorized as large-scale, medium-scale, and small-scale. This concept of scale is also for the scale of an image. Scale is directly proportional to the details on the map and image about features of the earth’s surface and their spatial resolution but inversely proportional to the ground coverage area. That means a larger-scale map and image have high spatial resolution and detail information about the scene because it covers a small portion of the ground like a map of a single ward of a city and images captured through drones for a small area. The large-scale map and images can distinguish between the most miniature objects, shapes, and boundaries.

On the other hand, small-scale maps and images do not provide minor details about features. However, they cover a large ground area with many objects and features and depict only major ones like highways, airports, rivers, general vegetated areas, and built-up areas. Map and images of medium-scale cover a geographic area relatively larger than large-scale and smaller than small-scale with details that lie in between the both like covering the minor roads, footbridges, differentiate between forest and agricultural land. Scale is a ratio, therefore a dimensionless metric in geoscience but of great importance. The scale of map or image is often about the spatial scale, but geoscience also uses the time scale (temporal scale), which is used to study and understand the changes on the earth’s surface. The temporal scale depends on how frequently the geospatial data is acquired. The higher the data collecting frequency, the higher the temporal resolution of the geospatial data and its temporal scale. Therefore, any

geospatial data’s scale (spatial and temporal) is one of the bases in any application. Even in an image, the number and size of pixels are based on the scale of that image. It again decides the density of information present in that image. Scale is also used for map projections, measuring the distortions while projecting the map, for geometric transformation, editing and analysis of geospatial data (Chang 2008). In Fig. 2, it is seen the difference between a medium and high-scale image, i.e., medium-scale image is only able to differentiate built-ups, river, and vegetation in gray, blue, and green color, respectively (Fig. 2, left). In contrast, individual entities can be identified through high-scale images like roads, rooftops of each house, tree canopies, and even bikes on the road (Fig. 2, right).

Example 3: Contrast Ratio (CR)

It is the ratio of brightest and darkest pixels of the image used to detect and resolve the objects by distinguishing the brightness value of the pixels (Sabins 2000). The ratio is represented as:

$$CR = \frac{B_{\max}}{B_{\min}}$$

where $B_{\max}$ is the maximum brightness value in the image and $B_{\min}$ is the minimum brightness value. CR provides the brightness profile in the scene, which is helpful to differentiate the objects and background in the image. CR is much beneficial for object detection through a monochrome or panchromatic image. Here, brightness values (pixel value) and CR (ratio of two dimensionless quantities), all measures are dimensionless.



Dimensionless Measures, Fig. 2 Medium-scale and resolution image (left) and high-scale and resolution image (right)

Example 4: Spectral Index or Band Ratio

Band ratio is similar to CR in detecting and highlighting features but through a multispectral image. Here, multiple spectral bands are used to form a ratio to extract features. Therefore, it is also known as spectral ratio and spectral index because it measures the changes between two or more spectral bands for retrieving various features. A spectral index is calculated through the ratio of spectral bands where DNs in one spectral band are divided by DNs in another spectral band. Therefore, the ratio is calculated pixel by pixel through a nonlinear transformation (Schowengerdt 2007) represented as:

$$\text{Spectral index} = \frac{\text{DNs of band } m}{\text{DNs of band } n}$$

The spectral index is widely used in remote sensing, but it is not that simple as the above expression. Sometimes other mathematical operators are used between two or more spectral bands to form the numerator and denominator of the ratio, which is also referred to as modulation ratio. The spectral ratio is also normalized to provide the range to the BVs of the final image, which is commonly from -1 to $+1$ to define the particular feature and other features. “The image obtained through spectral index poses a less dynamic range than the original image, resulting in a reduction of radiance extremes due to topography. Therefore, the reflectance contrast between various features is enhanced through different spectral ratios” (Schowengerdt 2007).

Numerous spectral ratios are present in the literature for several features like vegetation indices, built-up indices, water indices, soil indices, and many more (Lillesand et al. 2008; Jensen 2016; Baranwal and Ahmad 2021). Below are a few examples of band ratios widely used for multispectral images.

$$\text{Ratio vegetation index : RVI} = \frac{\text{NIR band}}{\text{RED band}}$$

$$\begin{aligned} \text{Normalized difference vegetation index : NDVI} \\ = \frac{\text{NIR band} - \text{RED band}}{\text{NIR band} + \text{RED band}} \end{aligned}$$

$$\begin{aligned} \text{Normalized difference built-up index : NDBI} \\ = \frac{\text{SWIR1 band} - \text{NIR band}}{\text{SWIR1 band} + \text{NIR band}} \end{aligned}$$

$$\begin{aligned} \text{Normalized built-up area index : NBAI} \\ = \frac{\left[\text{SWIR2 band} - \frac{\text{SWIR1 band}}{\text{GREEN band}} \right]}{\left[\text{SWIR2 band} + \frac{\text{SWIR1 band}}{\text{GREEN band}} \right]} \end{aligned}$$

$$\text{Modified normalized difference water index : NDWI}$$

$$= \frac{\text{GREEN band} - \text{MIR band}}{\text{GREEN band} + \text{MIR band}}$$

NIR stands for near-infrared, SWIR stands for short-wave near-infrared, and MIR for mid-near-infrared; SWIR and MIR both represent the same spectral band.

Example 5: Error Matrix and Measures Through It

Geospatial data are analyzed through several visual, mathematical, and statistical methods, but before making any decision, analyzed geospatial data is checked for its accuracy, which is commonly done through error matrix or confusion matrix. This error matrix is created based on the agreement and disagreement of each sample data of the geospatial dataset validated through high spatial ground reference data (Jensen 2016). Thus, the error matrix itself is a dimensionless

measure used to determine producer's and user's accuracy in deriving the error of omission and commission, respectively. Overall accuracy and Kappa coefficient are also calculated through this error matrix to measure samples' agreement. Producer's, user's, and overall accuracy are the proportion of the samples that fulfill the agreement of sample in the column, in the row, and for the entire error matrix, respectively. Kappa analysis is another dimensionless measure evaluated through an error matrix to compare and get the difference between the analyzed and the reference map or image (Lillesand et al. 2008; Jensen 2016). Six types of Kappa coefficients are used in literature to determine agreement and disagreement of quantity and allocation of samples through two different approaches of an error matrix, i.e., the traditional approach of Kappa analysis (Chang 2008; Lillesand et al. 2008; Jensen 2016) and new approach developed a few years back (Pontius and Millones 2011).

Summary and Conclusion

Data analysis in geoscience is dependent on dimensionless metrics. Users use these measures in geospatial data analysis and decision-making even without knowing that they are dealing with dimensionless quantities. The examples mentioned here are elementary, frequently used, and essential in geoscience.

Cross-References

- Accuracy and Precision
- Cluster Analysis and Classification
- Geographical Information Science
- Geostatistics
- Object-Based Image Analysis
- Remote Sensing
- Spatial Analysis
- Spatiotemporal Analysis
- Spectral Analysis
- Time Series Analysis
- Time Series Analysis in the Geosciences

Bibliography

- Baranwal E, Ahmad S (2021) Spectral ratioing: a computational model for quick information retrieval of Earth's surface dynamics. In: Singh SK, Kanga S, Meraj G, Farooq M, Sudhanshu S (eds) Geographic information science for land resource management. Scrivener Publishing LLC, Beverly, MA, pp 119–146
- Bhatti SS, Reis JP, Silva EA (2018) Spatial metrics: the static and dynamic perspectives. In: Comprehensive geographic information systems. Elsevier, Amsterdam, Netherlands, pp 181–196

- Chang KT (2008) Introduction to geographic information systems, 4th edn. Tata McGraw-Hill
- Jensen JR (2016) Introductory digital image processing – a remote sensing perspective. Pearson Education, Inc., Glenview
- Jones RR, Wawrzyniec TF, Holliman NS, McCaffrey KJ, Imber J, Holdsworth RE (2008) Describing the dimensionality of geospatial data in the earth sciences—recommendations for nomenclature. *Geosphere* 4(2):354–359
- Lillesand TM, Kiefer RW, Chipman JW (2008) Remote sensing and image interpretation, 6th edn. John Wiley & Sons, Inc., New Delhi, India
- Lo CP, Yeung AKW (2007) Concepts and techniques of geographic information systems, 2nd edn. Pearson Education Inc., New Jersey
- Pontius RG, Millones M (2011) Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *Int J Remote Sens* 32(15):4407–4429
- Sabins FF (2000) Remote sensing: principles and interpretation, 3rd edn. W. H. Freeman and Company, New York
- Schowengerdt RA (2007) Remote sensing: models and methods for image processing, 3rd edn. Elsevier, USA
- Tiihonen T (2016) Dimensional analysis. In: Pohjolainen S (ed) Mathematical modelling. Springer International Publishing, Cham, pp 111–119

Discrete Prolate Spheroidal Sequence

Dionissios T. Hristopulos

School of Electrical and Computer Engineering, Technical University of Crete, Chania, Crete, Greece

Abbreviations

- DFT Discrete Fourier Transform
 DPSS Discrete Prolate Spheroidal Sequence
 MTM Multitaper Method

Definition

The *discrete prolate spheroidal sequences (DPSSs)* are maximally concentrated in both the time and frequency domains. This is a crucial property for applications in power spectrum estimation. The DPSSs comprise sets of real-valued orthonormal sequences (discrete functions) with the following properties: (i) They are limited within the spectral band $[-W, W]$, where $W > 0$. (ii) The total energy of the sequence (i.e., the sum of the squares of the amplitudes) over a finite time interval is maximized. The *time-limited DPSSs*, also known as the *Slepian sequences*, are orthonormal sequences with the following properties: (i) They are *time-limited* within a finite time interval. (ii) They exhibit maximal spectral concentration in the frequency band $[-W, W]$. The DPSSs are parametrized in terms of the length N of the temporal sequence, the bandwidth W , and the order $k = 0, 1, \dots, N - 1$ of the Slepian sequence.

Overview

In Fourier analysis, it is known that functions cannot be fully localized in both time and frequency. As an extreme example, consider the *cosine* function $\cos(2\pi f_0 t)$ with frequency f_0 . Its Fourier transform involves the sum of two Dirac delta functions, $\delta(f \pm f_0)$. The *cosine* is fully extended in time, but the delta functions are completely localized at $\pm f_0$. In less extreme cases, functions that are concentrated (localized) in either the time or frequency space are extended (delocalized) in the other. This trade-off between localization in the direct (space or time) versus the spectral domain is also evident in Heisenberg's uncertainty principle of quantum physics.

Since functions cannot be completely confined both in the temporal and spectral domains, a fundamental problem is how to optimally concentrate the energy of a signal in the time (frequency) domain if the signal is spectrally (temporally) confined. The optimal concentration problem was addressed in a series of papers by the mathematicians Slepian, Pollack, and Landau (Landau and Pollak 1961, 1962; Slepian and Pollak 1961; Slepian 1978, 1983). These theoretical advances established the framework for the development of powerful spectral estimation methods (Thomson 1982; Percival and Walden 1993). The DPSSs solve the maximal concentration problem in the case of discrete functions (i.e., for time series). The solution of the concentration problem for continuous functions is provided by the so-called *prolate spheroidal wave functions* (Slepian 1978). The DPSS development is reviewed by Slepian (1983).

Methodology

Consider a finite-length discrete time sequence $t_n = n\delta t$, $n = 0, 1, \dots, N-1$, where $N > 1$ and the time step is $\delta t > 0$. In addition, let $x(t)$ represent a square-summable (finite-energy) signal, sampled at the times t_n , i.e., $x_n = x(t_n)$, $n = 0, \dots, N-1$. The *discrete-time Fourier transform* (DFT) of the sequence is given by

$$\tilde{x}(f) = \sum_{n=0}^{N-1} \exp(-2\pi i f t_n) x_n. \quad (1)$$

The DFT is defined over the frequency interval $-f_N < f \leq f_N$, where $f_N = f_s/2$ is the *Nyquist frequency* and $f_s = 1/\delta t$ is the sampling frequency.

Formulation of the Concentration Problem

The concentration problem consists of finding all the finite-energy time sequences $\{x_n\}_{n=0}^{N-1}$ which maximize the *spectral concentration ratio*

$$\lambda = \frac{\int_{-W}^W |\tilde{x}(f)|^2 df}{\int_{-f_N}^{f_N} |\tilde{x}(f)|^2 df}, \quad \text{where } 0 < W < f_N. \quad (2)$$

The parameter λ , $0 \leq \lambda \leq 1$, measures the ratio of the energy contained within the spectral band $[-W, W]$ over the total energy. The DPSSs comprise N discrete functions, denoted by $v_n^{(k)}(W, N)$, $k = 0, 1, \dots, N-1$, that maximize λ and are also orthogonal to each other. The DPSSs are the real-valued solutions of the following set of equations (Slepian 1978)

$$\sum_{m=0}^{N-1} \frac{\sin 2\pi W(n-m)}{\pi(n-m)} v_n^{(k)}(W, N) = \lambda_k(N, W) v_n^{(k)}(W, N), \quad n = 0, 1, \dots, N-1. \quad (3)$$

The above system is equivalent to the following eigenvalue problem:

$$\mathbf{A} \mathbf{v}^{(k)} = \lambda_k \mathbf{v}^{(k)}. \quad (4)$$

In Eq. (4), $\mathbf{A}$ is the *prolate matrix* with elements

$$A_{n,m} = \frac{\sin 2\pi W(n-m)}{\pi(n-m)}, \quad n, m = 0, \dots, N-1,$$

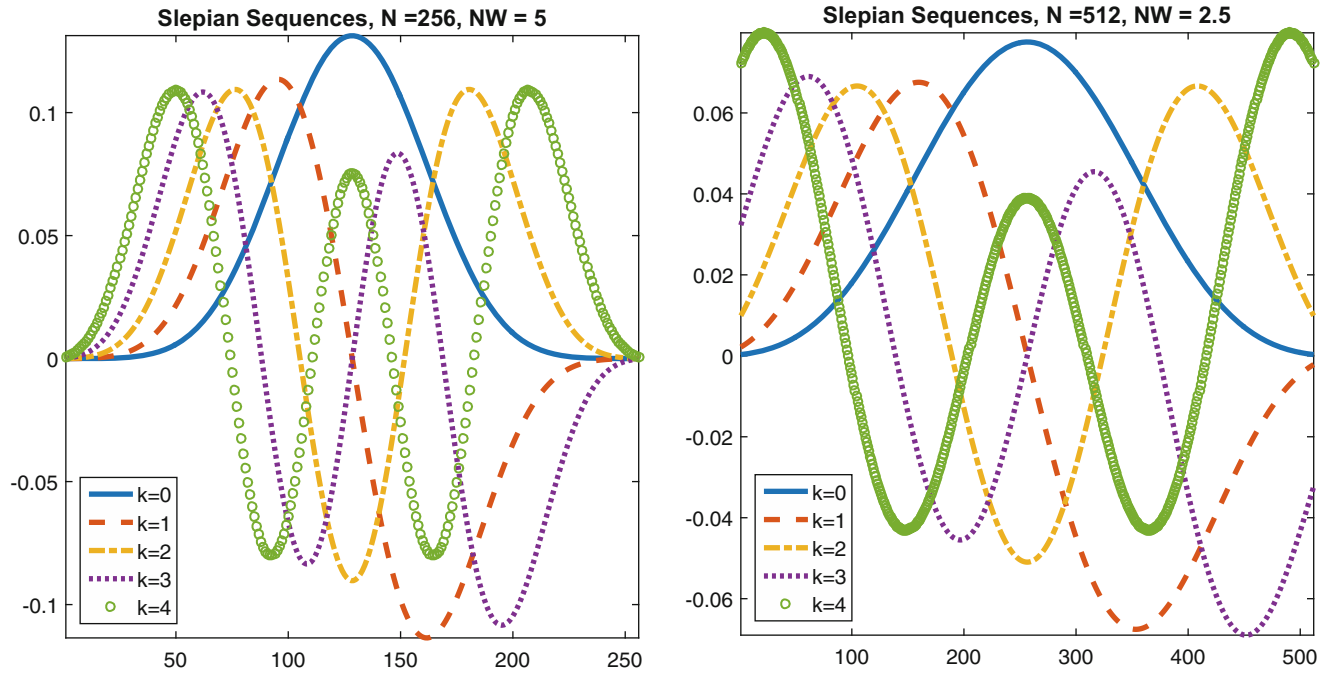
$\{\lambda_k\}_{k=0}^{N-1}$ is the set of N eigenvalues, and the DPSS vectors $\mathbf{v}^{(k)} = (v_0^{(k)}, \dots, v_{N-1}^{(k)})^\top$ are the respective eigenvectors (" $\top$ " denotes the transpose), which form the *Slepian basis*. The eigenvalues λ_k represent the spectral concentration ratios for the respective eigenvectors $\mathbf{v}^{(k)}$.

For continuous functions of time, the concentration problem is solved by an integral equation whose eigenfunctions are the *discrete prolate spheroidal wavefunctions* (Slepian and Pollak 1961; Slepian 1978).

Properties of the Slepian Eigenvectors and Eigenvalues

The DPSSs possess certain useful mathematical properties (Slepian 1983; Percival and Walden 1993; Lii and Rosenblatt 2008):

1. The eigenvalues are distinct, real-valued, and positive: $1 > \lambda_0 > \lambda_1 > \dots > \lambda_{N-1} > 0$. The positive eigenvalues reflect the positive-definiteness of the prolate matrix $\mathbf{A}$.
2. The eigenvectors $\mathbf{v}^{(k)}$ are even or odd according to whether k is even or odd.
3. The eigenvector $\mathbf{v}^{(k)}$ has exactly k zeros in the open interval $(0, N-1)$.



Discrete Prolate Spheroidal Sequence, Fig. 1 The first five DPSS eigenvectors for $N = 256$, $W = 5/N$ (left plot) and $N = 512$, $W = 2.5/N$ (right plot). Horizontal axes: time sequence ($\delta t = 1$). Vertical axes: DPSS sequences $v_n^{(k)}(W, N)$

4. The eigenvectors that correspond to different eigenvalues are mutually orthogonal, i.e., each eigenvector $v^{(k)}$ is orthogonal to all eigenvectors $v^{(l)}$, $l = 0, \dots, k-1$.
5. The eigenvalues λ_k exhibit *clustering behavior*: Slightly fewer than $N(W/f_N)$ eigenvalues are very close to one, slightly fewer than $N(1 - W/f_N)$ eigenvalues are very close to zero, while the remaining (very few) eigenvalues are not close to either one or zero (Lii and Rosenblatt 2008). Hence, the number $N(W/f_N)$ (which equals $2NW$, if $\delta t = 1$) is a rough estimate (upper bound) of the number of significant eigenvalues.
6. The integer part of $2NW$ is an approximate estimate for the dimension of the space of functions that are “essentially concentrated” in both time and frequency.

The first five eigenvectors for two Slepian bases displayed in Fig. 1 illustrate the above properties. The plot on the left-hand side of Fig. 1 corresponds to $N = 256$, $NW = 5$, and $\delta t = 1$. The five respective eigenvalues λ_k are all very close to one. The plot on the right-hand side corresponds to $N = 512$, $NW = 2.5$, and $\delta t = 1$. The respective eigenvalues λ_k are given by 1, 0.99984, 0.99622, 0.95213, and 0.71389. Note that λ_0 appears as 1 because $1 - \lambda_0 \approx 3 \cdot 10^{-6}$.

Applications in Spectral Estimation

An important practical problem is the estimation of the *power spectrum* (*power spectral density*) of a process based on a finite set of observations (sample) of the process.

The following focuses on *wide-sense stationary, ergodic random processes* $Z(t)$, which are assumed to extend over the infinite time domain, i.e., $t \in (-\infty, \infty)$. The spectral density $S(f)$ of $Z(t)$ is well-defined and coincides with the Fourier transform of the autocorrelation function according to the Bochner-Khinchin-Wiener theorem. The time series $\{z_n\}_{n=0}^{N-1}$, where $z_n = z(t_n)$, represents a finite sample of $Z(t)$ (“ z ” is used herein to avoid confusion with maximally concentrated functions denoted by “ x ”).

Periodogram Estimator

The two-sided *periodogram estimate* $\hat{S}(f)$ of the spectral density based on $\{z_n\}_{n=0}^{N-1}$ is given by

$$\hat{S}(f) = \frac{\delta t}{N} \left| \sum_{n=0}^{N-1} z_n \exp(-2\pi i f t_n) \right|^2, \quad -f_N < f \leq f_N. \quad (5)$$

To obtain the one-sided estimate (for $0 \leq f \leq f_N$), $\hat{S}(f)$ in Eq. (5) is multiplied by two for all frequencies except for $f = 0$ and $f = f_N$.

An estimator is called *consistent* if its variance tends to zero as $N \rightarrow \infty$. Unfortunately, the periodogram $\hat{S}(f)$ is not consistent, because its standard deviation is as large as the expectation (mean), even as the sample size $N \rightarrow \infty$.

The periodogram also suffers from *spectral leakage*. This occurs because the finite observation window implies that $Z(t)$ is multiplied by a boxcar kernel (rectangular time window) which localizes the signal within the finite sampling interval.

Multiplication with the window function implies the convolution of the Fourier transform $\tilde{z}(f)$ with that of the boxcar kernel in the spectral domain; the convolution perturbs $\tilde{z}(f)$ by mixing values at neighboring frequencies.

Welch's Method

The method proposed by Welch reduces the variance of $\hat{S}(f)$ by dividing the sample into different segments which are allowed to overlap. A modified periodogram (see below) is estimated for each time segment, and the segment estimates are averaged to generate the Welch estimate of the power spectral density. The assumed stationarity of $Z(t)$ ensures that the modified periodograms represent approximately uncorrelated estimates of $S(f)$. Thus, averaging over different segments reduces the variance of the Welch estimate.

The *modified periodograms* involve the multiplication of the time series (signal) in each time segment with a *window function*. The latter vanishes outside the specific segment, peaks in the middle of this segment, and is symmetric around the peak. Windows thus suppress the values of the time series near the segments' edges. Overlapping segments prevent loss of information, since signal values near the edges of one window are also near the center of neighboring windows. In addition, windowing reduces potential inter-segment correlations that could arise due to segment overlap. Various windows with subtle differences that impact *spectral leakage* are used in the literature, including the Slepian zero-order DPSS. The Kaiser (also known as Kaiser–Bessel) window provides a simple approximation of the Slepian window based on Bessel functions.

Multitaper Method for Spectral Density Estimation

The *multitaper method* (MTM) was also developed to address spectral leakage in power spectrum estimation (Thomson 1982; Percival and Walden 1993). In MTM, the DPSSs are used to construct low-bias, statistically consistent spectral estimators that reduce spectral leakage.

The MTM is a *nonparametric* estimation method, because it is data driven and does not need a parametric model of the process. For example, the classical periodogram and Welch's methods are also nonparametric, while the *maximum entropy method* is parametric (Percival and Walden 1993).

The variance reduction of the spectral estimates is achieved in MTM by using a small set of mutually orthogonal windows (tapers), while Welch's modified periodogram uses a single data taper. The optimal MTM tapers consist of DPSS eigenvectors, which minimize spectral leakage by maximizing energy concentration in the main spectral lobe. Both the mutual orthogonality and the optimal time-frequency concentration are critical for the success of the multitaper technique.

The MTM estimate is obtained by averaging K modified periodogram estimates based on a respective set of K Slepian

tapers. Averaging reduces the variance of the power spectrum. Only the fraction of tapers $\sim 2N \frac{W}{f_N}$ that generate small spectral leakage are used. Asymptotic properties for the bias and variance of MTM spectral density estimates are studied in Lii and Rosenblatt (2008).

Let $S_k(f)$ denote the *modified periodogram* obtained with the k -th Slepian sequence, $v^{(k)}$:

$$S_k(f) = \delta t \left| \sum_{n=0}^{N-1} v_n^{(k)} z_n \exp(-2\pi i f t_n) \right|^2. \quad (6)$$

In the simplest MTM version, the power spectral density is estimated by averaging the K modified periodograms:

$$S_{\text{MTM}}(f) = \frac{1}{K} \sum_{k=0}^{K-1} S_k(f). \quad (7)$$

Other MTM versions use taper-dependent weights w_k (Percival and Walden 1993):

$$S_{\text{MTM}}(f) = \frac{\sum_{k=0}^{K-1} w_k S_k(f)}{\sum_{k=0}^{K-1} w_k}. \quad (8)$$

Both the MTM and Welch's method average over approximately uncorrelated estimates of the power spectral density. However, the two approaches differ with respect to how they decorrelate the modified periodograms. Welch's method uses different segments of the signal for each periodogram. MTM uses the entire signal, but the decorrelation is enforced by the orthogonality of the Slepian tapers.

Selection of the MTM Parameters

The MTM involves two competing parameters: the number of tapers K and the half-bandwidth W . In order to understand their role, recall that DFT has a frequency resolution equal to the *Rayleigh frequency* $f_R = f_s/N$; this is the minimum frequency difference that can be resolved by the DFT of a finite-duration signal.

For MTM estimation, a large value of K is desirable to reduce the estimate for the variance of the power spectrum. However, in practice, only the first $K = N(W/f_N) - 1$ of the DPSS eigenvectors provide negligible spectral leakage (Slepian 1978; Thomson 1982; Ghil et al. 2002). Hence, if p is the largest integer that does not exceed $N(W/2f_N)$, K is bounded by $K \leq 2p - 1$. On the other hand, the frequency resolution of the MTM is $f'_R \approx pf_R = W$. This implies a trade-off between spectral resolution and variance reduction. Optimal values for p and K thus depend on the length N of the time series and the desired spectral resolution.

DPSS in the Geosciences

The Discrete Prolate Spheroidal Sequences find numerous applications in the spectral estimation of climatic time series, geochemical, paleoclimatic, oceanographic, stratigraphic, and seismological data [Percival and Walden (1993), Ghil et al. (2002) and references therein]. Extension to spherical Slepian functions for spatial and spectral concentration problems on the sphere has applications in geodesy, geophysics, and oceanography (Simons and Wang 2011).

Software Implementations

In **MATLAB** and R software, the DPSS is implemented by means of functions called **DPSS**; multitaper power spectral density estimation is implemented by means of the **PMTM** function in **MATLAB** and the package **MULTITAPER** in R. In Python, this functionality can be found in the **SPECTRUM** module.

Summary or Conclusions

Discrete prolate spherical (Slepian) sequences provide a mathematically elegant solution to the problem of time-frequency concentration: They are limited within a specific time (spectral) window and also maximize the energy content within a finite spectral (time) window. Slepian sequences are used as window functions in Welch's modified periodogram spectral estimation. The DPSSs are also an integral component of the multitaper method of spectral estimation. The MTM leads to considerable improvements over the classical periodogram in terms of variance and spectral leakage reduction.

Cross-References

- [Fast Fourier Transform](#)
- [Maximum Entropy Method](#)
- [Power Spectral Density](#)
- [Spectral Analysis](#)
- [Stochastic Geometry in the Geosciences](#)
- [Time Series Analysis](#)

Bibliography

- Ghil M, Allen MR, Dettinger MD, Ide K, Kondrashov D, Mann ME, Robertson AW, Saunders A, Tian Y, Varadi F, Yiou P (2002) Advanced spectral methods for climatic time series. *Rev Geophys* 40(1):1–41. <https://doi.org/10.1029/2000RG000092>
- Landau HJ, Pollak HO (1961) Prolate spheroidal wave functions, Fourier analysis and uncertainty – II. *Bell Syst Tech J* 40(1):65–84. <https://doi.org/10.1002/j.1538-7305.1961.tb03977.x>

- Landau HJ, Pollak HO (1962) Prolate spheroidal wave functions, Fourier analysis and uncertainty – III: the dimension of the space of essentially time- and band-limited signals. *Bell Syst Tech J* 41(4):1295–1336. <https://doi.org/10.1002/j.1538-7305.1962.tb03279.x>
- Lii K, Rosenblatt M (2008) Prolate spheroidal spectral estimates. *Stat Probabil Lett* 78(11):1339–1348. <https://doi.org/10.1016/j.spl.2008.05.022>
- Percival DB, Walden AT (1993) *Spectral analysis for physical applications*. Cambridge University Press, New York
- Simons FJ, Wang DV (2011) Spatiospectral concentration in the Cartesian plane. *GEM Int J Geomath* 2(1):1–36. <https://doi.org/10.1007/s13137-011-0016-z>
- Slepian D (1978) Prolate spheroidal wave functions, Fourier analysis, and uncertainty – V: the discrete case. *Bell Syst Tech J* 57(5):1371–1430. <https://doi.org/10.1002/j.1538-7305.1978.tb02104.x>
- Slepian D (1983) Some comments on Fourier analysis, uncertainty and modeling. *SIAM Rev* 25(3):379–393. <https://doi.org/10.1137/1025078>
- Slepian D, Pollak HO (1961) Prolate spheroidal wave functions, Fourier analysis and uncertainty – I. *Bell Syst Tech J* 40(1):43–63. <https://doi.org/10.1002/j.1538-7305.1961.tb03976.x>
- Thomson DJ (1982) Spectrum estimation and harmonic analysis. *Proc IEEE* 70(9):1055–1096. <https://doi.org/10.1109/PROC.1982.12433>

Discriminant Analysis

D. Arun Kumar

Department of Electronics and Communication Engineering and Center for Research and Innovation, KSRM College of Engineering, Kadapa, Andhra Pradesh, India

Definition

Discriminant Analysis, also called Linear Discriminant Analysis (LDA), is a mathematical approach to project the data points (in digital images data points are named as pixel values) from higher dimensional space to a lower dimensional space by removing the redundant features of data points. Removal of redundant features in the data points is identified as dimensionality reduction. The informative data in lower dimensions is obtained based on maximizing between-class variance and minimizing within-class variance. LDA is implemented in three important steps: (i) calculation of between-class variance; (ii) calculation of within-class variance; and (iii) constructing lower dimensional space with the data points having maximum between-class variance and minimum within-class variance.

Introduction

The utilization of Earth resources for various human needs has increased over the years. Assessment and monitoring of Earth resources are useful to plan for sustainable

development. Remote sensing is used to get the information about the resources through digital images with the advantages like spatial, spectral, temporal, and radiometric resolutions (Jensen 2004). Digital images consist of finite elements called pixels. The pixel values are also called digital numbers (DNs) (Gonzalez and Woods 2008). The pixel values in multispectral images are called pixel vector or a pattern. A pattern is associated and characterized by features. A pattern is a point in N dimensional feature space (Richards and Jia 2006).

Advancement in remote sensing technology has produced data with a higher number of features (also called higher dimensional data) (Richards and Jia 2006). Extracting the information from the higher dimensional data is a challenge in remote sensing data processing (Campbell 1987). One such type of data processing approach is data dimensionality reduction by preserving the information content (Lillesand et al. 2014). The dimensionality reduction techniques are broadly classified as (i) Unsupervised and (ii) Supervised (Fisher 1936). In Unsupervised dimensionality reduction approach, the techniques are applied to unlabeled data to extract the meaningful information. The methods like non negative matrix factorization (NMF), independent component analysis, (ICA) and principal component analysis (PCA) are mostly used unsupervised dimensionality reduction techniques (Jensen 2004). In supervised approach, the reduction techniques are applied on the labeled dataset. There are various methods of dimensionality reduction in labeled dataset such as neural networks (NNs), mixture discriminant analysis (MDA), and fuzzy rough set based functional dependency (FRFD) (Duda et al. 2000). Linear discriminant analysis (LDA) is mostly used dimensionality reduction technique for labeled dataset.

Linear Discriminant Analysis

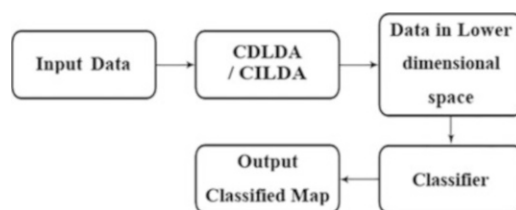
LDA is also identified as a data classification technique that converts features of pixels into a lower dimensional feature space from a higher dimensional feature space (Bishop et al. 1995). The approach in LDA is based on increasing the ratio of between-class variance to within-class variance of a labeled dataset (Tharwat et al. 2017). In LDA, dimensionality methods are classified into two categories: (i) Class-Independent LDA (CILDA) and (ii) Class-Dependent LDA (CDLDA) (Fisher 1936). In CILDA, the pixels of all the classes are projected into a common lower dimensional feature space. Due to this reason, the CILDA method is called Class-independent approach. In CDLDA, the pixels of each class are projected to class-specific lower dimensional feature space (Fisher 1936). In CILDA and CDLDA, the pixels are projected into lower dimensional feature space and later, the

pixels are classified using various linear or nonlinear classification approaches (Shahdoosti and Mirzapour 2017).

There exist various LDA approaches, such as removal of null space within-class matrix as a solution of finding lower dimensional space in the case, if the number of features are higher than the number of samples in the dataset. If the number of features is greater than the number of samples in the dataset, it leads to small sample problem (SS) (Tharwat et al. 2017). The other approach of LDA is to use kernel functions to transform within-class matrix to a full-rank matrix. There are different variants of LDA technique to solve the SS problem such as direct LDA, PCA + LDA, regularized LDA, Null LDA, and kernel DLDA (Fisher 1936). The application of LDA is high in diversified fields like land use/land cover classification, face recognition, and cancer classification (Bishop et al. 1995).

The functional block diagram of LDA for dimensionality reduction is given in Fig. 1. In Fig. 1, the dimensionality of input data is reduced using CILDA/CDLDA approaches. In the next stage, the data points are projected into lower dimensional feature space. Later, the pixels in the lower dimensional space are classified using classifiers such as K-Nearest Neighbor (KNN), Minimum Distance to Mean classifier (MDM). The classifier generates the classified output map which is evaluated using ground truth data.

In the present study, the detailed description of various stages of LDA is provided. In the first stage, the between-class variance of data points from different classes is calculated to know the distribution of points in the feature space. In the second stage, within-class variance among the data points of same class is calculated to know the distribution of data points within the class. In the third stage, the details related to the construction of lower dimensional space with the data points having maximum between-class variance and minimum within-class variance is presented. An example of LDA on the labeled dataset is given in section “[Example](#).” The evaluation metrics of LDA is given in section “[Evaluation Metrics](#)”. The conclusions related to LDA are provided in section “[Conclusions](#).”



Discriminant Analysis, Fig. 1 Schematic flow diagram of linear discriminant analysis. The flow diagram consists of steps like input data, processing input data using CDLDA/CILDA, obtaining data points in lower dimensional space, classifying the data, and obtaining output classified map

Stage I: Calculation of Between-Class Variance

In a dataset of c classes, the between-class variance of i th class is the distance between the mean of i th class (μ_i) and the mean of the dataset (μ). The overall between-class variance (V_D) of dataset is given as:

$$V_D = \sum_{k=1}^c n_k V_k, \quad (1)$$

where n_k is the number of samples or data points or pixels in i th class, c is the number of classes in the dataset. V_i is the variance of i th class. The variance of i th class (V_i) is calculated as:

$$(\mu_i - \mu)^2 = W^T V_i W, \quad (2)$$

where W is the transformation matrix of LDA. The LDA approach maximize the between-class variance, that is, maximize the distance between classes and project the points into lower dimensional space (Tharwat et al. 2017).

Stage II: Calculation of Within-Class Variance

The within-class variance of i th class is defined as the difference between the mean and the pixel values of respective class. The LDA technique projects the pixels into lower dimensional space such that the pixels have minimum within-class variance. The within-class variance of each class in the dataset (V_{Wj}) is calculated as:

$$V_{Wj} = \sum_{k=1}^{n_j} (P_{kj} - \mu_j)(P_{kj} - \mu_j)^T, \quad (3)$$

where P_{kj} represents the k th pixel of j th class, μ_j is the mean of j th class. The detailed explanation of LDA with an example is given in (Tharwat et al. 2017).

Stage III: Construct Lower Dimensional Space

The lower dimensional space for projecting the pixels is computed using transformation matrix (W) in LDA. The transformation matrix (W) is calculated based on Fisher's criterion as:

$$V_{Wj} W = \lambda V_D W, \quad (4)$$

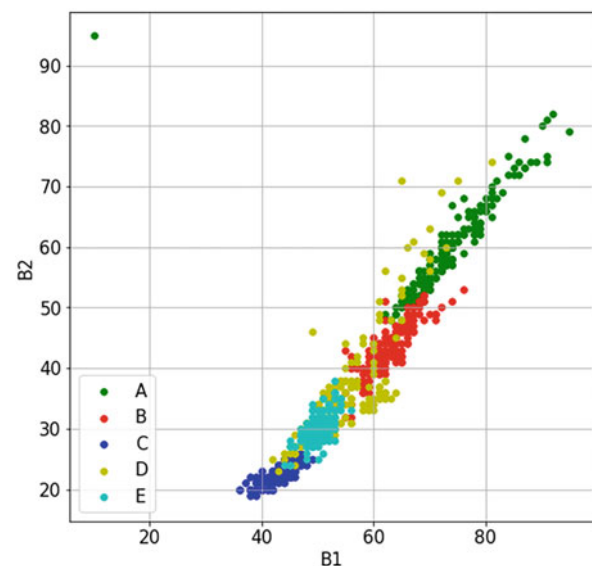
where λ is the eigenvalues of matrix W . The solution of Eq. 4 is obtained by calculating the eigenvalues ($\lambda = \lambda_1, \lambda_2,$

$\lambda_3, \lambda_4, \dots, \lambda_p$) and eigenvectors ($v = v_1, v_2, v_3, v_4, \dots, v_p$). Eigenvalues are scalar and the eigenvectors are derived from the eigenvalues which are nonzero vectors. The eigenvalues and eigenvectors satisfy the Eq. 4 and provide the information related to reduced feature space after the application of LDA.

In the multidimensional space, the eigenvectors represent the directions of the new space, and the corresponding eigenvalues represent the magnitude of the eigenvectors. Each eigenvector represents a feature axis of LDA space and the eigenvalues represent the intensity of eigenvector. The intensity of eigen vector represents its ability to discriminate different classes in the feature space by increasing between-class variance, and by decreasing within-class variance of each class. The eigenvectors with highest eigenvalues are considered to construct lower dimensional feature space and the eigenvectors with lowest eigenvalues are neglected. The selected eigenvectors are also called principal components and each principal component (a vector) is a feature axis in lower dimensional space. The data points or pixels are then projected into lower dimensional space using the transformation matrix. The pixels that are projected into lower

Discriminant Analysis,
Table 1 IRS LISS III
dataset and number of
samples

Land use/land cover type	Samples
Urban dense	200
Urban sparse	200
Forest	200
Water	200
Agriculture	200



Discriminant Analysis, Fig. 2 Distribution of highly overlapping data points in Band 1 and Band 2 feature space. Data points with label A are Urban Dense, B – Urban Sparse, C – Forest, D – Water, E – Agriculture

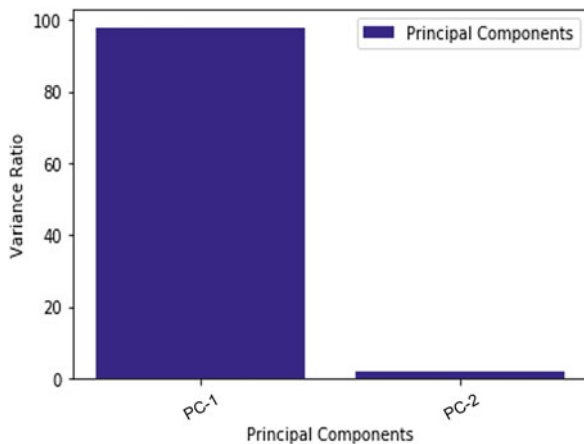
dimensional space are classified using the classifiers such as KNN, MDM, Neural networks, etc.

Example

In the present study, five major land use/cover classes for IRS LISS III dataset of Mumbai region, India, were considered. The classes in the dataset are Urban Dense, Urban Sparse, Water, Forest, and Agriculture. The dataset consists of 1000 samples/pixels such as 200 samples from each class. The

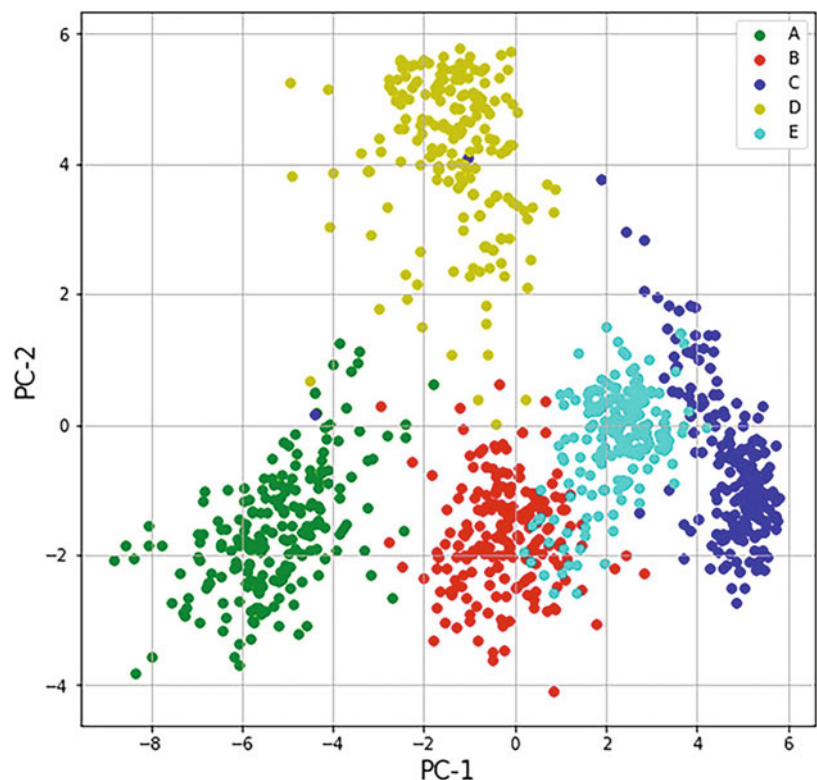
details of the dataset are given in Table 1. Each pixel is characterized by four features (four spectral bands such as Band 1, Band 2, Band 3, and Band 4) and 1000 pixels are projected as 1000 points in 4D feature space. The distribution of 1000 pixels in Band 1 and Band 2 feature space is given in Fig. 2. In Fig. 2, the pixels of five classes are highly overlapped with less chances of separability. The dense distribution of pixels in the feature space (Fig. 2) is transformed to lower dimensional feature space with good separability between the pixels of different classes.

LDA is applied on the input dataset and the corresponding principal components are computed (as mentioned in the above sections) to project the points in to lower dimensional feature space. The two principal components PC1 and PC2 with the corresponding variance ratio are given in Fig. 3. In Fig. 3, PC1 has maximum variance ratio (98%) which indicate that the transformed points have high separability along the PC1 axis. The separability of data points in the lower dimensional space along the PC1 axis with respect to PC2 axis is given in Fig. 4. Later, the data points with high separability in the lower dimensional feature space are classified using various type of linear/non-linear classifiers or parametric/non-parametric classifiers. The classifiers such as KNN, MDM, NN, decision trees, and random forest are frequently used for data classification (Richards and Jia 2006). In the present study, the overall accuracy of KNN classifier ($K = 5$) before the application of LDA on the dataset is 95% while the overall accuracy of KNN classifier ($K = 5$)



Discriminant Analysis, Fig. 3 The two principal components namely PC1 and PC2 with corresponding variance ratio

Discriminant Analysis, Fig. 4 Data points in 2-D feature space with maximum separability along PC1 axis. Data points with label A are Urban Dense, B – Urban Sparse, C – Forest, D – Water, E – Agriculture



after LDA on the dataset is 98%. Application of LDA on LISS III dataset improved the class separability which resulted in 3% improvement in the overall accuracy for KNN classifier.

Evaluation Metrics

The accuracy of classified pixels/data points is measured using different metrics like users' accuracy, producers' accuracy, overall accuracy, kappa coefficient, and dispersion score. These metrics are quantified from the confusion matrix or accuracy matrix (Jensen 2004).

Conclusions

Advancement of remote sensing sensor technology produced a large amount of spatial data. There are various data processing techniques such as classification, clustering, and regression analysis. Data/pixel classification is mostly used preprocessing technique in remote sensing data analysis. Pixel classification in the highly overlapping classes is a challenging task. LDA is used to improve the separability among the classes by reducing the number of features. In the present study, the application of LDA on LISS III dataset reduced the number of feature from four to two. Also, the data points in lower dimensional feature space have high separability. Application of LDA followed by KNN classifier improved the overall accuracy by 3% (with 98% overall accuracy value LDA + KNN) for IRS LISS III dataset.

Cross-References

- [Geoinformatics](#)
- [Geomatics](#)
- [Remote Sensing](#)
- [Spatial Data](#)
- [Spatial Statistics](#)

Bibliography

- Bishop CM et al (1995) Neural networks for pattern recognition. Oxford University Press, Oxford
- Campbell JB (1987) Introduction to remote sensing. The Guilford Press, New York
- Duda RO, Hart PE, Stork DG (2000) Pattern classification, 2nd edn. Wiley, New York
- Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Ann Eugenics* 7(2):179–188
- Gonzalez RC, Woods RE (2008) Digital image processing, 3rd edn. Prentice Hall, Upper Saddle River

- Jensen JR (2004) Introductory digital image processing: a remote sensing perspective, 3rd edn. Prentice Hall, Upper Saddle River
- Lillesand T, Kiefer RW, Chipman J (2014) Remote sensing and image interpretation. John Wiley & Sons, Hoboken
- Richards JA, Jia X (2006) Remote sensing digital image analysis: an introduction, 4th edn. Springer Verlag, Berlin
- Shahdoosti HR, Mirzapour F (2017) Spectral spatial feature extraction using orthogonal linear discriminant analysis for classification of hyperspectral data. *Eur J Remote Sens* 50(1):111–124
- Tharwat A, Gaber T, Ibrahim A, Hassanien AE (2017) Linear discriminant analysis: a detailed tutorial. *AI Commun* 30:169–190. <https://doi.org/10.3233/AIC-170729>

Discriminant Function Analysis

Norman MacLeod

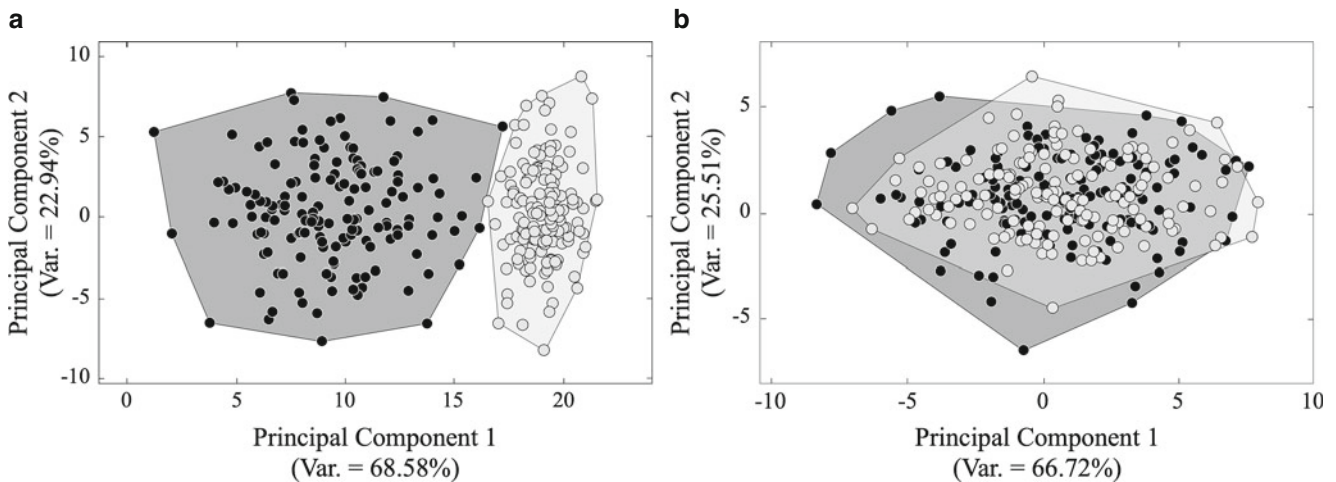
Department of Earth Sciences and Engineering, Nanjing University, Nanjing, Jiangsu, China

Definition

Discriminant Function Analysis – A family of data analysis methods that attempt to find a linear combination of variables that optimally separates two or more groups, categories, or classes of objects, samples, or events.

Introduction

It is often the case that earth scientists become involved with the multivariate analysis of datasets that exhibit subgroup-level structure that has been established a priori, at the outset of an investigation. If this knowledge is irrelevant to the question(s) being investigated, or if the purpose of the investigation is exploratory in the sense of determining whether differences within the pooled sample reflect the structure of these a priori subgroupings, it can be appropriate to employ pooled-sample, ordination-based, data-analysis methods (e.g., principal components analysis, principal coordinates analysis, factor analysis). For example, if a dataset is composed of three or more variables drawn from two groups of specimens, and the variable(s) exhibiting largest difference between group means is/are also the variable(s) with the largest variance(s), it will often be the case that an eigenanalysis of these data will reveal the subgroup-level structure on the first few eigenvectors (Fig. 1a). However, if a dataset is composed of three or more variables drawn from two groups of specimens, and the variable(s) exhibiting largest between-groups difference(s) is/are not the variable(s) with the largest variance(s) (Fig. 1b), the existence of group-level structure may only be apparent on higher-level eigenvectors whose



Discriminant Function Analysis, Fig. 1 Simulated data demonstrating the inability of eigenvector-based ordination methods (e.g., PCA, PCoord. Factor Analysis) to be relied on to identify subgroup-level structure in multivariate datasets. (a) Results from the analysis of a three-variable dataset in which a discontinuity in variation was placed in the variable exhibiting the greatest variance. (b) Results from the analysis of a three-variable dataset in which a discontinuity in variation was placed in the variable exhibiting the least variance. Note that, despite

PCs 1 and 2 including information from all three variables, there is no hint of the existence of subgroup structure in B. This discontinuity would be revealed in ordination plots that included PC-3 for this dataset. But in higher dimensional data it would very easy to overlook this structure if it was located in low-variance variables. This discrepancy demonstrates why discriminant analysis is required if subgroup structure is suspected and is a target in many data-analysis situations

interpretive significance may be overlooked, or perhaps on none at all. In these instances it would be grievous error to conclude that subgroup-level structure either did, or did not, exist in these data on the basis of an analysis that was based only on consideration of the pooled-sample structure. Unfortunately, the published earth-science literature is replete with examples of researchers who have used pooled-sample data-analysis methods to test hypotheses concerning the existence of subgroup-level structure or, even more seriously, the lack thereof. It was to address the complications that arise in the consideration of datasets in which there is reason to believe subgroup-level structure exists that methods of linear discriminant function analysis were developed.

Discrimination, Classification, and Identification

Before embarking on a discussion of linear discriminant analysis, it is necessary to distinguish between the processes of discrimination, classification, and identification as these terms are often confused and, indeed, have different meanings in the fields of earth science, mathematics, and computer science. Discrimination represents any attempt, by means of quantitative procedures, to discover and define a combination of variables that can be used to maximize the differences between two or more a priori-defined groups or subgroups belonging to a single group. The mathematical functions or models best suited to the discrimination task may be linear or nonlinear, continuous or discontinuous, etc. Irrespective of

these differences, the discrimination problem is a problem of optimization.

Discrimination requires knowing how many groups need to be separated from one another at the outset of the investigation. In order for the discriminant functions so defined to have any practical utility outside the context of the data used to determine them, it is also required that the samples used to estimate these functions be representative statistically of the populations from which they were drawn as it is these populations that are the actual targets of the investigation. This requires careful attention be paid to the collection of the samples used in a discriminant analysis. If these samples are biased or nonrepresentative of the larger populations from which they were drawn all results obtained from their analysis will only apply to the samples themselves, not to the larger populations. Finally, if the actual targets of the hypothesis tests under consideration are these larger populations, performance of the estimated discriminant functions must be tested via reference to independent, external samples that were not used to determine the discriminant functions in question.

Classification is the act of determining how many subgroups a larger group may be subdivided into and/or the act of determining which larger group to which a subgroup may be assigned (e.g., fossil species attributed to a genus, minerals attributed to a mineral family). As such, it is a completely different and distinct procedure from that of discrimination. If classification is the goal the procedure is not one of optimization, but rather one of discovery. In quantitative data analysis this discovery is often made when methods of ordination

or similarity/dissimilarity clustering detect subgroupings of objects whose variable values are either completely or relatively distinct from other subgroupings. Accordingly, a much wider array of numerical procedures can be brought to bear on the question of classification than is the case for discrimination. Moreover, in the vast majority of instances it will be impossible to know how many completely or relatively distinct subgroupings of objects are present in any given dataset. It must also be assumed that different types of quantitative rule sets might be involved in the recognition of different subgroupings that exist, especially when the dimensionality of the data is high. Complicating this task further is the fact that the number of specimens required by an analysis must increase near exponentially as dimensionality increases in order to maintain a density of observations in the variable space sufficient to be able to recognize subgroupings of objects stably (= the curse of dimensionality, see Altman and Krzywinski 2018).

Once a classification has been devised, qualitative, simple qualitative or complex numerical procedures can be employed to facilitate the placement of unknown objects into the classification scheme. This is identification. Classifications are utilitarian in the sense of the same objects being allocated to different categories within different classification systems that may be based on different types of variables, observations attributes, and/or measurements. Identification procedures need neither be quantitative nor optimized though there is obvious value in placing such decisions on an optimized, quantitative footing where possible. Identification can also become part of the discovery/classification process in the sense that objects (or collections of variable values) falling outside the ranges of groups or subgroupings already in the classification system may signal the presence or existence of new subgroup-level categories.

These distinctions between discrimination, classification, and identification must be kept in mind when undertaking either research or applications in any of these areas. In addition, when reading the works of researchers that employ these terms care must be taken to understand how the terms they use map onto the concepts described above, lest great and unnecessary confusion result (e.g., computer scientists typically use the term “classification” to refer to the concept of discrimination).

Fisher (Two-Group) Linear Discriminant Analysis

Linear discriminant analysis was developed originally by R. A. Fisher (1936) to address the problem of optimally allocating individual sets of observations to one of two subgroupings using a linear classification function that maximized the F ratio of $\mathbf{B}/\mathbf{T}$ where $\mathbf{B}$ represents the between subgroupings variance-

covariance matrix and $\mathbf{T}$ represents the total (= pooled sample) variance-covariance matrix.

Fisher's original method was restricted to the consideration of only two subgroupings and geometrically amounted to solving the problem by maximizing the separation of their multivariate means relative to their pooled sums of squares and cross products. This calculation resulted in a function of the general form:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \cdots + \beta_p x_p \quad (1)$$

where the x values represent real-variable values of collected data and the β coefficients refer to weights assigned to the variables that represent the angular relation between the variable and the linear discriminant vector. Under Fisher's original model it is assumed that the optimal linear discriminant vector is oriented at some angle to all variable vectors such that no single variable accounts for the maximum separation of multivariate means. In matrix notation this two-subgroup linear discriminant vector can be found by solving the following equation for λ .

$$S\lambda = D \quad (2)$$

where

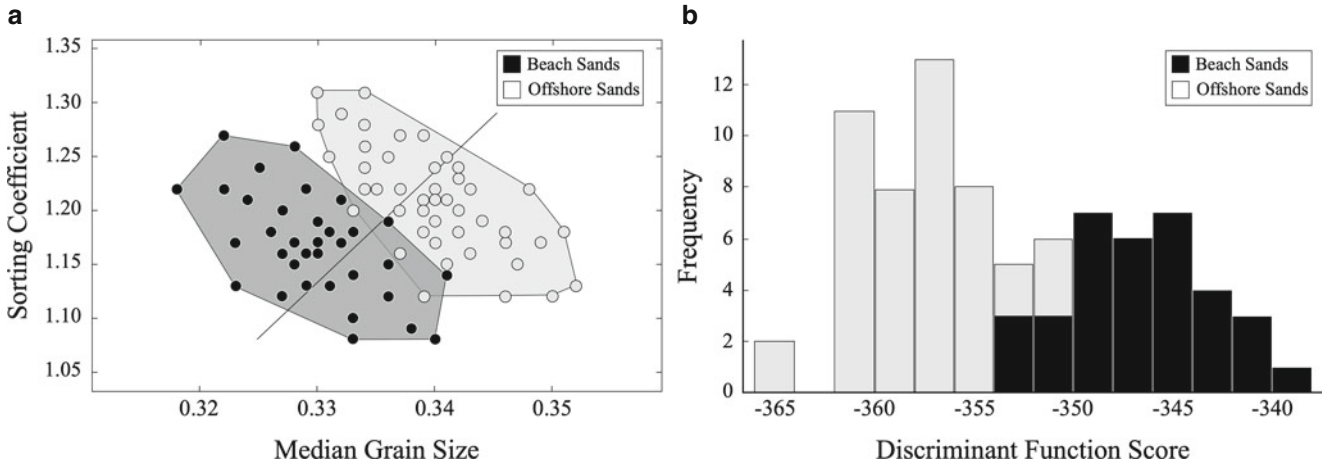
$$S = \frac{S_A + S_B}{n_a + n_b - 2} \quad (3)$$

$$S_A = \sum_{j=1}^m \sum_{i=1}^{n_a} x_{aj} x_{aj} - \frac{\sum_{i=1}^{n_a} x_{ai} \sum_{i=1}^{n_a} x_{ai}}{n_a} \quad (4)$$

$$S_B = \sum_{j=1}^m \sum_{i=1}^{n_b} x_{bj} x_{bj} - \frac{\sum_{i=1}^{n_b} x_{bi} \sum_{i=1}^{n_b} x_{bi}}{n_b} \quad (5)$$

$$D = \bar{A} - \bar{B} \quad (6)$$

Objects, samples, or events described by the set of variables may be projected onto the discriminant axis by pre-multiplying λ by the observed data values. By convention, the decision point that separates subgroup domains along the linear discriminant function is usually taken to be the midpoint between the subgroup centroids. For a simple analysis of linear discrimination between beach sands and offshore sands collected from the Texas Gulf Coast (see the Davis 2002 SANDS.txt datafile; <https://bcs.wiley.com/he-bcs/Books?action=index&bcsId=1479&itemId=0471172758>) application of these equations produces the results shown in Fig. 2.



Discriminant Function Analysis, Fig. 2 Results from a classic Fisher (1936) (two-group) linear discriminant analysis applied to a marine sand size, sorting dataset. (a) data values plotted in the space of the original variables with the Fisher linear discriminant function. (b) Histogram of

the beach sand and offshore sand data projected onto the linear discriminant function. Note the broad range of distribution overlap that characterize each of the original variables in **a** and the manner in which this overlap has been minimized in **b**

Multiple Group Linear Discriminant Analysis (Canonical Variates Analysis)

While the classic, formulation of linear discriminant analysis established by Fisher (1936) is perfectly serviceable for the simple two-group case, most practical discriminant analysis problems involve more than two groups or subgroups of a single group. Fortunately this procedure has been generalized and, providing sufficient sample sizes are available, can be used to analyze any number of subgroups. The modern formulation of this extension, termed canonical variates analysis (CVA), takes advantage of several important improvements suggested by Bartlett (1951) and Rao (1952).

The general form of Eq. 1 is used to define a set functions that will yield the maximum possible F value while not being correlated the other discriminant functions. In this way the discriminant functions calculated as a result of CVA can be regarded as being similar to principal components. However, an important difference between PCA and CVA that must to kept in mind is that, while the canonical variates are orthogonal to one another in their own space, they will not be orthogonal to one another in the space of the original variables. This unfortunate feature can make the qualitative interpretation of individual canonical variates difficult, but interpretation can often be aided for some types of data (e.g., morphological) via modeling of the canonical variates space. Another difference between CVA and PCA is that, for a discriminant problem involving k groups, only $k-1$ canonical variates will result irrespective of the number of variables included in the data.

Finding the equations of the canonical variates is a surprisingly simple procedure involving a logical partitioning of the

total (pooled-sample) variance-covariance matrix T , into two components, a within-groups variance-covariance matrix W and a between-groups variance-covariance matrix B , as follows.

$$T = B + W \quad (7)$$

where

$i = i^{\text{th}}$ object or sample

$j = j^{\text{th}}$ variable

$k = k^{\text{th}}$ group

$$\bar{X}_{kj} = \frac{\sum_{i=1}^{n_k} x_{kij}}{n_k} \quad (8)$$

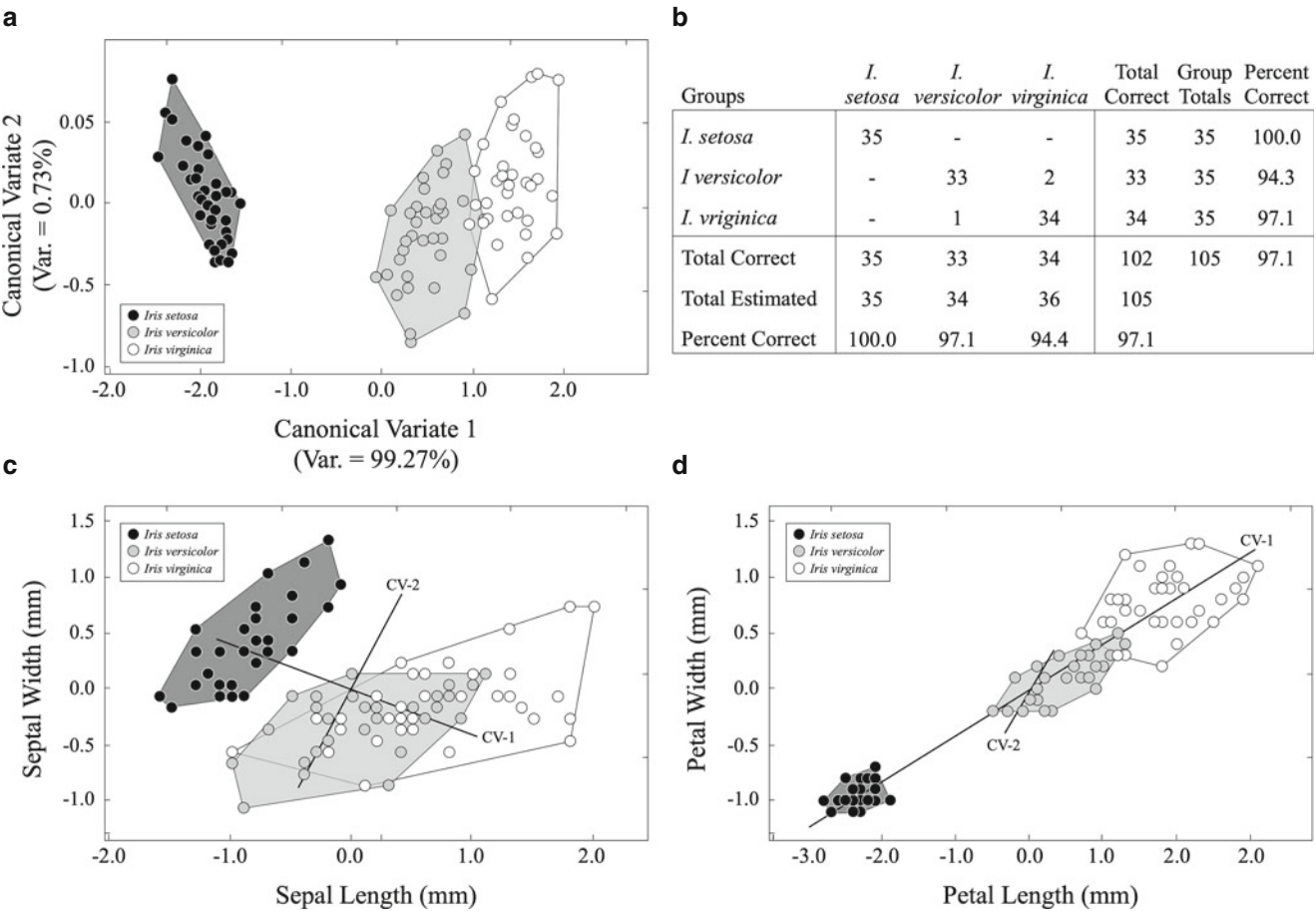
$$\bar{\bar{X}}_j = \frac{\sum_{k=1}^g \sum_{i=1}^{n_k} x_{kij}}{N} \quad (9)$$

$$T_{pq} = \sum_{k=1}^g \sum_{i=1}^{n_k} (x_{kij} - \bar{\bar{X}}_p) (x_{kij} - \bar{\bar{X}}_q) \quad (10)$$

$$W_{pq} = \sum_{k=1}^g \sum_{i=1}^{n_k} (x_{kij} - \bar{X}_{kp}) (x_{kij} - \bar{X}_{kq}) \quad (11)$$

$$B_{pq} = \sum_{k=1}^g n_k (\bar{X}_{kp} - \bar{\bar{X}}_p) (\bar{X}_{kq} - \bar{\bar{X}}_q) \quad (12)$$

Of course, once the T matrix has been calculated, calculation of either the W or the B matrix will be sufficient to determine the final matrix via subtraction. Next, the



Discriminant Function Analysis, Fig. 3 Results of an example CVA of the Anderson-Fisher *Iris* data. (a) Ordination of the first 35 specimens of each of the three *Iris* species in the space of the two canonical variates. Note extent of the between-groups separation and the fact that, in this dataset, virtually all of the between-groups separation is subsumed on CV-1. (b) Confusion matrix for the set of training specimens based on the proximity for projected specimen dataset with respect to group centroids in the CV space. While group discrimination appears to be excellent for

this specimen set, such a result does not necessarily indicate a comparable level of performance will be achieved for non-training datasets. (c and d). scatterplots of sepal length/width (c) and petal length/width (d) datasets along with the back-projected positions, lengths, and orientations of the canonical variate vectors. Note that, despite these vectors exhibiting an orthogonal orientation relative to each other in the canonical variates space (a), they are not orthogonal in the space of the original variables

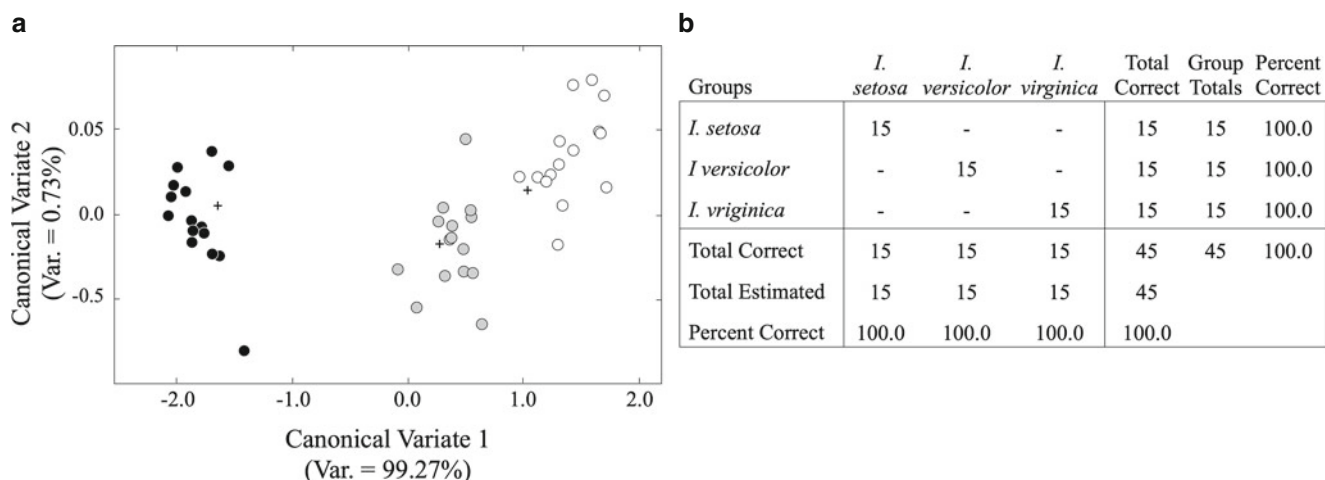
eigenvectors of the $W^{-1}B$ matrix are calculated. These eigenvectors will contain the weight loading (β) values of the set of $k-1$ canonical variates (= discriminant function) while their eigenvalues will represent an indication of how much separation between group centroids is accounted for by each canonical variate. For illustrative purposes, the set of original data values may be projected into the space formed by the canonical variates by premultiplying them by the matrix of eigenvectors. Campbell and Atchley (1981) have pointed out that CVA is, in essence a two-stage eigenanalysis-based rotation of the original data, the first of which treats the dataset as a whole and the second of which focuses on the group centroids with an intermediate eigenvalue scaling step that serves to equalize across-group variances.

Fisher (1936) originally used a set petal and sepal lengths collected from three different *Iris* species by Edgar Anderson

and published in 1935 in the Bulletin of the American *Iris* Society to illustrate his concept of discriminant function analysis. Since then the “Fisher” *Iris* data has been used by countless statisticians to develop and illustrate the use of new methods of quantitative data analysis. Figure 3 illustrates results obtained from a classic CVA of these data.

Performance Testing and Statistical Significance

As with Fisher’s LDA, the classic (but not the only) method of allocating unknown objects or samples (= sets of variable values) is to determine which group centroid in the CV space an object’s or sample’s projected position in the full CV space is closest to. Such calculations are summarized typically in tabular form (see Fig. 3b), in what is termed a “confusion



Discriminant Function Analysis, Fig. 4 Results of use of a validation set of 15 specimens from each of the Anderson-Fisher *Iris* species to test use of the CVs shown in Fig. 3 as tools for accurate species identification. (a) scatterplot of the projected positions of the validation-set specimens into the training-set CV space. (b) Confusion matrix for the

validation-set specimens based on proximity of their projected positions in the CV space with respect to the training-set group centroids. In this case all validation set specimens projected to positions closest to the training-set centroids (+) of the species to which they had been assigned a priori

matrix.” From this table a wide variety of descriptive summaries may be calculated that, together, provide an indication of the degree to which the discriminant analysis was successful in separating the subgroups (see Tharwat 2018). While informative in and of itself, a confusion matrix calculated from the training set should be interpreted with a fair amount of skepticism. Such results are valid for the training-set data but may not provide an accurate indication of the degree to which the estimated discriminant functions will be successful in achieving accurate identifications of objects that were not included in the training set. For this reason it is always preferable to randomly subdivide a sample of observations or measurements into a training dataset and a (usually smaller) testing or validation data set, so the performance of the discriminant functions in a real-world situation can be assessed objectively. Of course, this test implicitly assumes the total sample to hand represents a statistically valid sample of the populations from which the samples were drawn, and so underscores the absolute importance of sample quality in any discriminant analysis situation. Sets of discriminant functions that return excellent results for the training set, but much poorer results for the testing or validation set, are said to be “overtrained.” Figure 4 illustrates results of an independent validation set test of the “Fisher” *Iris* discriminant functions shown in Fig. 3.

If, as is often the case, an insufficient number of objects or samples are available to create both a training dataset and a testing or validation dataset, performance testing of a sort can still be carried out using the calculation-intensive “jackknife” strategy (Manly 2004). Under this procedure, a full set of n CVAs are carried out on datasets from which a different object or sample has been sequestered successively. This

sequestered object or sample is then identified by the resulting discriminant functions and the process repeated until all members of the training set have been sequestered and then re-identified independently by discrimination functions estimated on the non-sequestered data. Nevertheless, care must always be taken in interpreting jackknifed results as their use makes the statistical validity of the training set data even more of an issue. In all cases valid testing or validation datasets drawn independently from the populations of interest should be collected if at all possible and the direct analysis of such data preferred over the jackknife-estimation alternative.

A wide range of statistical tests are available for both Fisher’s LDA and CVA analysis in order to determine whether the separation between subgroup means along the LDA axis and/or within the CVA space is sufficient, relative to subgroup variances, to warrant rejection of the null hypothesis of no difference. These tests include Hotelling’s T^2 , the log likelihood ratio (ϕ), Wilk’s lambda (λ), Roy’s largest root (λ), Pillai’s trace (V), and the Lawes-Hotelling trace (U) (see Manly and Alberto 2017). Each of these tests assumes multivariate normality and equality of covariance matrixes across all groups/subgroups included in the sample. Advocates of each test claim theirs is robust to violations of these assumptions, but no guidance is provided typically as to how serious such violations might need to be before the test results become unreliable. If questions arise regarding the validity of these assumptions, perhaps the best alternative is to avail oneself of either the Monte Carlo or bootstrapping re-sampling strategies in order to create a valid empirical distribution “on the fly” for use in formal hypothesis testing (see Manly 2004). However, once again, this strategy raises

the issue of how representative of the parent population(s) sets of samples might be. Manly and Alberto (2017) also describe a Mahalanobis distance (D^2) test that may be used to evaluate the question of whether any individual set of measurements really belongs to the group to which it has been assigned statistically, via reference to the χ^2 distribution.

Conclusion

Many earth-science data analysts seem to assume that subgroup-level structure will be aligned with the major direction of variance within the pooled sample. In many cases this will be true and, when it is true, aspects of such structural relations can be revealed by pooled-sample ordination methods (e.g., PCA, PCoord, factor analysis). But this structural arrangement will not be true invariably. In order to visualize and/or test hypotheses of group or subgroup difference, it is necessary to employ some form of discriminant analysis irrespective of what pooled-sample ordination plots may, or may not, show. Fisher (two-group) and multi-group linear discriminant analysis represent two powerful approaches to the analysis of subgroup structure in multivariate datasets. Care must be taken when applying these methods to very high dimensional datasets and datasets characterized by small and/or unbalanced sample sizes. Nevertheless, these linear approaches represent traditional and (still) attractive alternatives to more powerful machine-learning methods (which often require enormous sample sizes) in many earth-science contexts.

Bibliography

- Altman N, Krzywinski M (2018) The curse(s) of dimensionality. *Nat Methods* 15:399–400
- Anderson E (1935) The irises of the Graspé peninsula. *Bull Am Iris Soc* 59:2–5
- Bartlett MS (1951) An inverse matrix adjustment arising in discriminant analysis. *Ann Math Stat* 22:107–111
- Campbell NA, Atchley WR (1981) The geometry of canonical variate analysis. *Syst Zool* 30:268–280
- Davis JC (2002) *Statistics and data analysis in geology*, 3rd edn. Wiley, New York
- Fisher RA (1936) The utilization of multiple measurements in taxonomic problems. *Ann Eugenics* 7:179–188
- Manly BFJ (2004) *Multivariate statistical methods: a primer*, 3rd edn. Chapman Hall/CRC, Boca Raton
- Manly BFJ, Alberto JAN (2017) *Multivariate statistical methods: a primer*, 4th edn. CRC Press, Boca Raton
- Rao CR (1952) *Advanced statistical methods in biometric research*. Wiley, New York
- Tharwat A (2018) Classification assessment methods. *Appl Comput Inform* 17:168–192

Doveton, John Holroyd

John C. Davis

Heinemann Oil GmbH, Baldwin City, KS, USA



Fig. 1 John Holroyd Doveton, courtesy of Prof. Doveton

Biography

John H. Doveton is a geologist best known for his pioneering use of computers in subsurface investigations, and especially in the quantitative interpretation of well logs. John has been instrumental in expanding the scope of well log analysis, with its narrow focus on porosity and permeability, into the broad expanse of petrophysics dealing with all aspects of composition and texture of rocks in the subsurface.

John was born in Croydon, Surrey, England in 1944. He studied geology at Oxford University, receiving BA and MA degrees. His doctoral studies (PhD, 1969) were done at the University of Edinburgh, employing the university's computer for analyses of sedimentological data. This experience led to a yearlong postdoctoral fellowship at the Kansas Geological Survey, and in turn to a position in Calgary with Mobil Oil Canada.

As a new recruit to the oil business, Mobil provided John an intense schooling in all aspects of applied petroleum geology that an exploration geologist might need, including the interpretation and analysis of well logs. This proved fortuitous, because John recognized that this was an area where his mathematical skills and computing background could be especially useful. Over the next several years, in addition to his well site duties, John prepared a number of quantitative studies for Mobil, which eventually resulted in a return to

Kansas for a cooperative study. This quickly evolved into a permanent position in the Kansas Geological Survey, where John remained for the next 45 years.

John's first years at Kansas were devoted to probabilistic assessment in petroleum exploration and investigation of a wide variety of multivariate procedures for their potential utility in geologic studies. He also developed materials for a graduate course in well log analysis for engineers and geologists, a subject he continued to teach throughout his career.

Experience with statistics lead John to regard wireline well logs as mathematical matrices that could be manipulated by computer. On this basis, he designed an innovative computer program for well log analysis that became preeminent in both the petroleum industry and academia. Subsequently, John wrote algorithms for petrophysics that could be implemented on ordinary spreadsheet programs. These were used to provide students with practical experience in his graduate and professional courses on log analysis.

In addition to his regular courses on log analysis at the University of Kansas, John served as visiting professor at numerous other universities, both in the USA and abroad. He also offered his short courses on petrophysics to oil companies and professional groups all over the world. His extensive course notes were provided in versions customized for each audience. John has published extensively on topics in mathematical geology; his resume includes hundreds of articles, reviews, and three textbooks on log analysis and petrophysics.

Bibliography

- Doveton JH (1986) Log analysis of subsurface geology. Wiley-Interscience, New York, 273 pp
- Doveton JH (1994) Geologic log analysis using computer methods. AAPG Applications in Geology, Tulsa, No. 2, 169 pp
- Doveton JH (2014) Principles of mathematical petrophysics. Oxford University Press, New York, 248 pp

Earth Surface Processes

Vikrant Jain¹, Ramendra Sahoo¹ and R. N. Singh²

¹Discipline of Earth Sciences, IIT Gandhinagar, Gandhinagar, India

²Discipline of Earth Sciences, Indian Institute of Technology, Gandhinagar, Palaj, Gandhinagar, India

Definition

Earth surface processes (ESP) shape earth's topography by adding or removing mass from earth's surface by erosion, transportation, and deposition. Records of earth surface process have been of interest as archives of past events like tectonic activity, climate change, and eustatic sea level change. Advent of new chronological tools led toward quantitative understanding of ESP, which further resulted in a new set of research questions about cause–effect relationship between landforms, processes, and controls. Hence, recently, mathematical concepts have been widely used to quantify and model these processes, which provide a comprehensible idea into the physics of these processes. The mathematical concepts also enable to model and infer the highly nonlinear cause–effect characteristics prevalent in ESP.

Earth Surface Processes (ESP) and Landforms

Earth's landscape is sculpted by both surface and subsurface components. These processes are influenced by the tectonic motion, base-level change, and climate change, which supply and deduct mass from the surface and create various patterns on the earth's surface. ESP are driven primarily by some form of fluid (i.e., air, glacier, and water) motion and operate at a decadal to millennial time scales. Recently, anthropogenic activities have significantly affected the ESP. The role of various controls on ESP through different thresholds leads

to a nonlinear and complex response of geomorphic systems to external controls.

Complexities in geomorphic systems are also caused by (1) the action of various drivers, i.e., wind, glacial, fluvial, hillslopes, or coastal, and (2) their operations and interactions at different spatiotemporal scales. Understanding ESP and its interactions at cross-over of scales is the key to unravel the complexities in landscape patterns and its evolutionary trajectory. Interestingly, mathematical concepts provide these tools to study and understand the across-scale linkages and inherent complexities present in ESP. For example, basic equations of sediment grain transport are being applied to study evolutionary trajectories of landscapes in different scenarios of climate change and tectonics. Hence, the application of mathematical concepts in ESP studies is providing new insight to earth surface science. This entry focuses on the advancement and application of mathematical concepts in ESP studies.

Mathematical Concepts to Study Various ESP and Landforms

Mathematical approaches for ESP range from modeling the dynamic processes resulting in the complex landscape in terms of partial differential equations (PDE) to the representation of the earth surface forms in terms of known mathematical functions. PDE models are derived from the conservation of mass and specific transport laws relating mass flux to physical characteristics of topography (such as gradient) (Tucker and Hancock 2010). In the following section, we compile a few of the major mathematical concepts, which are being implemented in modern-day ESP studies.

Differential Equation

The general governing equation for the spatiotemporal evolution of earth surface topography in a fluvial domain is a conservation equation:

$$\frac{\partial h}{\partial t} = -\nabla \cdot q_s + U \quad (1)$$

Here, h is elevation of topography, q_s is the sediment flux across a cross-section, and U is a source (or sink) term. Sediment flux is related to elevation through the following constitutive relation:

$$q_s = uh - D\nabla h \quad (2)$$

where u is the advection velocity (celerity), and D is the diffusivity. Correspondingly, the first term is the advective flux and second is the diffusive flux. Governing equation for elevation is obtained by combining Eqs. 1 and 2 as an advection-diffusion equation:

$$\frac{\partial h}{\partial t} = -\nabla \cdot (uh) + \nabla \cdot (D\nabla h) + U \quad (3)$$

To obtain a unique solution, we need a set of boundary conditions (BC) and an initial condition (IC). While the IC can take any arbitrary form, owing to the specifics of the system and the problem, generally we have for the BC: (1) Dirichlet (fixed boundary elevation) (2) Neumann (fixed flux), and (3) Robin condition (a combination of flux and elevation as fixed). The solution to this PDE (Eq. 3) gives us a transient form of elevation at any point of time and space. There are two special cases of the general governing equation, (1) hillslope evolution process, when diffusion is important and (2) bedrock channel evolution, when advection is important, as discussed below.

Hillslope Evolution

Evolution of a hillslope can be mathematically represented as a linear diffusion process, and the governing equation in 1-D is:

$$\frac{\partial h}{\partial t} = U(x, t) + D \frac{\partial^2 h}{\partial x^2} \quad (4)$$

The transient form of the diffusion equation can be solved using the method of separation of variables (Fourier series method) or integral transform methods. The transient solution can then be used for morphological dating of geomorphic landforms. For instance, a general solution for the evolution of the gradient of a fault scarp (in the absence of uplift) is given as (Pelletier 2008):

$$\frac{\partial h}{\partial x} = \frac{\alpha - b}{2} \left(\operatorname{erf} \left(\frac{x + a/(\alpha - b)}{\sqrt{4Dt}} \right) - \operatorname{erf} \left(\frac{x + a/(\alpha - b)}{\sqrt{4Dt}} \right) \right) + b \quad (5)$$

where α is the initial scarp slope, $2a$ the scarp offset, and b is the regional slope, and erf is error function. The product of the

age (t) and diffusivity (D), morphologic age (Dt), is related to the length scale of diffusion and dictates the development of the final form of the scarp. With a calibrated value of D , we can use the morphologic age to calculate the numerical age of a fault scarp and slip rate along the fault.

Further, in the case of a nonlinear transport, the governing equation will be modified to (Roering et al. 1999):

$$\frac{\partial h}{\partial t} = U - \frac{\partial}{\partial x} \left(\frac{-D(\partial h / \partial x)}{1 - [(\partial h / \partial x) / S_c]^2} \right) \quad (6)$$

Here, S_c is the critical hillslope gradient at which down-slope sediment fluxes become infinite. Nonlinearity can also arise from considering nonlocal effects of topographic gradient on flux, as opposed to the locally defined sediment flux as in the case for Eqs. 4 and 6. This nonlocal sediment flux can be expressed in the form of flux at a given point as a weighted average of the upslope topographic attributes:

$$q_s(x) = -K \int_0^x g(l) \nabla h(x-l) dl \quad (7)$$

where $g(l)$ is a kernel performing a weighted average of local gradients upslope of the point of interest x as they contribute to the sediment flux at the point x . It has been shown (Foufoula-Georgiou et al. 2010) that when the weighting function $g(l)$ has no characteristic length scale, i.e., $g(l) \sim l^{1-\infty}$, Eq. 7 takes the form of a fractional derivative:

$$q_s(x) = -K \nabla^{\infty-1} h(x) \quad (8)$$

The resulting governing equation for topography thus becomes

$$\frac{\partial h}{\partial t} = U + K \nabla^{\infty} h \quad (9)$$

The transient solution to Eqs. 6 and 9 can be achieved using numerical methods (Foufoula-Georgiou et al. 2010; Perron 2011).

Besides the fluvial domain, the diffusion equation has also been used to model glacial erosion (Herman et al. 2018) and aeolian transport (Pelletier 2018). This highlights the generality and adaptability of mathematical concepts to carryout physics-based modeling across multiple domains of ESP studies.

Bedrock Channel Evolution

Bedrock channel incision is a fundamental process for understanding the trajectories of landscape evolution. It is commonly represented by the area-based Stream Power Incision

Model (SPIM), explicitly representing the universally observed inverse power relation between channel slope and drainage area (Howard et al. 1994), and can be presented as follows:

$$\frac{\partial h}{\partial t} = U - K'A^{\beta'}S^{\zeta} \quad (10)$$

If we substitute area (A) with distance from channel head (x) using Hack's Law (section "Power-Law Function"), the governing equation transforms to

$$\frac{\partial h}{\partial t} = U - Kx^{\beta}S^{\zeta} \quad (11)$$

where $\beta = \beta'/H$ and $K = K'/K_H^{\beta'/H}$. K' and K are the erodibility coefficient, while β' (β) and ζ are area (distance) and slope exponents, respectively. H is the Hack's exponent and K_H is the Hack's coefficient. Equation 11 can be compared to a 1-D advection equation of the following form:

$$\frac{\partial h}{\partial t} = U(x, t) - u(x, t) \frac{\partial h}{\partial x} \quad (12)$$

which suggests that Eq. 11 is basically an advection equation with a celerity as $Kx^{\beta}S^{\zeta} - 1$. These equations of channel incision process at reach-scale further become the fundamental equations in Landscape evaluation models (LEMs) to understand landscape dynamics in response to external controls.

The general form of transient solution to Eq. 11 can be achieved using method of characteristics (Royden and Taylor 2013, Appendix A) and can be written in the following integral form (for $\zeta = 1$):

$$h(t, x) = \int_{t-\tau(x)}^t U[t', \tau^{-1}(\tau(x) - t + t')] dt' \quad (13)$$

where t' is an integration variable. $\tau(x)$ is the response time for any perturbation (e.g., surface uplift) and can be defined as:

$$\tau(x) = \int_0^x \frac{dx'}{Kx'^m} \quad (14)$$

where x' is another integration variable. Equations 13 and 14 can be used to get the tectonic or climate information from river long profile (Goren 2016) using linear inversion technique.

Exponential Function

The exponential equation has great utility in mathematical modeling since both its integration and differentiation gives the same original form. In ESP studies, elevation (h) of topography is often related to horizontal distance (x) as:

$$h(x) = h_0 \exp(-ax) \quad (15)$$

where h_0 is the elevation of topographic ridge and a is a decay constant. The exponential function is also a solution of the following differential equation for eroding topography:

$$\frac{dh}{dx} = c - ah \quad (16)$$

River long profile has been modeled by exponential functions to infer the control of sediment transport process, geology, and climate on long profile evolution (Sonam and Jain 2018). Exponential functions are also prevalent in the morphometric analyses of the drainage network (Strahler 1957).

The reciprocal of the exponential function is the logarithmic function. The logarithmic function is famous in the field of ESP in the form of Hack profile (Hack 1973), which can be derived by assuming a gradual downstream drop in elevation along a river profile:

$$h = h_u + c \log\left(\frac{x_u}{x}\right) \quad (17)$$

where x_u and h_u are the distance of channel head from drainage divide and channel head elevation, respectively. c is a proportionality constant.

Power-Law Function

Power-law successfully captures the very essence of the structure of a natural system's nonlinear behavior. It is widely used to characterize the statistical property of any natural system and can be regarded as a first-hand proof of self-organized criticality (Van De Wiel and Coulthard 2010), which widely governs the dynamic processes in nature.

Power-law is also a quintessential example of a fat-tailed distribution, which can be observed for most of the extreme phenomena, such as frequency of earthquake (Bak and Tang 1989) or flood (Malamud and Turcotte 2006), among other natural hazards. Power-law functions are very common in fluvial geomorphology. They arise in drainage basin characteristics, such as Hack's law (Hack 1957), which relates the upstream area and length of a basin and indicates the evolutionary pattern of a river basin:

$$L = K_H A^H \quad (18)$$

One of the other important examples in fluvial geomorphology is the relation between slope and upstream area of an equilibrated landscape (Wobus et al. 2006):

$$S = k_s A^{-\theta} \quad (19)$$

where k_s and θ are channel steepness and concavity, respectively.

Fractals in ESP

Coupled interactions of governing factors lead to complex natural forms, which can be mathematically represented as fractal objects, which shows statistical self-similarity at different scales of observation. These fractal objects or forms can then be characterized with fractal dimension. Estimation of the fractal dimension of natural objects utilizes the power-law dependence of some basic physical properties of the objects (e.g., length of a coastline) on the scale.

$$l \propto r^{1-d} \quad (20)$$

where l is the total length of a fractal object (coastline, Mandelbrot 1967) and r is the scale of observation. The slope of this line on a log-log scale gives us the fractal dimension, d . Recent geomorphological studies have related the fractal dimension of the topography to governing controls, such as tectonics, climate, and lithology (e.g., Sahoo et al. 2020).

Multifractals, on the other hand, can be regarded as objects comprising a union of subsets, each characterized by different scaling exponent (Berkowitz and Hadad 1997). An appropriate description of multifractal nature of sets of any measure is the singularity spectrum $f(\alpha)$ (Halsey et al. 1986). The width of this spectrum has been used in literature to interpret the governing factors in landscape and drainage network evolution (e.g., Dutta 2017).

Matrices

The primary spatial data often used in any ESP analysis is the digital elevation model (DEM), which is a synthetic model of the earth's landscape in the form of matrices. Each grid (pixel) of the DEM contains the average elevation of the corresponding ground pixel of a certain area depending on the resolution of the DEM product. ESP studies use DEMs to estimate transport of flux across the landscape, for which spatial relation of a grid to its neighborhood is of primary importance. The matrix form of DEM provides the opportunity to avail the powerful techniques of linear algebra for modeling. For instance, the upstream contributing area is a proxy for the flux at any cell of the DEM and can be estimated using the following relation in matrix form (Schwanghart and Kuhn 2010):

$$a = [I - M^T]^{-1} w(0) \quad (21)$$

where a is the upstream contributing area vector, I is the identity matrix, M^T is the transpose of the flow direction or transfer matrix, which is slope-dependent, and $w(0)$ is the initial amount of water in each cell. Similarly, set theory-based morphological operations (e.g., erosion, dilation), which contains simple linear operation between images (2-D sets) are widely used for feature extraction from DEM. For example, river networks can be extracted from DEM using skeletonization technique, and can be mathematically defined as follows (Sagar et al. 2000):

$$CH(I) = \bigcup_{k=1}^{i-1} \left[\bigcup_{j=0}^J CH_m(A_k) / A_{k+1} \right] \quad (22a)$$

$$CH_j(A) = [A \ominus B_j] / [(A \ominus B_j) \ominus B] \oplus B \quad (22b)$$

where $CH(I)$ represents the river channel network corresponding to gray scale image (DEM), I, j is the size of the structuring element B , k is the number of binary sets, A , derived from I using thresholding operation. Sagar et al. (2000) can be referred for the definitions of the binary operations used in Eqs. 22a and 22b.

Further, in ESP studies, we have a huge set of output, in terms of elevation, sediment load, the chemical signature of sediment, etc. Hence, we mostly deal with inverse problems to get an estimate of model parameters of the governing equations of such geomorphic entities, such as source, diffusivity, and BC. These inverse problems also make use of principles of linear algebra. The governing equations for any ESP problem can be generally written as a system of linear equations, which in matrix form generally renders itself as follows:

$$d = Am \quad (23)$$

Here, d is a data vector to study effects and A represents the causal connection between cause (m) and effect (d). In inverse problems, we solve this system of linear equations. The model parameters (m) can be tectonic uplift rate in the context of landscape evolution (Goren et al. 2014), chemical erosion rate in the context of sediment source fingerprinting (Tripathy and Singh 2010), and so on. We can minimize both the error (inherent in data) and the length of the solution vectors (penalty) for a near-singular matrix A . With an apriori set of values for the parameters (m_p) and a penalty factor (δ^2), the solution becomes (Tarantola 2005):

$$m = m_p + (AA^T + \delta^2 I)^{-1} A^T (d - Am_p) \quad (24)$$

Statistical Approaches in ESP

Statistics in ESP is primarily used for regression analysis, sensitivity analysis, uncertainties estimation, among other types of quantification of geomorphological processes. Statistical measures (e.g., central measures or dispersion) are also often used to characterize ESP, as well as the landscape features. Additionally, in modern-day landscape evolution models, there is scope to use stochastic time-varying forcings (climate, tectonics), which can be drawn from a characteristic probability density function (pdf) for climate forcings [rain-fall intensity (P), storm duration (T_r), and interstorm period (T_b)] (Tucker and Bras 2000):

$$f(P) = \frac{1}{\langle P \rangle} e^{-P/\langle P \rangle} \quad (25a)$$

$$f(T_r) = \frac{1}{\langle T_r \rangle} e^{-T_r/\langle T_r \rangle} \quad (25b)$$

$$f(T_b) = \frac{1}{\langle T_b \rangle} e^{-T_b/\langle T_b \rangle} \quad (25c)$$

where $\langle - \rangle$ represents long-term average of a stochastic variable. Central limit theorem (CLT) constitutes a fundamental component of large-sample statistics. State-of-art resampling techniques, such as Monte-Carlo or bootstrap techniques, make use of the CLT and are widely used for error propagation and uncertainty analysis (Blaauw 2010) in ESP models.

Numerical Methods in ESP

Numerical methods enable us to solve evolutionary problems, usually comprising of multiple differential equations, as a set of algebraic equations for a discrete set of points in space and time. These methods discretize the time (and space in case of PDE) over which we try to solve our problems into different nodes. Then it uses a numerical approximation of differentials (e.g., Newton's Difference quotient) to solve the governing equations for each of the nodes.

Finite Difference (FD) method is one of the easiest to use in family of techniques to solve any differential equation. For instance, in order to solve a linear advection-diffusion equation (Eq. 3 in 1-D), we can use an explicit central difference scheme as follows:

$$\frac{h_i^{j+1} - h_i^j}{\Delta t} = D \frac{h_{i+1}^j - 2h_i^j + h_{i-1}^j}{\Delta x^2} - K \frac{h_{i+1}^j - h_i^j}{\Delta x} \quad (26)$$

where h^{j+1} depends on h^j . Here, i and j are the spatial and temporal indices, respectively. Δx and Δt represent the grid spacing and the time step of the numerical scheme. There is also an implicit scheme in FD method with a distinct advantage of unconditional stability (Perron 2011).

Finite Volume (FV) method is another technique being widely used in landscape evolution models (Tucker et al. 2001). In the FV method, the whole domain is divided into cells, and the basic continuity equation (without any source or uplift for our example) is integrated over the surface area of a cell as follows:

$$\iint_A \frac{\partial h}{\partial t} dA = - \iint_a (\nabla \cdot q_s) dA \quad (27)$$

Using Gauss Divergence theorem and integrating the first term as an area-average, this equation transforms into

$$A_i \frac{\partial \bar{h}_i}{\partial t} = \oint_p (q_s \cdot \hat{n}) dp \quad (28)$$

Here, A_i is the area of a cell, p represents position along the perimeter of the cell, and $\hat{n}$ is a unit vector normal to the perimeter and pointing outward and $\bar{h}_i$ is the average elevation of the cell. Both time derivative and surface integrals for a diffusion problem are then approximated as (Tucker et al. 2001):

$$\frac{h_i^{j+1} - h_i^j}{\Delta t} = \frac{D}{A_i} \sum_{\alpha=1}^{np} \lambda_{i\alpha} \frac{h_\alpha^j - h_i^j}{L_{i\alpha}} \quad (29)$$

where A_i is the area of a Voronoi cell, the sum is over all the np points adjacent to cell i , the ratio inside the sum represents the slope of a segment (between i and one of the neighbors α), and $\lambda_{i\alpha}$ is the length of the corresponding Voronoi edge. The FV numerical scheme further provides the base to LANDLAB, a python-based landscape evolution model, which is being used extensively in teaching and research (Hobley et al. 2017).

Machine Learning Application in ESP

ESP community is witnessing a paradigm shift in the past couple of decades caused by the big data pool generated from the rapid advancement in the field of remote sensing techniques. The rapidly evolving field of machine learning (ML) is tipped to play a vital role in the effort to extract useful information from the big data. ML application in ESP studies have been applied most notably in landslide hazard mapping, soil classification, and landform classification (e.g., Bergen et al. 2019). For instance, artificial neural network (ANN) is widely used for landslide susceptibility mapping (Bragagnolo et al. 2020). The mathematics behind ML techniques primarily constitutes linear algebra and differential calculus.

A unique challenge in ML for ESP studies is that ESP is governed by physical laws, and the resulting events are highly correlated at different scales in both space and time. A recent review by Reichstein et al. (2019) highlighted this challenge

and suggested that these novel aspects of ESP problems should be used as part of deep learning to gain further process understanding of Earth system science problems.

Conclusions

Mathematical concepts provide important tools to analyze the process interactions and feedback at different scales across various geomorphic agents. It also helps to integrate ESPs driven by different geomorphic agents. Application of mathematical methodologies has not only quantitatively expressed the patterns of landscape/landforms but also provided new insights about cause–effect relationships. These methods are now fundamental to understand the complex and nonlinear behavior of geomorphic systems.

Bibliography

- Bak P, Tang C (1989) Earthquakes as a self-organised critical phenomenon. *J Geophys Res Solid Earth* 94(B11):15635–15637
- Bergen KJ, Johnson PA, Maarten V, Beroza GC (2019) Machine learning for data-driven discovery in solid earth geoscience. *Science* 363(6433):eaau0323
- Berkowitz B, Hadad A (1997) Fractal and multifractal measures of natural and synthetic fracture networks. *J Geophys Res Solid Earth* 102(B6):12205–12218
- Blaauw M (2010) Methods and code for ‘classical’ age-modelling of radiocarbon sequences. *Quat Geochronol* 5(5):512–518
- Bragagnolo L, da Silva RV, Grzybowski JMV (2020) Artificial neural network ensembles applied to the mapping of landslide susceptibility. *Catena* 184:104240
- Dutta S (2017) Decoding the morphological differences between Himalayan glacial and fluvial landscapes using multifractal analysis. *Sci Rep* 7(1):1–8
- Foufoula-Georgiou E, Ganti V, Dietrich WE (2010) A nonlocal theory of sediment transport on hillslopes. *J Geophys Res Earth* 115(F2):F00A16
- Goren L (2016) A theoretical model for fluvial channel response time during time-dependent climatic and tectonic forcing and its inverse applications. *Geophys Res Lett* 43(20):10–753
- Goren L, Fox M, Willett SD (2014) Tectonics from fluvial topography using formal linear inversion: theory and applications to the Inyo Mountains, California. *J Geophys Res Earth* 119(8):1651–1681
- Hack JT (1957) Studies of longitudinal stream profiles in Virginia and Maryland, vol 294. US Government Printing Office, Washington
- Hack JT (1973) Stream-profile analysis and stream-gradient index. *J Res US Geol Surv* 1(4):421–429
- Halsey TC, Jensen MH, Kadanoff LP, Procaccia I, Shraiman BI (1986) Fractal measures and their singularities: the characterisation of strange sets. *Phys Rev A* 33(2):1141–1151
- Herman F, Braun J, Deal E, Prasicek G (2018) The response time of glacial erosion. *J Geophys Res Earth* 123(4):801–817
- Hobley D, Adams JM, Siddhartha Nudurupati S, Hutton EW, Gasparini NM, Istanbuloglu E, Tucker GE (2017) Creative computing with Landlab: an open-source toolkit for building, coupling, and exploring two-dimensional numerical models of earth-surface dynamics. *Earth Surf Dyn* 5:21–46
- Howard AD, Dietrich WE, Seidl MA (1994) Modeling fluvial erosion on regional to continental scales. *J Geophys Res Solid Earth* 99(B7):13971–13986
- Malamud BD, Turcotte DL (2006) The applicability of power-law frequency statistics to floods. *J Hydrol* 322(1–4):168–180
- Mandelbrot B (1967) How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science* 156:636–638
- Pelletier J (2008) Quantitative modeling of earth surface processes. Cambridge University Press, Cambridge
- Pelletier JD (2018) Controls on yardang development and morphology: 2. Numerical modeling. *J Geophys Res Earth* 123(4):723–743
- Perron JT (2011) Numerical methods for nonlinear hillslope transport laws. *J Geophys Res Earth* 116(F2):F02021
- Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N (2019) Deep learning and process understanding for data-driven earth system science. *Nature* 566(7743):195–204
- Roering JJ, Kirchner JW, Dietrich WE (1999) Evidence for nonlinear, diffusive sediment transport on hillslopes and implications for landscape morphology. *Water Resour Res* 35(3):853–870
- Royden L, Taylor PJ (2013) Solutions of the stream power equation and application to the evolution of river longitudinal profiles. *J Geophys Res Earth* 118(2):497–518
- Sagar BSD, Venu M, Srinivas D (2000) Morphological operators to extract channel networks from digital elevation models. *Int J Remote Sens* 21(1):21–29
- Sahoo R, Singh RN, Jain V (2020) Process inference from topographic fractal characteristics in the tectonically active northwest Himalaya, India. *Earth Surf Process Landf*. <https://doi.org/10.1002/esp.4984>
- Schwanghart W, Kuhn NJ (2010) TopoToolbox: a set of Matlab functions for topographic analysis. *Environ Model Softw* 25(6):770–781
- Sonam, Jain V (2018) Geomorphic effectiveness of a long profile shape and the role of inherent geological controls in the Himalayan hinterland area of the Ganga River basin, India. *Geomorphology* 304: 15–29
- Strahler AN (1957) Quantitative analysis of watershed geomorphology. *EOS Trans Am Geophys Union* 38(6):913–920
- Tarantola A (2005) Inverse problem theory and methods for model parameter estimation. *Soc Ind Appl Math*. <https://doi.org/10.1137/1.9780898717921>
- Tripathy GR, Singh SK (2010) Chemical erosion rates of river basins of the Ganga system in the Himalaya: reanalysis based on inversion of dissolved major ions, Sr, and ⁸⁷Sr/⁸⁶Sr. *Geochem Geophys Geosyst* 11(3)
- Tucker GE, Bras RL (2000) A stochastic approach to modeling the role of rainfall variability in drainage basin evolution. *Water Resour Res* 36(7):1953–1964
- Tucker GE, Hancock GR (2010) Modelling landscape evolution. *Earth Surf Process Landf* 35(1):28–50
- Tucker G, Lancaster S, Gasparini N, Bras R (2001) The channel-hillslope integrated landscape development model (CHILD). In: *Landscape erosion and evolution modeling*. Springer, Boston, pp 349–388
- Van De Wiel MJ, Coulthard TJ (2010) Self-organised criticality in river basins: challenging sedimentary records of environmental change. *Geology* 38(1):87–90
- Wobus C, Whipple KX, Kirby E, Snyder N, Johnson J, Spyropoulou K, Crosby B, Sheehan D (2006) Tectonics from topography: procedures, promise, and pitfalls. *Spec Pap - Geol Soc Am* 398:55–74

Earth System Science

Gang Liu^{1,2} and Hongfei Zhang^{1,3}

¹State Key Laboratory of Biogeology and Environmental Geology, Wuhan, Hubei, China

²School of Computer Science, China University of Geosciences, Wuhan, Hubei, China

³School of Earth Science, China University of Geosciences, Wuhan, Hubei, China

Definition

The Earth system refers to an organic whole composed of the atmosphere, hydrosphere, geosphere (lithosphere, mantle and Earth core), and biosphere (including human beings). Earth system science (ESS) is to study the mechanism of the interaction between these subsystems of the Earth system, the law of the change of the Earth system and the mechanism of controlling these changes, so as to establish a scientific basis for the prediction of global environmental change and provide a basis for the scientific management of the Earth system. The space scope of Earth system science research ranges from the center of the Earth to the outer space of the Earth, and the time scale ranges from hundreds of years to millions of years.

Goal of Earth System Science

The research purpose of Earth system science is to understand the processes involved in the Earth system, the connections, and interactions between various components, maintain sufficient supply of natural resources, reduce geological disasters, regulate global environmental changes and minimize hazards, and obtain a scientific understanding of the whole Earth system on a global scale.

Human beings have become one of the forces that change the topography, geochemistry, and life evolution on a global scale. Taking human beings as an independent sphere, we should strengthen the research on the interaction between geosphere, biosphere, and anthroposphere and reveal the interaction between the Earth system, including the Earth's spheres. This marks a major change in the research object, guiding theory and research methods of Earth Science, paying more attention to the research of man-Earth relationship, highlighting the characteristics and evolution of the Earth system, adjusting man-Earth relationship, and promoting man-Earth harmony (Steffen 2004).

History of Earth System Science

Earth system science has experienced four stages: the embryonic stage before the 1970s, the discipline establishment stage in the 1980s, the global stage in the early twenty-first century, and the contemporary development after 2015.

In the middle of the twentieth century, geoscience began to be internationalized. Marked by the International Geophysical Year (IGY) from 1957 to 1958, this unprecedented research integrated the research works of 67 countries and improved the comprehensive understanding of the Earth's sphere system. The important achievement of this stage is the GAIA hypothesis put forward by James Lovelock in 1972, which has produced a new way of thinking about the Earth: recognizing the important impact of biota on the global environment and the importance of correlation and mutual feedback among the main components of the Earth system.

The NASA's Earth System Science Committee was established in 1983 (National Research Council 1986) and published the report "Earth System Science" in 1988 (NASA Advisory Council 1988). It formally put forward the concept of Earth system science and promoted the development of ESS through observation, modeling, and process research and supported Earth observation system satellites and related research. The research project led by NASA had developed the Bretherton structure diagram, which was the first system dynamics display of the Earth system. Through a series of complex external conditions and feedback, it coupled the physical climate system with the biogeochemical cycle, visualizing the physical, chemical, and biological processes connecting the interaction of various components of the Earth system and to recognize that human activities were an important driving force for system change (Shikazono 2009; Cornell et al. 2012).

From the 1990s to the early twenty-first century, the global research system supported by international programs such as World Climate Research Programme (WCRP), International Geosphere Biosphere Program (IGBP), International Biodiversity Program (DIVERSITAS), and International Human Dimensions Program on Global Environmental Change (IHDP), provided a "workspace" for the cooperation of scientists from different disciplines around the world, which was very important to the development of ESS.

After 2015, ESS has been relatively mature and began to carry out major institutional restructuring and interdisciplinary research on the basis of higher-level integration. IGBP, IHDP, and DIVERSITAS were merged into the new plan "Future Earth" in 2015, aiming to accelerate the transformation to global sustainability research through innovation (Steffen et al. 2020; Castree 2021).

Research Tasks of Earth System Science

1. Observation of Earth system phenomena and accumulation of Earth system multisource and multimodal data.
2. Analyze and interpret the observation data, and establish quantitative relationship among physical phenomena, chemical phenomena, and biological activities.
3. Establish Earth system conceptual model and numerical model.
4. Verify the model and use it to make predictions and forecasts of future trends.

New Concepts in Earth System Science

Since the beginning of the twenty-first century, with the development of Earth observation and detection technology, the concept of Earth system has been strengthened, which has promoted the intersection and integration of multi-disciplines. The development and evolution of Earth system science have also spawned a series of new concepts. The concept of geological diversity is formed on the macro scale, the concept of Anthropocene is formed on the time scale, the concept of critical zone of the Earth is formed on the spatial scale, and the concept of tipping element is formed on the element scale. These concepts promote the development of Earth system science in depth.

1. Geological diversity: one of the important theoretical cornerstones of Earth system science. Geological diversity refers to the diversity of geological (rocks, minerals, fossils), geomorphology (topography, physical process), soil, hydrology, sediment, and other element characteristics and the combination and process of elements. Geological diversity, biological diversity, climate diversity, and marine diversity constitute the basis of sustainable development. Geological diversity and biodiversity constitute natural capital, bring ecosystem services and geological system services to mankind, and jointly supply and maintain the whole natural system.
2. Anthropocene: the key time scale of Earth system science research. In view of the fact that human activities have become a very active geological force, Dutch atmospheric chemist Paul Kruzen and ecologist Eugene Stoermer jointly put forward the concept of “Anthropocene” in 2000 and regarded it as a geological Holocene period that was replaced by Pleistocene and Holocene. Once the concept of Anthropocene was put forward, it was widely recognized by the Earth science community. Geoscientists have discussed the starting point of the Anthropocene.

There are four main understandings: (a) the lower limit of the original Holocene; (b) in the early Holocene, the lower limit was pushed to thousands of years ago; (c) the beginning of the industrial revolution, that is, the second half of the eighteenth century; (d) the middle of the twentieth century, that is, the time when mankind’s first atomic bomb exploded in 1945. On May 21, 2018, after voting, it was determined that the starting point of Anthropocene was set in the middle of the twentieth century.

3. Earth critical zone: the entry point of Earth system science research. The concept of Earth critical zone was first seen in the strategic report of the National Research Council on Opportunities for Basic Research in Geosciences in 2001. It refers to the zone where the near surface environment in the shallow lithosphere intersects and interacts with the hydrosphere, atmosphere, and biosphere. It takes the vegetation canopy and deep underground aquifer as the upper and lower interfaces (affected by altitude and landform). It is also a composite space for the interaction between chemical processes of Earth materials and biological activities. Earth critical zone requires comprehensive observation and research of geology, geochemistry, hydrology, pedology, biology, geomorphology, and other disciplines, so as to understand the global and millennium span human adaptability to environmental changes. The National Science Foundation of the United States had set up 10 observation stations, whose observation contents include external geological processes in different evolution stages, weathering crust and soil formation, biological and microbial activities, aquifer and surface water, energy and material processes, landscape evolution, and water and carbon cycle. The concept of Earth critical zone has created conditions for the Earth system to scientifically solve major geological events in local areas (important water resources areas, geological fragile areas and important urban areas). Through the geological mapping, monitoring, modeling, and prediction of the Earth critical zone, the formation and evolution law of the Earth critical zone are revealed, and then the protection and restoration scheme of the ecological environment is put forward.
4. Tipping elements: nonlinear characteristics of the Earth system. Tipping elements refer to subsystems or system elements that can change fundamentally in the Earth system, such as global climate warming and environmental change caused by human activities. Arctic sea ice and Greenland ice sheet are critical elements. Tipping element is an important concept to describe the characteristics of the Earth system, which shows that the Earth system has strong characteristics of nonlinearity, complexity, and openness (NASEM 2021).

Information Processing and Expression in Earth System Science

The Earth is a complex giant system. The span of time and space varies greatly. Only by using high-speed computers can we simulate some unobservable phenomena. Using data mining technology, we will be able to better understand and analyze the observed massive data and find out the laws and knowledge. Scientific computing can break through the limitations of experimental and theoretical science. Modeling and simulation can explore the collected data about the Earth more deeply.

In the era of big data and artificial intelligence, Earth system data, information system platform, and Earth system numerical model are the infrastructure of Earth system science research. Digital Earth is regarded as an effective carrier and platform of Earth big data, which can provide a series of technologies, methods, and integrated environment for Earth system scientific research (Reichstein et al. 2019; Irrgang et al. 2021).

Digital Earth is the digital expression of Earth System Science, which will be conducive to the development from the qualitative description of natural phenomena to quantification. Digital Earth takes the Earth system as the prototype, the Earth coordinate as the reference system, and the Earth system science, information science, and computing science as the theoretical basis and realizes the integration of a series of prototypes, mathematical models, physical models, mechanical models, information models, and computer models at different levels. At the same time, with the support of Earth observation and network new technology, it builds the fusion of multi-resolution, massive data, and various data and uses multimedia and simulation virtual technology for multidimensional expression based on a digital, networked, intelligent, and visual technology system, so as to understand the information process involved in the whole Earth system and pay special attention to the law of information connection and interaction between various spheres of the Earth system. The main research contents include digital Earth prototype, theoretical basis of Earth system field, digital Earth material model, digital geodynamics model, digital Earth mathematical model, digital Earth information model, digital Earth information acquisition technology, digital geospatial information infrastructure, and digital Earth engineering.

Next Generation of Earth System Science

In May 2021, the European Commission announced that it would fund two new Earth system model (ESM) development projects through the “Horizon 2020” program to further improve the expression of the Earth system model on key processes and coupling systems and improve the simulation, prediction, and prediction ability of climate models.

In 2021, National Academies of Sciences, Engineering, and Medicine, USA, released the report “Next Generation Earth Systems Science at the National Science Foundation”. The main contents include (1) at present, human technologies and activities in economic development and resource utilization have begun to compete with and even exceed the scale of natural processes in driving the process of the Earth system. Our understanding of some key issues has been slow, such as the rapid changes in the Earth’s natural systems, the extent of human impact on them, the impact of the Earth’s natural systems on the sustainability and resilience of humans and ecosystems, and the effectiveness of different ways to meet these challenges; (2) The relationship between nature and society is the core of many complex problems faced by society and the Earth. It is necessary to recognize the multi-directional relationship between these natural and social subsystems in the context of science and technology. Integrated research provides a method to develop comprehensive science, construct research problems from the perspective of social problems, and integrate the knowledge and methods from nature, society, computing, and engineering science from the beginning; (3) Infrastructure such as observation, computing, and modeling must work together to support the integration of Earth system science. Observation and monitoring reveal changes in the Earth system. The model representing the natural and social system processes and their interactions of the whole Earth system absorbs data from a variety of sources. Computing provides a framework for integrating complex parts of Earth system science, supporting data collection and analysis, generating predictions, and interpreting model results (NASSEM, 2021).

Conclusions

Earth system science is a new view of the earth and a revolutionary change of human view of the earth. This change is reflected in the following five aspects:

1. The Earth is an organic system, which can be understood from the spheres and its interactions of the earth.
2. The earth is a symbiotic community of human and nature, full of complex changes.
3. Humans are becoming the driving force of global change.
4. Mission on sustainable development: it is impossible for human beings to transcend the limitation of the Earth’s living space and available resources, and let alone change the law of the earth system itself. Mankind is duty-bound to manage the earth.
5. New methods and technologies based on big data and artificial intelligence provide a new research paradigm for earth system science.

Bibliography

- National Research Council (1986) Earth system science: overview: a program for global change. The National Academies Press, Washington, DC. <https://doi.org/10.17226/19210>
- NASA Advisory Council (1988) Earth system science: a closer view. Report of the NASA earth system Sciences committee (F. Bretherton, chair). National Aeronautics and Space Agency, Washington DC
- Steffen W, Sanderson A, Tyson PD et al (2004) Global change and the earth system: a planet under pressure. Springer
- Shikazono N (2009) Introduction to earth and planetary system science. Springer
- Cornell S, Prentice IC, House J, Downy C (2012) Understanding the earth system: global change science for application. Cambridge University Press
- Reichstein M, Camps-Valls G, Stevens B et al (2019) Deep learning and process understanding for data-driven earth system science. *Nature* 566:195–204
- Irrgang C, Boers N, Sonnewald M et al (2021) Towards neural earth system modelling by integrating artificial intelligence in earth system science. *Nat Mach Intell* 3(8):667–674
- Steffen W, Richardson K, Rockström J et al (2020) The emergence and evolution of earth system science. *Nat Rev Earth Environ* 1(4768): 54–63
- Castree N (2021). Earth System Science. In *International Encyclopedia of Geography* (eds D. Richardson, N. Castree, M.F. Goodchild, A. Kobayashi, W. Liu and R.A. Marston). <https://doi.org/10.1002/9781118786352.wbieg1087.pub2>
- National Academies of Sciences, Engineering, and Medicine (2021) Next generation earth systems science at the National Science Foundation. The National Academies Press, Washington, DC. <https://doi.org/10.17226/26042>

Earthquake Complexity

William I. Newman

Departments of Earth, Planetary, & Space Sciences, Physics & Astronomy, and Mathematics, University of California, Los Angeles, CA, USA

Definition

Earthquakes belong to a class of phenomena referred to as complex where the hallmark of complexity is the interplay of a variety of nonlinear factors. These factors embrace behaviors such as self-organization, scaling and fractal behavior, chaos in high degree-of-freedom systems, including sensitivity to initial conditions, and the emergence of hierarchical networks wherein these factors can mutually influence each other, even in the presence of extraordinary heterogeneity.

Introduction

Earthquakes continue to pose catastrophic risks to major population centers. Historically, they have had devastating

effects with the most extreme events taking place in Shangxi, China (designated as 1556 M8.0 Shangxi earthquake), which possibly killed over 800,000, to the twentieth century with the Tangshan, China, event (designated as 1976 M7.6 Tangshan earthquake) where estimates of fatalities range from over 240,000 to as many as 650,000. During this century, the 2004 Indian Ocean earthquake and tsunami (designated 2004 M9.1 Sumatra-Andaman earthquake), as well as the 2010 Haitian earthquakes, have each taken approximately 200,000 lives, with many more lost due to the unavailability of medical assistance and proper sanitation and hygiene. The heightened awareness of the earthquake threat has become especially prominent in assessments of potential seismic catastrophes in both northern and southern California (along the San Andreas Fault), the Pacific Northwest of the United States and Canada (Juan de Fuca Plate, commonly called the Cascadia subduction zone), and in Japan (Tōkai region between Nagoya and Tokyo). Taken together, these events and our expanding knowledge of the earthquake mechanism gave rise to the hope that earthquakes of substantial size could be predicted with some specificity regarding their magnitude, location, and time. For reasons that we shall discuss, these goals were quickly dashed owing to their failure to make reliable predictions. It is important to note, quite apart from the attendant scientific and engineering issues, that the prediction of major events that do not occur and, even more significantly, the failure to predict major destructive events have dramatic legal–political–social outcomes. In the aftermath of the failure to develop reliable prediction techniques, there is now emerging a renewed hope in forecasting such events using pattern recognition schemes, possibly in conjunction with machine intelligence. In contrast with prediction, earthquake forecasting has a limited capacity to provide a probability-based assessment of seismic hazards in relatively large spatial and temporal domains, for instance, that southern California might expect a significant earthquake (magnitude 6.7 or greater) during the coming 30 years of 59%, with a 37% chance of a magnitude 7.5 or greater event during that time window. An excellent nontechnical overview of our general understanding of earthquake phenomenology and its inherently unpredictable nature may be found in the semi-popular books by Hough (2009) and Jones (2019). These estimates undergo continual updating and revision, but it is unlikely that we will ever be able to predict events in the way that we routinely predict the weather.

Ironically, we sought to improve the quality of local or synoptic weather forecasting using numerical weather prediction models (Browning 1980) in conjunction with routine updates of prevailing conditions via “data assimilation” techniques. This has given rise to the concept of “nowcasting,” literally the amalgamation of “now” with “forecasting,” the term coined by Browning (1980). Very short or mesoscale prediction over periods of 2 hours is a methodology accepted

by the World Meteorological Society, while the blending of econometric data with statistical models has become popular in economics as a modality for assessing the state of the economy. In a similar fashion, geophysicists such as Rundle et al. (2021) seek to blend existing observations of long-term trends with potentially observable changes in the (local) state of the Earth employing a variety of possible “precursors,” manifestations that likely can be causally connected with the underlying physics of rock fracture. The application of new methodologies arising from machine learning, data mining over complex hierarchical networks, and artificial intelligence could lead to the ability to obtain improved estimates of short-term, on a scale of hours to days, major seismic events. As we will observe later, earthquakes and the fault networks on which they occur are “complex systems” and, like other complex systems (Malik and Ozturk 2020) – including the brain, human society, and infrastructure systems – experience events that are rare in frequency but can have catastrophic consequences. The complex system framework provides both new challenges and opportunities.

Finally, there are some aspects of seismic wave propagation that could be exploited by “warning systems” that utilize the difference in propagation velocities of the first-to-arrive P (for primary or compression) waves, and the much more destructive but significantly later-to-arrive S (for secondary or shear) waves. As an illustration, P waves travel with velocities ranging from 5 to 8 km/s, depending on rock type, in contrast with S waves whose seismic velocities are typically 60% of P waves. In essence, the initial detection of a P wave with potentially much larger S wave to come could allow major infrastructure components, such as electrical power to be switched to emergency generators while natural gas supplies would be terminated to reduce the risk of fire, as well as other mitigation actions.

The Implausibility of Earthquake Prediction

The focus of this chapter will be on understanding why prediction is impossible while forecasting has limited value. The immediate and profound implication here is that construction standards in earthquake-prone areas must of necessity be increased in homes, businesses, schools, hospitals, and other life-critical agencies. They must be designed, equipped, and prepared to remain resilient in their response to earthquakes and quickly recover in their aftermath (Hough 2009; Jones 2019). While there are many other issues pertaining to earthquakes and their effects, we will provide a glimpse into the nature of the uncertainties underlying the earthquake mechanism via the 1906 San Francisco earthquake. We will observe how they transcend simple notions of chaotic behavior and require that seismic events be regarded as illustrations of dynamic complexity. The implications of that paradigm

shift are profound. Our approach is to focus on an audience trained in applied mathematics, including some application to the geosciences, but having only modest exposure to ideas emergent from geology as well as new concepts arising from what is presently called complexity theory. A good introduction to the essential aspects of complexity can be found in the semi-popular books by Waldrop (1993) and Mitchell (2009). Unlike many other arenas in the mathematical geosciences, such as geophysical fluid dynamics, the nature of earthquake faults and their distribution and associated phenomenology are not known a priori. Friction parameters and stress distributions on faults are known to some degree based on seismicity patterns, stress drops, and slip orientation identification. Hence, modeling such phenomena is only possible in a very approximate way through a variety of phenomenological models arising from empirical observations of earthquake events. However, they are not known well enough to be used as reliable initial conditions for simulations in the context of prediction.

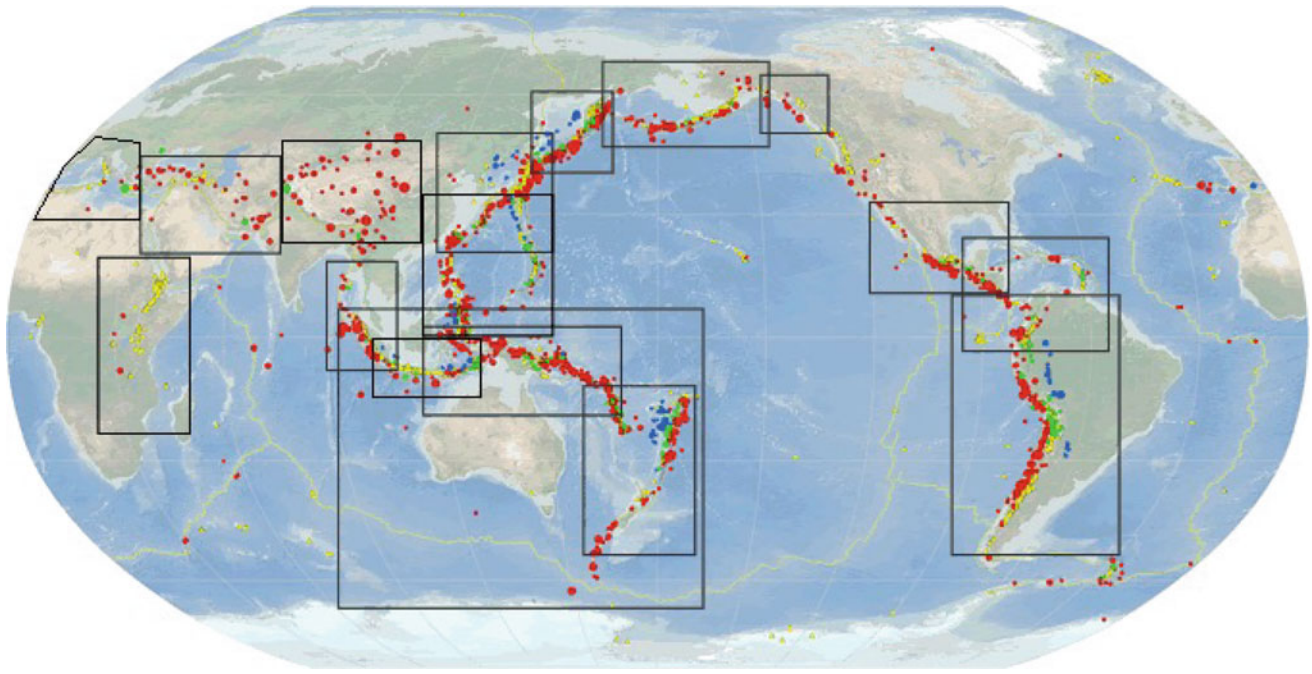
With these caveats, let us now consider some of the phenomenological aspects that characterize regions that are particularly prone to earthquakes as well as introduce some of the nomenclature, before proceeding to explore global aspects of such behaviors. Earthquakes refer to the shaking that is experienced intermittently in such localities, but global factors underscore their selection. The shaking is often described by its magnitude, which, loosely speaking, is a logarithmic measure of the energy released in the process, although there are better quantitative descriptors such as the seismic moment. The magnitude of events was invented by Gutenberg and Richter in 1958, who observed a scaling law relating the magnitude to the frequency of such events, a theme we will return to later, inasmuch as it is manifestly universal. (Ironically, Richter has been quoted as saying that “only fools and charlatans” predict earthquakes – a litany of false predictions and failure to predict are a fixture of contemporary life.) The cause of earthquakes is generally recurrent with the time interval between events and the energy released being variable. Qualitatively speaking, they are related to fracture or rupture of earth materials beginning at a location referred to as the hypocenter at approximately 10–20 km depth; the projection of the hypocenter on the Earth’s surface is referred to as the epicenter. For example, suppose a given region with an area of several hundred square kilometers experiences a magnitude 4 earthquake approximately once a month, then it will experience magnitude 5 being 10 times less frequently, and so on. Similarly, the energy released in a magnitude 5 event is approximately 30 times greater, more accurately $\sqrt{1000}$, than a magnitude 4 event. While the amplitude of earth motion varies with respect to the magnitude M for low-magnitude events, say, $M \leq 4$, the amplitude of body-waves will saturate as events begin to exceed $M \approx 6.5$ but shaking quantitatively

observed as body waves will be sustained for longer intervals of time than smaller events. Meanwhile, the shaking associated with surface waves saturates at $M \approx 8.0$ and above. We qualitatively distinguish among earthquakes as being foreshocks, main shocks, and aftershocks – each of these designations is largely artificial inasmuch as they present the same qualitative seismic waveform and refer solely to their respective magnitudes. The main shock is preceded by a foreshock 30% of the time – and this underscores an underlying limitation to prediction: there is no a priori way to distinguish the category in which a given event occurs. Most of the time, around 70% of events, an earthquake is a “main” shock, but 30% of the time that earthquake is a premonitor of a larger one to come. Meanwhile, most shocks are followed by lower-magnitude aftershocks with a diminishing rate of occurrence (Omori’s law) and also display a drop in magnitude of approximately 1–1.2 (Bath’s law). Having noted the energetic character of destructive earthquakes, it is noteworthy that the January 17, 1994, Northridge, CA, earthquake with a magnitude of 6.7 released more energy than a 5 megaton bomb, while the Sumatra-Andaman earthquake and tsunami of December 26, 2004, with a magnitude of approximately 9.1 released 1000 times more energy than the largest hydrogen bombs detonated during the depths of the cold war. Associated with the initiation of an earthquake with the local failure of rock materials at depth, a variety of related phenomena can occur. For example, gases associated with the geochemical synthesis of those materials can be trapped during their formation in the interstices between grains in rocks but are released when those materials are subject to extreme stress and the gases released might be measurable. Similarly, the changes in the rock environment can influence its electrical resistivity. The failure of rock materials subject to micro-cracking could result in ultrasonic emission of sound and possibly other phenomena. This in turn has been related to observations of various mammals and even fish manifesting “skittish” behavior prior to substantial seismic events. The operative question, however, is whether any of these or possible other physical–chemical process provide a set of precursory phenomena that regularly and reliably serve as prognosticators of impending earthquake activity. The answer, sadly, is no – though certain precursors appear to occur about 30% of the time – the books by Hough (2009) and Jones (2019) offer good resources accessible to lay audiences into the limitations of prediction.

Pervasive Patterns Consistent with Complexity

Let us now turn our attention to the global distribution of earthquake events and how these must be incorporated into developing a mechanistic understanding of the earthquake process and its multifactorial manifestations, and the

fundamental limitations to their prediction. Prior to the 1906 San Francisco earthquake, the prevailing view of the processes associated with the formation of the Earth’s surface and topography arose from ideas relating to the formation of solar system bodies through a process of gravitational accretion of granular and icy materials present in the primitive solar nebula. It was widely accepted then that the collision of such materials and their subsequent gravitational accretion produced the planets that slowly cooled, leaving a wrinkled surface, metaphorically like a baked apple taken out of an oven. Meanwhile, the accumulation of global insights into the distribution of the Earth’s continents and a variety of other inherently geological features led by the pioneering work of Wegener (1923) and others led to the establishment of the paradigms of continental drift and of plate tectonics. The premier feature of plate tectonics is evident from Fig. 1, which describes the global seismicity of the Earth, 1900–2007 (USGS Scientific Investigations Map 3064, public domain), providing a comprehensive overview of great earthquakes (M8.0 and larger). Recalling our reference to the notion of a cooling primordial Earth, we appreciate that the radioactive decay of heavy isotopes in its deep interior preserves a molten liquid state below its crust and mantle. This material in turn is undergoing convection at a rate of several centimeters per year, propelling the cooler-solidified materials toward the surface where they collect and form “tectonic plates.” The critical feature here is that the locus of these epicenters identifies the major convergent tectonic plate boundaries, where the oceanic and continental crust can collide and are forcibly driven together. Depending on location, those contact points can be associated at this time with the plunging of some plates below others, a process known as “subduction” (e.g., the Indian and Eurasian plates and the formation of the Himalayas via dip-slip events) or with shearing resisted by friction (e.g., the San Andreas fault in the Americas with strike-slip events). The East African Rift System, as well as the North and East Anatolian Faults, together with the western edge of the Pacific Plate, including the 2011 Tohoku earthquake and tsunami, has had a profound influence in human historical terms. While most of the relative motion observed on the plates is in the horizontal direction, crudely 10% is in the vertical, contributing to the rise, for example, of the Sierra Nevada mountain range by a few tens of centimeters every century. Approximately 70% of seismic events – as well as volcanic events that are also associated with tectonic plate boundaries and very hot sources of molten rock or magma residing below some descending plates – are associated with such convergent boundaries. Seismic events that take place under the Pacific, especially, can produce tsunamis (the term “tidal wave” is a misnomer), particularly if the relative motion of the two plates has a substantial vertical component. The remaining 30% of seismic events are associated with so-called “hot spots,” including, for



Earthquake Complexity, Fig. 1 Global seismicity 1900–2007 $M \geq 8.0$ with filled colored circles designating earthquake events with hypocenters ranging from 0–69 km (red), 70–299 km (green), and 300–700 km (blue).

The diameter of each circle designates the moment magnitude (M_w) with larger circles identifying larger events.

example, the Hawaiian Islands, which were formed by the injection of magmatic materials through the overlying oceanic crust to form a new land mass, and an even smaller number of “intra-plate” events, likely associated once again with deep-rooted convective cells, such as that which produced the four New Madrid earthquakes of 1811–1812 with $M \approx 7.2$ – 8.2 and resulted in the displacement of the Mississippi River by about 2 kilometers. Each of these different types of earthquake events displays distinct characteristics and is even less amenable to prediction.

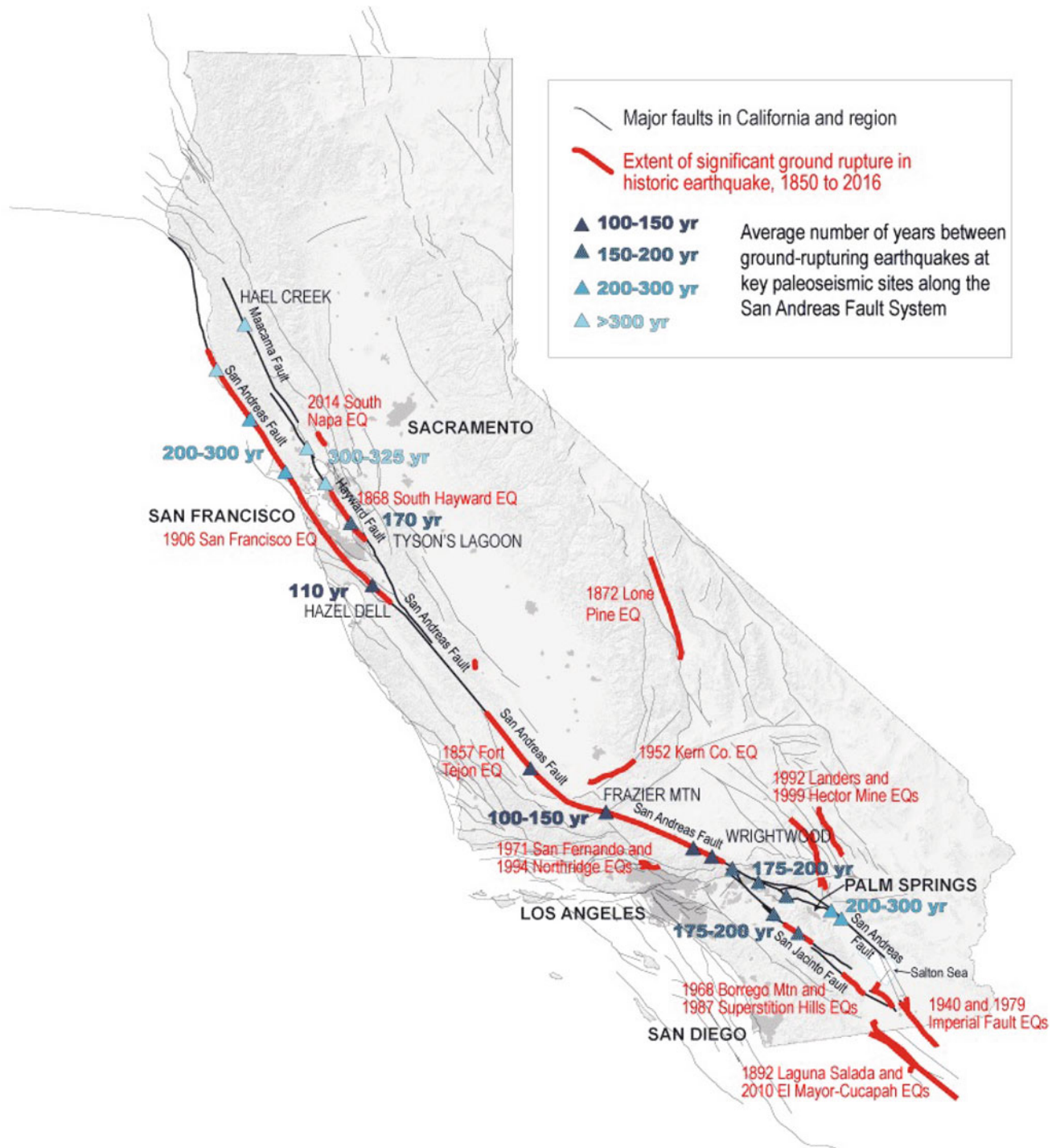
The preceding discussion highlighted the remarkably complicated character of global tectonics driven by fluid motion in the Earth’s mantle. While there are nearly a dozen major plates and many more smaller plates, their dynamic interactions are substantially more complicated. While the mass media focuses on the nature of the San Andreas Fault and the concerns that it raises among residents of Los Angeles and San Francisco, the situation is geologically highly intricate, as shown in the map displaying a complex hierarchical network of fault ensembles and its presentation of California seismicity (Fig. 2). Indeed, we observe that California’s earthquake faults constitute a highly irregular yet hierarchical interlocking network, much of it concealed below the Earth’s surface. There is a significant literature addressing the geometric complexity of fault networks, demonstrating that they satisfy through their scaling statistics the nature of a fractal. See the textbook by Turcotte (Turcotte 1997) for an

introduction to issues attendant to fractals and chaos in geological settings. We will return to the theme of complexity momentarily.

Finally, we wish to assemble in a more quantitative way the statistical behavior observed in the energy released, through the proxy of earthquake magnitudes. Then, we will explore the statistical distribution describing the likelihood that a given event can be associated with specified energy via its logarithm (the magnitude), the so-called Gutenberg–Richter law, and other earthquake-related scaling laws such as Omori’s law and Bath’s law mentioned earlier. This figure reveals that the number N of events larger than a magnitude M satisfies the empirical Gutenberg–Richter law in the form

$$\log_{10} N = -bM + \log_{10} \dot{a} \quad (1)$$

where $\dot{a}$ and b are constants. While the rate $\dot{a}$ depends upon the geographic region and area considered among other factors, the value of b appears to be slightly above 1 without exception. We show this in the following plot, where we have normalized the number count and fit the data using the prevailing standard using the numbers of events above a given magnitude per year. We observe for this global dataset that a “b-value” of approximately 1.025 emerges. Events above magnitude 7.5 fall in frequency below the expression shown, likely because the largest events are associated with ruptures that extend to significant depths below the surface



Earthquake Complexity, Fig. 2 USGS map of the principal faults in California demonstrating the interplay between different fault zones over geological time. This kind of topographical character is an indicator of

fractal geometry. Certain faults displayed could be relics of faults that existed in prior geological periods as the western or leading edge of the North American Plate migrated away from the mid-Atlantic ridge

where temperatures are much higher and the rocks are much more viscoelastic than at the surface. We noted the statistical uniformity of large seismic events. Paradoxically, this unvarying behavior persists despite a remarkable degree of heterogeneity in the rock compositional types, mechanical

properties, and geometries present along the faults that make up the “ring of fire,” the pattern evident in the epicenters that reside on the periphery of the Pacific Ocean. The added irony here is that universal scaling laws are observed statistically, just as we see in fully developed isotropic turbulence in fluid

dynamics. Nevertheless, nonuniformity and heterogeneity prevail, thereby compounding the challenge of developing any reliable predictive capacity.

The 1906 San Francisco Earthquake: Unequivocally Unpredictable

Our understanding of the earthquake mechanism underwent a fundamental advance in the wake of the October 18, 1906, San Francisco earthquake, the first major earthquake to strike a large population center in the twentieth century. An important geographic aspect of the “ring of fire” is that it resides in coastal areas that are very attractive for human habitation with some of the world’s greatest cities exposed to risk from major earthquakes and their associated tsunamis. With an estimated magnitude of $\approx 7.9 \pm 0.2$ and a maximum Mercalli intensity of XI (a semi-quantitative metric designated by Roman numerals and designating the degree of shaking affecting structures), more than 3000 people died and more than 80% of the city was destroyed. This event occurred long before the ring of fire and plate tectonics came to be widely recognized and accepted in the 1970s. Nevertheless, the State Earthquake Investigation Commission issued a remarkably prescient and comprehensive 451-page report including 146 photographic plates. This monumental study documented the highly varied geological setting and manifestations of the earthquake that were directly observable over the >450 km extent of the rupture. Evidently, the region affected originated at ≈ 8 km depth below part of the feature called at the time the “San Andreas rift.” Geologists also referred to it as a fault, having no comprehension of how it came to be nor especially how it was part of a global-scale phenomenon. The report describes the geological signatures observed along ruptured segment of the fault. In particular, geologists noted that the geological signatures were highly variable. Indeed, they were witnessing evidence that the two sides of a relatively linear segment of the fault had undergone transport in opposing directions extending over hundreds of kilometers. They recognized northward movement on the west side (now identified with the Pacific Plate) relative to southward movement on the east side (now identified with the North American Plate). The heterogeneous and nonuniform character of the different rock types – with different compositions, mineralogical content, and history – underscored how complicated and varied were the geochemical and geophysical signatures present.

An amazing outcome of this report was the development by one of its members, Professor Harry Fielding Reid of Johns Hopkins University, of the “elastic rebound theory.” He concluded that the crust of the Earth can gradually store elastic stress – due to what we now know is due to the underlying mantle motion driving the tectonic plates above – and that the earthquake must have involved an “elastic rebound” of

previously stored elastic stress. His conjecture was supported by numerous photographs showing fences that were literally sheared by the movement below of the segments of earth material to which they were attached. The resulting offset was approximately 2.5 m, although the relative displacement observed along the rupture was somewhat variable. Figure 4 illustrates this using an effective time sequence of one of those fences as seen from above and the evolution of the seismic event from (1) before the event, (2) during the event, and finally (3) after movement stopped. One memorable image shows two displaced segments of the fence with a “patch” constructed connecting the two decoupled segments.

A hint as to the complicated nature of failure is offered by the analogy that one can make with a long thin piece of wood being forcibly bent at each end in opposing directions to form an arc with steadily increasing curvature with the progress of time. Ultimately, we expect that it must break – but when it will and where along its length remains unpredictable. This metaphor illustrates a possible link to “critical point phenomena” and why earthquake prediction remains highly elusive.

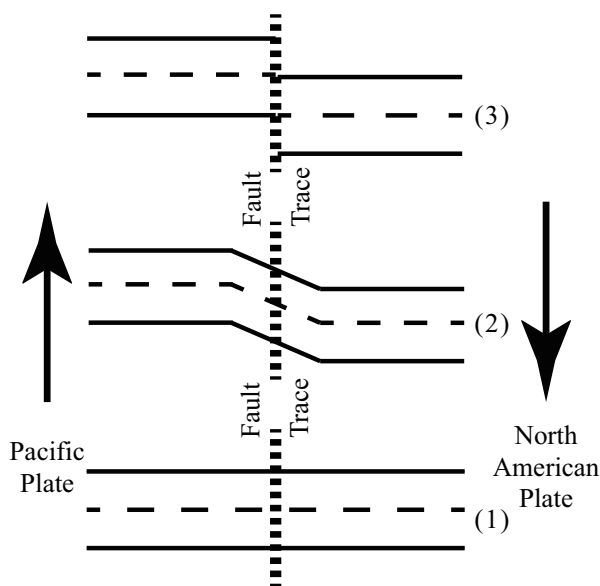
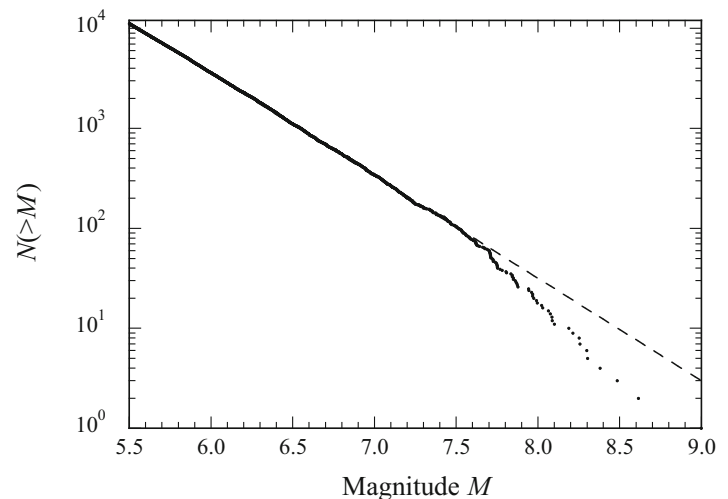
An additional aspect of the elastic rebound conjecture is that we can obtain evidence identified via “paleoseismology” that establishes an existing pattern. Previous earthquakes indicate that recurrent events can occur showing a crude measure of periodicity, although the time window for a major event remains highly uncertain. Nevertheless, probabilistic approaches have been utilized to estimate approximately the maximum magnitude, return period, and geological fault slip for future events and produce a “probability” of an earthquake in a certain area during a specified window of time, as we mentioned previously with respect to southern California.

This narrative, given the highly fragmented nature of California faults displayed in Fig. 3, is a gross oversimplification of what we appreciate better today. The notion that a simple frictional model can be employed to describe the interaction between the two “faces” of a fault is oversimplified. Those faces are themselves the product of seemingly countless seismic events and are highly irregular and crenulated, thereby presenting an indescribably complicated picture of the fault geometry and how it varies along its length. Indeed, we expect fragments on many length scales to form between the rock faces, making the interaction of the rock faces even more complex. Thus, the San Andreas near the site of the 1857 Fort Tejon earthquake near present-day Palmdale has the two fault surfaces separated by 1–2 km with the intervening volume filled by the rock rubble and detritus of materials fractured earlier. This rubble is referred to as “fault gouge” and that volume as the “damage zone.” Interestingly, the rubble assembly will typically drop at various locations below the level of the two opposing rock faces, establishing locations where rainwater can accumulate and, thereby, form “sag ponds.”

Rock mechanics as it has come to be known, quite unlike fluid mechanics, is a very approximate phenomenological

Earthquake Complexity,

Fig. 3 Global seismicity from 1977 to 2007 and the Gutenberg–Richter empirical law. (Adapted from Newman (2012) with data from the Global Centroid-Moment-Tensor (CMT) Project at Columbia University, which is available online)



Earthquake Complexity, Fig. 4 Illustration of elastic rebound theory for the 1906 San Francisco earthquake developed by H.F. Reid (1911). (Adapted from <https://earthquake.usgs.gov/earthquakes/events/1906calif/18april/reid.php>)

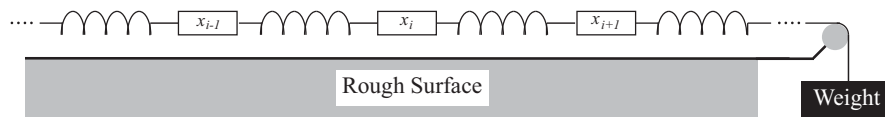
theory. The added reality of heterogeneous rock types, fluidization, variable thermal histories, and so on makes it all the more clear that earthquake prediction is an unrealistic and unattainable goal. Modern continuum mechanics, which provides a mathematical first-principles but approximate basis for consideration, provides the equations underlying elastic wave as well as some insight into “plastic” and other dissipative “non-Newtonian” systems. However, those features are largely empirical and parameter based at best – see, for instance, Newman (2012) – thereby leaving open the question of where can we go from here.

We have already introduced the concepts attendant to pattern formation, namely, the isolation and identification for long-time large-area risk assessment. Data assimilation,

allowing for the regular updating of inputs to models for events, can provide a route to nowcasting amidst emergent developments in machine intelligence utilized over hours or, possibly, days. The 1906 San Francisco earthquake was preceded by a foreshock ≈ 25 seconds earlier – but only $\approx 30\%$ of seismic events are foreshocks – and it provided no hint as to the coming main shock, which lasted 42 seconds. In contrast, the 1964 Great Alaska Good Friday earthquake with a magnitude of 9.2 produced shaking over several minutes and a tsunami that ultimately destroyed the Chilean navy. Fortuitously, the risk presented by tsunamis can be established following a great earthquake taking place primarily in the Pacific, but the earthquake that produced it cannot be predicted. Hence, we remain limited by our inability to predict complex events such as earthquakes, among others. The take-home lesson here is that we can and, of necessity, must prepare for seismic events in earthquake-prone areas, particularly around the ring of fire whose human population continues to grow.

Burridge–Knopoff Model (1967) as a Paradigm for Complexity

Simple models may provide additional insight into the complex nature of earthquakes, and, possibly, some of these insights could yield clues into the kinds of nonlinear phenomena present. A useful summary of some of these models and the characteristic nonlinear phenomena that they can help elucidate may be found in the review by Shcherbakov et al. (2015). We will end our treatment of earthquake complexity by presenting one of the earliest yet simplest models for the elastic rebound theory and its possible link to the Gutenberg–Richter (power-law) frequency–magnitude relation. Figure 5 illustrates the Burridge and Knopoff (1967) or slider-block model. While often described as a model for deterministic chaos, demonstrating sensitivity to initial conditions, it can



Earthquake Complexity, Fig. 5 Depiction of the Burridge and Knopoff (1967) model, commonly called the slider-block model, for the elastic rebound theory. (Adapted from Newman (2012))

incorporate a very large number of constituent parts and, therefore, becomes a high-degree-of-freedom system and, in the continuum limit, an infinite-degree-of-freedom or complex system. Let us provide a simple description of how this toy model operates – a more complete description may be found in Newman (2012) and, especially, Burridge and Knopoff (1967).

Specifically, this model is designed to describe the “stick–slip” behavior associated with earthquake faults. We described earlier the highly complicated geomorphological nature of earthquake faults and the intervening damage zone and fault gouge; clearly, the Burridge–Knopoff model is a skeletal or toy depiction of that problematic reality. The model incorporates a solid, massive block with a rough surface, depicting one fault surface and an amalgam of identical weights or slider blocks connected by springs that satisfy Hooke’s linear force law for the other fault that is being driven by a weight shown on the right-hand side. The weight, then, may be regarded as the plate tectonic forces arising from mantle motion below the plates forcing the plates to move relative to each other subject to irregular resistive “frictional” forces. Each slider block is assumed to have its horizontal motion resisted by a frictional force proportional to its weight. In this scheme, the connecting springs apply dynamic forces to each of the blocks; however, a given block will resist moving unless the combination of the two springs acting upon it provides a force greater than that of friction. A given block will not move unless the spring forces apply exceed the force of static friction. Should that happen, the block will move in accord with Newton’s force law, thereby causing the connecting springs to respond and “slip,” resulting in a reduction of the net force that springs provide below the frictional limit, whereupon the affected block will once again “stick.” Suppose, now, that the system of slider blocks is at rest and allow the weight on the right-hand side (representing tectonic stresses) to increase. Ultimately one block will slip and, in the process of its altering the extension of the adjoining springs, could cause an adjacent block to slip; this process could repeat itself among other blocks... Every time a block slips, we identify an earthquake since this presents a metaphor for the Earth shaking at that spot. Most of the time, only isolated individual blocks will slip, but there will be times when 2 or 3 or ... will slip. It is appropriate to regard this as a kind of “chain reaction” that can be interpreted as a form of pattern formation and self-organization, offering further the

possibility of some form of critical point behavior. The magnitude of an earthquake corresponds to the number (or the logarithm of the number) of blocks that suddenly move. Meanwhile, the frequency of events is simply a number count of events of a given size, that is, the number count of blocks that respond. Remarkably, power-law statistics not unlike the Gutenberg–Richter law emerge, often associated with fractal behavior. An infinitesimal change in the original position of any of the slider blocks will result after a short period of time in an unrecognizably different configuration, yet one that preserves the overall power-law frequency–magnitude statistic. This model can readily be modified to allow for heterogeneity and, not surprisingly, the linear chain can be extended to form hierarchical networks. In other words, this deceptively simple model can show indications of many, if not all, of the features observed in real earthquakes. It is highly instructive to construct such a model numerically or build a physical “toy” using blocks connected by springs and so on as shown in Fig. 5. Despite its overall simplicity, we can readily identify its overarching complexity and emerge in the process with an appreciation that earthquake prediction remains an impossible dream. However, we can also come away from this experiment with a belief that, with adequate updating of our knowledge of the state of the Earth and the application of data assimilation from all possible precursors, some limited success may be achieved via nowcasting.

Summary and Conclusions

Natural hazards and earthquakes, especially, result in major losses of life and property. Ideally, we wish to develop methods for predicting earthquakes and, thereby, reduce human suffering and loss. It is important to note that inaccurate prediction schemes are profoundly problematic since they will fail to predict major events and will claim to “predict” events that do not take place – both of these situations have substantial and potentially tragic outcomes.

Some patterns emerge when considering seismic activity, particularly their proclivity to occur in specific regions of the globe on earthquake faults. This applies to seismicity with a limited degree of temporal regularity but with substantial variation in recurrence time. These features are evident from the historical record and what has come to be called the “paleoseismic” record, which incorporates natural indicators

of drastic change. For instance, the severed roots of trees that straddled a fault and their annular tree rings provide indications of when the earthquake took place and the magnitude of the lateral displacement displayed by their broken roots on the two sides of the fault. Of particular importance, the ring of fire, which figuratively encircles the Pacific Ocean, is especially geologically active owing to plate tectonic dynamics while at the same time is home to rapidly growing population centers – approximately 70% of earthquake activity takes place there.

The geological features and processes underlying earthquake faults are also remarkably complicated. Earthquake faults exhibit a remarkable degree of compositional heterogeneity as well as highly irregular geometries and variable stress distributions along their length. We employed the detailed investigation of the 1906 San Francisco earthquake to illustrate this point. The tendency among many mathematical geoscientists to treat earthquakes solely as “events” represents earthquakes as points in a space–time continuum according to their epicenter locations and rupture initiation times. While rupture may be initiated at a point, it will grow and could propagate over hundreds of kilometers extending over seconds to possibly minutes of time. Collectively, these factors make it very difficult, if not impossible, to develop mathematical models to describe the evolution of earthquake faults and their potential for failure. However, mathematical models such as the pioneering Burridge and Knopoff (1967) model provide our intuition with a simple yet deeper appreciation for what nature is displaying, and also serves as an example of a mathematical “complex” system.

Seismic activity is an illustration of mathematical complexity that is observed in other natural and social phenomena: these are associated with (a) dynamically active systems with very large numbers of degrees of freedom and having substantial sensitivity to initial conditions, (b) statistical or probabilistic scaling features sometimes referred to a “laws,” (c) fractal geometric structure associated with hierarchical networks and topologies, and (d) collectively may appear to undergo self-organization. Simply put, these aspects of seismicity have defeated efforts to develop prediction schemes with any modicum of reliability. Earthquakes are fundamentally complex phenomena.

There have emerged advances recently in machine learning and artificial intelligence methodologies that offer some hope. In other arenas displaying complexity, such as numerical weather prediction and the advent of “nowcasting,” these methodological advances have demonstrated a limited capacity to make forecasts. By coupling the progress made using these new computational technologies with the ability to assemble and analyze massive data sets with time-adaptive algorithms, there is emerging a possibility that we could obtain some limited success in anticipating major seismic events.

Unless such efforts prove reliable, we have no alternative but to pursue a three-pronged approach to earthquake hazard

mitigation: (a) develop and enforce earthquake-proof construction standards together, (b) engage everyone in maintaining safe living and work environments, and (c) prepare for the aftermath of such events when social and commercial disruption are to be expected. Further, we should utilize the velocity differences in the initial but less destructive P-waves and the inevitable and much more destructive S-waves. When employed in major earthquake-prone urban centers, time intervals of tens of seconds or more could provide the margin of safety needed to switch on alternative, independent power supplies, turn off potentially dangerous natural gas and other fire hazards, and provide individuals an opportunity to take cover or to find appropriate shelter. If all of these measures are rigorously pursued, then we will be able to “ride out” such events with relatively few casualties and limited property loss, which is a fundamental benefit for the study of all natural hazards.

Bibliography

- Browning KA (1980) Review lecture: local weather forecasting. *Proc Royal Soc Lond A Math Phys Sci* 371(1745):179–211
- Burridge R, Knopoff L (1967) Model and theoretical seismicity. *Bull Seismol Soc Am* 57(3):341–371
- Hough SE (2009) *Predicting the unpredictable*. Princeton University Press
- Jones L (2019) The big ones: how natural disasters have shaped us (and what we can do about them). Anchor
- Malik N, Ozturk U (2020) Rare events in complex systems: understanding and prediction. *Chaos* 30(9):090,401
- Mitchell M (2009) *Complexity: a guided tour*. Oxford University Press
- Newman WI (2012) *Continuum mechanics in the earth sciences*. Cambridge University Press
- Rundle J, Stein S, Donnellan A, Turcotte DL, Klein W, Saylor C (2021) The complex dynamics of earthquake fault systems: new approaches to forecasting and nowcasting of earthquakes. *Reports on progress in physics*
- Shcherbakov R, Turcotte DL, Rundle JB (2015) Complexity and earthquakes. In: *Treatise on geophysics*, 2nd edn, pp 627–653
- Turcotte DL (1997) *Fractals and chaos in geology and geophysics*. Cambridge University Press
- Waldrop MM (1993) *Complexity: the emerging science at the edge of order and chaos*. Simon and Schuster

Eigenvalues and Eigenvectors

Jaya Sreevalsan-Nair

Graphics-Visualization-Computing Lab, International Institute of Information Technology Bangalore, Electronic City, Bangalore, Karnataka, India

Definition

In linear algebra, a linear transformation refers to a mapping between two vector spaces that preserve operations of vector

addition and scalar multiplication, such that $\mathcal{L} : \mathcal{V} \rightarrow \mathcal{W}$, where $(\mathcal{V}, \mathbb{F})$ and $(\mathcal{W}, \mathbb{F})$ are vector spaces defined over a field $\mathbb{F}$. The commonly known linear transformations that can be expressed geometrically are rotation at a fixed point about an axis and scaling along the basis vectors. To understand rotation and scaling, they can be understood with respect to a three-dimensional (3D) object and the basis vectors, i.e., x , y , and z axes in the Cartesian coordinate space. In computer graphics, an object is often represented using a surface mesh, discretized into triangles. The vertices of the triangles in the mesh can be rotated or scaled to rotate or scale the object, respectively. These vertices can be written as 3D column vectors. So, what is the relationship between linear transformations and *eigenvalues and eigenvectors*? To explain the relationship, the representation theorem for the linear transformations must be discussed. The theorem states that a linear transformation $\mathcal{L} : \mathcal{V} \rightarrow \mathcal{W}$ can be represented as a matrix A , where the matrix is computed from the bases of $\mathcal{V}$ and $\mathcal{W}$. The matrix $A = \{a_{ij}\} \in \mathbb{R}^{m \times n}$ gives the transformed basis vector of $\mathcal{V}$ as $\mathcal{L}v_i = \sum_{k=1}^m a_{ki} w_k$. Thus, $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$. When the dimensionality is the same for the domain and co-domain, A is a square matrix, i.e., $A \in \mathbb{R}^n \times \mathbb{R}^n$. Focusing on square matrices alone here, upon multiplying A with an arbitrary n -dimensional column vector $\mathbf{x}$, the matrix vector product $A\mathbf{x}$ is a transformation of x involving both rotation and scaling. However, there exist rotation-invariant vectors that can scale, which means the resultant vector $A\mathbf{x}$ is either a compressed or an elongated version of $\mathbf{x}$ itself, but with the same orientation as $\mathbf{x}$. Thus, such a vector $\mathbf{x}$ is a property of the matrix A . Thus, rewriting as $A\mathbf{x} = \lambda\mathbf{x}$, the vector $\mathbf{x}$ is referred to as the “right eigenvector of A ,” and λ is a scalar called as the “eigenvalue that corresponds to x ,” and $\lambda \in \mathbb{R}$. Similarly, for $y \in \mathbb{R}^n$, if $y^T A = \mu y^T$, for a scalar value μ , then y is referred to as the “left eigenvector of A ,” and μ is the eigenvalue that corresponds to y . Thus, the left eigenvector of A becomes the right eigenvector of its transpose, A^T , and vice versa.

Overview

The problem of finding the eigenvectors and corresponding eigenvalues of a square matrix is called eigenvalue decomposition of the matrix. It is also a way to factorize the matrix. The eigenvalue decomposition solves for an equation $A\mathbf{x} = \lambda\mathbf{x}$ (Heath 2018). Conventionally, the computational methods are designed to find the right eigenvectors, for which the set of eigenvalues is referred to as $\lambda(A) = \{\lambda\}$. The determinant of $(A - \lambda I_n)$ is referred to as the *characteristic polynomial* of A , p_A , where I_n is the identity matrix of size n . The solution to the equation is given by $(p_A = 0)$. Algebraic multiplicity of an eigenvalue λ of A is the number of times λ occurs as the root

of p_A . A similar property of eigenvalues is its geometric multiplicity, where it is the dimensionality of the eigenspace of A . Eigenspace ϵ_A is the nullspace $(A - \lambda I_n)$. The geometric multiplicity of an eigenvalue is also defined as the number of linearly independent eigenvectors that are associated with the eigenvalue. For every eigenvalue λ of A , the geometric multiplicity need not be the same as the algebraic multiplicity, but cannot exceed the algebraic multiplicity. A is diagonalizable when every eigenvalue has its geometric multiplicity equal to the algebraic one.

The equation is conventionally solved by transforming it to where the characteristic polynomial is zero and solving for the same. While this approach works for small values of n , for general cases, methods such as power iteration, inverse iteration, Rayleigh quotient iteration, simultaneous iteration, orthogonal iteration, and QR iteration are used. QR iteration is popularly used in this regard, where it is implemented as a two-stage process, involving the generation of tridiagonal and diagonal matrices. The eigenvalue problems in practice tend to take the form $A\mathbf{x} = \lambda B\mathbf{x}$, which is also referred to as the generalized eigenvalue problem. If A and B are both non-singular, then the generalized eigenvalue problem can be reduced to the standard format, as $(B^{-1}A)\mathbf{x} = \lambda\mathbf{x}$ or $(A^{-1}B)\mathbf{x} = (1/\lambda)\mathbf{x}$.

Compared to these iterative methods of finding all eigenvectors and eigenvalues, which are of $O(n^3)$ complexity, an alternative, efficient strategy is to find eigenvalues using an $O(n^2)$ algorithm and then using inverse iteration for computing the corresponding eigenvectors. This is the strategy used in the relatively robust representation (RRR) method. This method can also be adapted to determining selected eigenvalues instead of exhaustively finding all eigenvalues to make the algorithms more efficient and fulfilling the exact requirements of the eigenvalues. Additionally, in order to improve the numerical accuracy of the computed eigenvalues and eigenvectors, numerically stable libraries in EISPACK, LAPACK, etc., with extended precision arithmetic are used.

Eigenvalue decomposition is related to singular value decomposition (SVD) as the singular values of a rectangular matrix, say, C of size $m \times n$, are the non-negative square roots of the eigenvalues of $A^T A$. SVD gives $C = U\Sigma V$, where U and V are orthogonal matrices of sizes m and n , respectively. It is also known that the columns of U and V are orthonormal eigenvalues of AA^T and $A^T A$, respectively.

It is important to know that eigenvalue decomposition has a close connection with the *spectral theory*. The set $\lambda(A)$ is also referred to as the spectrum of A . The maximum absolute value of the eigenvalues is called spectral radius of A , $\rho(A) = \max \{|\lambda| : \lambda \in \lambda(A)\}$. If A is symmetric and diagonalizable, then the decomposition gives $A = Q\Lambda Q^T$, where Q is an orthogonal matrix with eigenvectors as rows, and Λ is a diagonal matrix with eigenvalues, in the same order as the eigenvectors are positioned in Q . This decomposition, thus,

provides the quadratic form of $\mathbf{A}$. If the eigenvalues are unique, then $\mathbf{Q}$, $\mathbf{\Lambda}$, and $\mathbf{A}$ are matrices of the same size. The introductory content so far discussed in this chapter has been summarized from the textbooks by Laub (2004) and Heath (2018).

Hawkins (1975) has documented the history of eigenvalue problems in the eighteenth century, including the key contributions by Cauchy, d'Alembert, Lagrange, and Laplace. The connections between spectral theory and eigenvalues stem from astronomical applications (Hawkins 1975). The history includes the principal axes theorem that uses eigenvalue decomposition for reduction to the normal form of a symmetric bilinear form and the diagonalization of a real symmetric matrix (Steen 1973). Hilbert had then extended the principal axes theorem by generalizing the concept of eigenvalue through the definition of the spectrum of symmetric matrices. The subset of $\lambda(\mathbf{A})$ for which the equation $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ has trivial solutions is referred to as the point spectrum of $\mathbf{A}$, and its complement is called the continuous spectrum. During 1904–1907, Hilbert–Schmidt spectral theory was developed, which explores the use of linear transformations in infinite-dimensional space by using the quadratic forms obtained during eigenvalue decomposition (Steen 1973). Since then, there have been several successors of the Hilbert–Schmidt spectral theory.

Applications with Focus on Covariance Matrix

Principal component analysis (PCA) is a dimensionality reduction algorithm that traces back its history to Karl Pearson and Hotelling. This involves identifying a new orthogonal space to describe a set of n -dimensional points, in such a way that the axes are ordered by decreasing levels of variance along the axis. These new dimensions are called *principal components*. The use of PCA in spatial data involves two approaches (Demšar et al. 2013), where in the first approach PCA is applied to nonspatial attribute data, and in the second approach, it is applied directly to the geographic data. PCA leads to factorization of a rectangular matrix of m measurements and n variables, $\mathbf{X}$, given by $\mathbf{X} = \mathbf{TP}^T$, where $\mathbf{T}$ is the projection of $\mathbf{X}$ onto an r -dimensional space, $\mathbf{P}$, $\mathbf{P}$ is an orthonormal projection matrix, and r is the matrix rank of $\mathbf{X}$. The score matrix $\mathbf{T} \in \mathbb{R}^n \times r$, and the loading matrix $\mathbf{P} \in \mathbb{R}^m \times r$. Dimensionality reduction occurs when only considering dimensions with significant eigenvalue, thus, truncating the eigenvalue set and modifying the original matrix $\mathbf{X}$ to $\mathbf{X}' \in \mathbb{R}^m \times r$. The principal components correspond to eigenvectors of the covariance or correlation matrix $\mathbf{\Sigma}$, given as $\mathbf{\Sigma} = \frac{\mathbf{X}'^T \mathbf{X}}{m-1} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T$.

In the case of spatial data, PCA gives the shape of local neighborhood, which is used in point cloud analysis for light

detection and ranging (LiDAR) (Jutzi and Gross 2009). The shape of the local neighborhood is quantified using a local geometric descriptor, computed as the covariance matrix $\mathbf{\Sigma}$ of position coordinates of the point with respect to its selected local neighbors. $\mathbf{\Sigma}$ of each point provides its geometric features, which are further used in supervised learning classifier for semantic classification.

In geospatial data, PCA is widely used on image or raster data. PCA can be applied to satellite images with N pixels and b bands to compute the covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{b \times b}$. The significant principal components are computed using eigenvalue decomposition of $\mathbf{\Sigma}$. For change detection in land-use cover owing to forest fires (Estornell et al. 2013), the eigenvalues of the second principal component helped in identifying the perimeter of the region affected by the forest fires. Similarly, for synthetic aperture radar (SAR) images, the covariance matrix is used for PCA, where the first and second eigenvalue maps of the pixels give the dominant and secondary backscattering mechanism selection (Fornaro et al. 2014). In polarimetric SAR (Pol-SAR) images, the eigenvalue decomposition of the polarimetric covariance matrices (PCMs) gives the segmentation of areas, classification of typology, and identification of scattering mechanisms (Addabbo et al. 2019).

Summary

Overall, eigenvalue decomposition of linear transformations or measurement matrices has been a powerful tool for information extraction from spatial as well as nonspatial data in geoscientific applications.

Cross-References

- LiDAR
- Matrix Algebra
- Polarimetric SAR Data Adaptive Filtering
- Principal Component Analysis

Bibliography

- Addabbo P, Biondi F, Clemente C, Orlando D, Pallotta L (2019) Classification of covariance matrix eigenvalues in polarimetric SAR for environmental monitoring applications. *IEEE Aerosp Electron Syst Mag* 34(6):28–43. <https://doi.org/10.1109/MAES.2019.2905924>
- Demšar U, Harris P, Brunsdon C, Fotheringham AS, McLoone S (2013) Principal component analysis on spatial data: an overview. *Ann Assoc Am Geogr* 103(1):106–128
- Estornell J, Martí-Gavilá JM, Sebastiá MT, Mengual J (2013) Principal component analysis applied to remote sensing. *Model Sci Educ Learn* 6:83–89

- Fornaro G, Verde S, Reale D, Pauciuolo A (2014) CAESAR: an approach based on covariance matrix decomposition to improve multi-baseline-multitemporal interferometric SAR processing. *IEEE Trans Geosci Remote Sens* 53(4):2050–2065
- Hawkins T (1975) Cauchy and the spectral theory of matrices. *Hist Math* 2(1):1–29
- Heath MT (2018) *Scientific computing: an introductory survey*, Revised Second Edition SIAM. <https://doi.org/10.1137/1.9781611975581>
- Jutzi B, Gross H (2009) Nearest neighbour classification on laser point clouds to gain object structures from buildings. *Int Archiv Photogramm Remote Sens Spat Inf Sci* 38(Part 1):4–7
- Laub, Alan J. (2004). *Matrix analysis for scientists and engineers* (Vol. 91). SIAM. ISBN: 978-0-898715-76-7. <http://bookstore.siam.org/ot91/>
- Steen LA (1973) Highlights in the history of spectral theory. *Am Math Mon* 80(4):359–381

Electrofacies

John H. Doveton
Kansas Geological Survey, University of Kansas, Lawrence,
KS, USA

Definition

The term electrofacies was first introduced and defined to be “the set of log responses which characterizes a bed and permits it to be distinguished from the other” (Serra and Abbott 1980).

Facies and Electrofacies

A widely quoted definition of facies is “A facies should ideally be a distinctive rock that forms under certain conditions of sedimentation reflecting a particular process or environment” (Reading 1978). In contrast, an electrofacies is a collective association of log responses that are linked with geological attributes either implicitly through interpretation or explicitly linked to a lithology database. Electrofacies are clearly determined by geology, because log responses are measurements of the physical properties of rocks. The intent of electrofacies identification is generally to match them with lithofacies identified in the core or an outcrop.

An important philosophical distinction between electrofacies and “classical” geological facies is that electrofacies are primarily observational in origin, and facies are traditionally rooted in genesis. Electrofacies can be distinctive empirical log-response associations, but whether they reflect

depositional environment, diagenetic over-prints, or other processes requires careful consideration on a case-by-case basis.

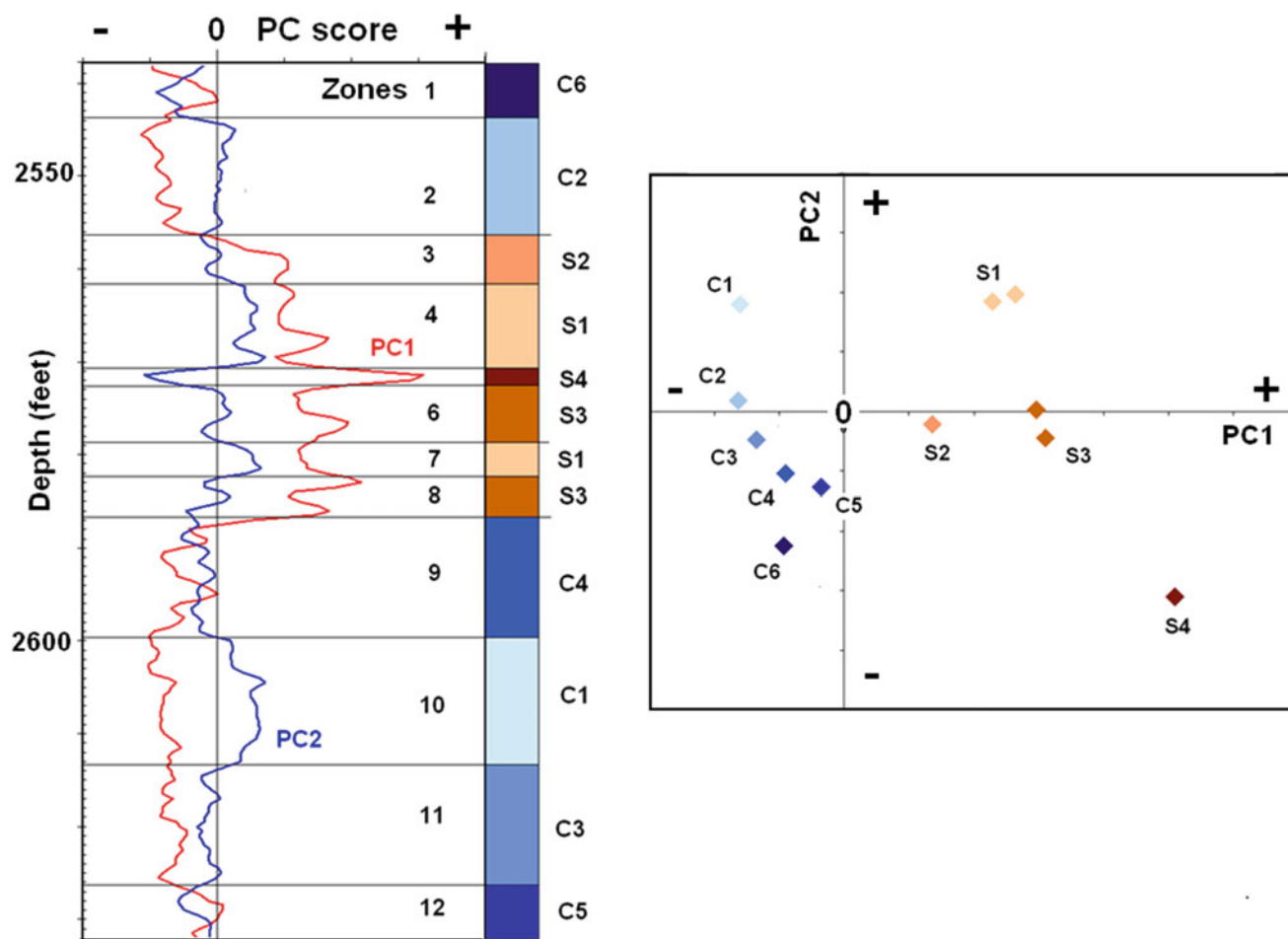
Supervised and Unsupervised Models

An electrofacies analysis that is given no prior information concerning group membership of individual observations is an “unsupervised” approach, where trends and clusters that are perceived in petrophysical measurements are subsequently assigned geological meaning. This inductive philosophy “allows the data to speak for themselves” and operates from the “bottom-up,” because the analysis originates with the observational data rather than being driven top-down by a deductive model. Advantages include the element of surprise, in that new associations may be revealed that represent useful information to augment the current geological model (Delfiner et al. 1987). An example is shown in Fig. 1, where a crossplot of the first two principal components of a petrophysical log suite (right) is used to identify electrofacies, which then are assigned to a Permian carbonate/clastic sequence (left). A comparison between this electrofacies succession and facies observed from core is shown in Fig. 2.

In a “supervised” method, different categories or groups are specified before the analysis; the goal of the method is to find the best function to distinguish the categories, based on the characteristics of the data. Subsequently, the function can be used to classify unknown observations on the basis of likelihood of membership in one or another of the groups. In this case, lithofacies are identified in the core and linked with log data from a cored well as a “training set,” from which a multivariate statistical method can “learn” the relationships between logs and core lithofacies. The matching of predicted lithofacies with modern depositional analogs means that dimensional measures, such as shape and lateral extent, can be used to correlate genetic layers across a field in a satisfactory geomodel. Most wells are logged rather than cored, so that lithofacies predicted from electrofacies are important in augmenting sparsely distributed cored wells.

Parametric Methods

A simple model of an electrofacies can be approximated by a cloud of points that are concentrated in the center and diffuse in density outward. A multivariate normal distribution provides a convenient means to model such a distribution efficiently. It also provides the basis for a probability model that



Electrofacies, Fig. 1 Crossplot of the first two principal components of a petrophysical log suite (right) from a Permian carbonate/clastic section is used to identify electrofacies, which then are assigned to a stratigraphic sequence (left)

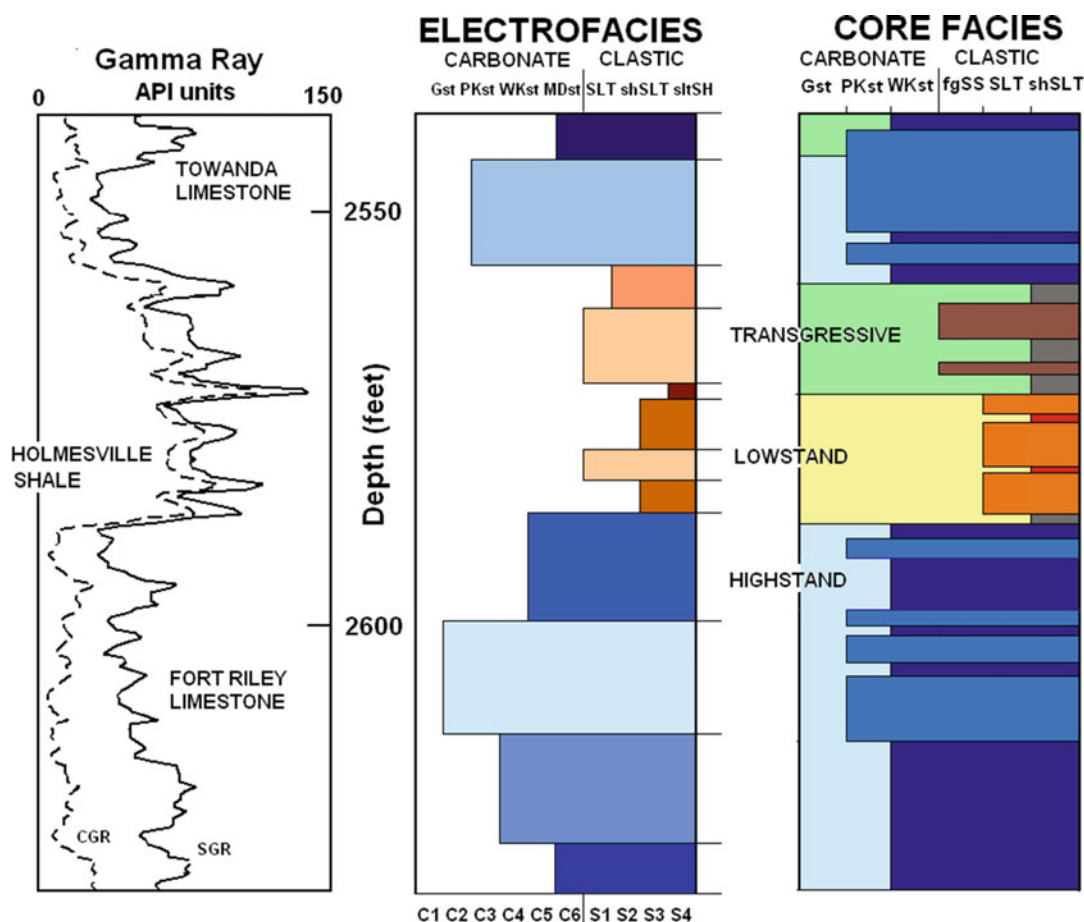
allows statistical classifications to be made from a database of multiple electrofacies. If approximately normal, then each electrofacies data cloud is a hyperellipsoid that can be specified completely by the statistics of the multivariate mean and the matrix of variances and covariances between the logs.

Although the locations of data points in multilog space collectively delineate clouds with the same number of dimensions as logs, they can often be mapped effectively in a much reduced dimensionality. This is because intercorrelations between the log variables cause the clouds to be extended along certain trends. Principal component analysis computes an ordered set of orthogonal axes that absorb the variation in a systematic manner (Wolff and Pelissier-Combescur 1982). Clustering is then run in two phases. In the first phase, the clustering is used to isolate local modes as zonal representatives of the digital data. In the second phase, the local modes are agglomerated into clusters that are equated with electrofacies. Decisions concerning potential cluster subdivision or fusion are monitored by referral to geological information from the core or from geological experience.

Alternatively, multiple discriminant analysis can be used to classify electrofacies that are keyed to a lithofacies database, first for classification and then for prediction (Doveton 1994).

Nonparametric and Neural Network Methods

The representation of electrofacies data clouds by multinormal hyperellipsoids is a simple model that is often a poorly adequate representation of real data variation. The richly connected structure of neural networks allows a more nuanced treatment of multiple log data (Doveton 2014). However, careful evaluations must be made of predictive results to ensure that the network is not overfitting the data. This is achieved by linking neural network training with validation data sets. The methodology is particularly appropriate when constructing large geomodels based on a well set that has been extensively logged, but with limited core data (Dubois et al. 2006).



Electrofacies, Fig. 2 A comparison between the unsupervised electrofacies succession and the facies observed from core examination

Summary

Electrofacies are clusters of multiple log responses that distinguish rock types in the subsurface. Their geological meaning can be interpreted by pattern recognition, but they more commonly are linked explicitly with lithofacies observed in core or from drill-cuttings. Electrofacies analysis is particularly important in the construction of geomodels, because core information is typically much less available than petrophysical logs in the majority of wells. Consequently, observations from core can be used to train logs in uncored wells in the prediction of lithofacies.

Cross-References

- [Discriminant Analysis](#)
- [Mathematical Minerals](#)
- [Neural Networks](#)
- [Principal Component Analysis](#)

- [Quantitative Stratigraphy](#)
- [Stratigraphic Sequence](#)

Bibliography

- Delfiner PC, Peyret O, Serra O (1987) Automatic determination of lithology from well logs: Society of Petroleum Engineers Formation Evaluation 2:303–310
- Doveton JH (1994) Geological log analysis using computer methods: computer applications in geology, No. 2. American Association of Petroleum Geologists, Tulsa, 169 pp
- Doveton JH (2014) Principles of mathematical petrophysics. Oxford University Press, 267 pp
- Dubois MK, Byrnes AP, Bohling GC, Doveton JH (2006) Multiscale geologic and petrophysical modeling of the giant Hugoton gas field (Permian), Kansas and Oklahoma. In: Harris PM, Weber LJ (eds) Giant hydrocarbon reservoirs of the world: From rocks to reservoir characterization and modeling. American Association of Petroleum Geologists Memoir 88, Tulsa, Oklahoma, pp 307–353
- Reading HG (1978) Facies. In: Reading HG (ed) Sedimentary environments and facies. Elsevier, New York, pp 4–14

- Serra, O, Abbott, HT (1980) The contribution of logging data to sedimentology and stratigraphy. Society of Petroleum Engineers, SPE 9270-MS, 19 p
- Wolff M, Pelissier-Combes J (1982) FACIOLOG – automatic electrofacies determination: transactions of the Society of Professional Well Log Analysts. In: 23rd annual logging symposium, paper FF, 22 p

Ensemble Kalman Filtering

J. Jaime Gómez-Hernández

Research Institute of Water and Environmental Engineering,
Universitat Politècnica de València, Valencia, Spain

Definition

The ensemble Kalman filter (EnKF) is an evolution of the Kalman filter for its application to nonlinear state-transition systems with a further extension to serve as a powerful parameter inversion method. Its main purpose is to improve the estimates of the system state as observations are acquired. As the Kalman filter, the EnKF is based on two steps, forecasting and updating (or filtering). During the forecasting step, the state of the system is forecasted on the basis of the latest state estimate; then, forecasts are compared with observations and a correction is made to the state (and possibly to the parameters of the state equation) that will serve as the basis for the next forecast.

The Kalman filter was developed by Kalman (1960) for the purpose of improving the trajectory estimation of spacecrafts and missiles and one of its very first implementations was in the Apollo program. The Kalman filter, as developed by Kalman, had a great success when it was first proposed since it provides optimal results for dynamic systems that are controlled by a linear state-transition equation; then, it went into oblivion because it did not work so well when dealing with nonlinear systems. New formulations such as the extended Kalman filter were postulated with limited success, but it was not until the work by Evensen (1994), who introduced the EnKF as a mean to deal with the nonlinearities, that the Kalman filter came back into the spotlight. This rebirth of the Kalman filter not only allowed a better modeling of nonlinear systems but also proved that it can be used as a very efficient inverse modeling approach for the identification of parameters of the state-transition equations.

The Kalman Filter

Consider a dynamic system evolving in time according to a state-transition equation. There are two ways to predict the

state system at time t : (i) by dead reckoning, that is, the state-transition model is run until time t given some known state at time 0, or (ii) by observing the state of the system using measuring probes. Neither way will provide an exact description of the state of the system at time t , the first one, because the state-transition model is always an approximation of the real system, and the second one, because the probes always have measurement errors and, generally, only sample sparsely the state itself. Kalman found a way to combine the two predictions into a better one in his formulation of his filter.

Kalman formulated his filter for a linear dynamic system

$$x^t = \mathcal{L}(x^{t-1}) \quad (1)$$

where x is the state of the system, and $\mathcal{L}$ the linear state-transition equation that gives the state of the system at time t given its state in a previous instant $t - 1$. Since most of the models used in the geosciences are built on discretized versions of the real system, a matrix notation can be used to describe the evolution of the system

$$\mathbf{x}^t = \mathbf{A} \cdot \mathbf{x}^{t-1} \quad (2)$$

where $\mathbf{x}$ is a column vector of size $N \times 1$ with the values of the state at the centers of the N voxels that discretize the study area, and $\mathbf{A}$ is a state-transition matrix of size $N \times N$.

Adopting a random function model, we can interpret each element of $\mathbf{x}$, at any time t , as a random variable, and the entire vector, at any time t , as a random function. As a consequence, the expected value of the random function propagates in time according to

$$E \{ \mathbf{x}^t \} = \mathbf{A} \cdot E \{ \mathbf{x}^{t-1} \}, \quad (3)$$

and its covariance, according to

$$C(\mathbf{x}^t) = \mathbf{A} \cdot C(\mathbf{x}^{t-1}) \cdot \mathbf{A}', \quad (4)$$

where C is a covariance matrix of size $N \times N$ containing all pairwise covariances between any two elements in the vector state and the $'$ symbol is used to denote transpose.

Now, consider that the state has been observed at M locations (generally $M \ll N$). These observations are represented by vector $\mathbf{y}$. For the sake of simplicity, assume that these observations have been taken at the centers of the voxels that discretize the study area and that measure directly the state of the system not a proxy. This simplification implies that the observation matrix $\mathbf{H}$ that will be used later should be made up of just zeroes and ones and it serves to extract the specific rows and columns from the vectors and matrices used in the formulation that follows corresponding to the observation locations. Consider also that these observations have un-

correlated measurement errors with zero mean; the error covariance $\mathbf{R}$ is, therefore, diagonal. Under these considerations, the conditional expectation of the state at time t given the observations taken at that same time is the optimal estimate in a least-square sense of the state of the system and is given by the following expression:

$$E\{\mathbf{x}^t|\mathbf{y}^t\} = E\{\mathbf{x}^t\} + \mathbf{K} \cdot (\mathbf{y}^t - \mathbf{H} \cdot E\{\mathbf{x}^t\}) \quad (5)$$

where $\mathbf{K}$ is the so-called Kalman gain matrix of size $N \times M$, the expression of which is

$$\mathbf{K} = C(\mathbf{x}^t) \cdot \mathbf{H}' \cdot (\mathbf{H} \cdot C(\mathbf{x}^t) \cdot \mathbf{H}' + \mathbf{R}^t)^{-1}. \quad (6)$$

The analysis of expression (5) shows that the conditional expectation is equal to the unconditional one corrected by a factor that is proportional to the discrepancy between the observations and the predicted expected data at time t obtained with Eq. (3). The proportionality term is itself proportional to the predicted state covariance at time t , and inverse proportional to the measurement error covariance and to the covariance between the locations at which the state has been observed.

The conditional covariance is given by

$$C(\mathbf{x}^t|\mathbf{y}^t) = (\mathbf{I} - \mathbf{K} \cdot \mathbf{H}) \cdot C(\mathbf{x}^t); \quad (7)$$

it is the covariance associated with the conditional expectation.

The Kalman filter consists of two steps. A forecast step, given by Eq. (3) and an update step given by Eq. (5). Alongside, the covariances are also forecasted and updated (Eqs. (4) and (7), respectively). The predictions are based on the linear state transition equation and the updates are the result of an optimization in some least squares sense. Once the updates are made, they become the basis for the next forecast and update from t to $t + 1$.

As described, the method is straightforward to implement and apply. One of its first applications was to help the guidance of the Apollo spacecrafts and it was coded in the Apollo computer, a computer with only 2 kilobytes of RAM, 36 kilobytes of ROM, and a chip running at 100 MHz.

The Extended Kalman Filter

Given its success with linear dynamic systems, an attempt to expand the Kalman filter to deal with nonlinear state transition equations was done with the Extended Kalman filter (Jazwinski 1970). When the system does not evolve linearly

$$\mathbf{x}^t = \Phi(\mathbf{x}^{t-1}) \quad (8)$$

with Φ being a nonlinear operator, a Taylor expansion can be applied to get a linear approximation

$$\mathbf{x}^t \approx \mathbf{A} \cdot \mathbf{x}^{t-1} + \dots \quad (9)$$

and

$$C(\mathbf{x}^t) \approx \mathbf{A} \cdot C(\mathbf{x}^{t-1}) \cdot \mathbf{A}' + \dots, \quad (10)$$

with

$$\mathbf{A} = \frac{\partial \Phi}{\partial \mathbf{x}^t} \quad (11)$$

being the Jacobian of the state transition equation. When only the first term of the expansion is retained, the formulation defaults to the Kalman filter formulation, which is then applied. The main problem with this approach is that, even if the Jacobian is reevaluated after each update step, the covariance estimate deteriorates quickly in time when Eq. (10) is used, yielding very poor results.

The Ensemble Kalman Filter

An improved method to compute the state covariances for nonlinear systems was proposed by Evensen (1994) circumventing the main problem of the extended Kalman filter. A random function can be defined by the statistics of the random variables that conform it, or as a collection of realizations. Evensen replaced working directly with the expected value and covariance of the random function by working with an ensemble of realizations, which are forecasted and updated as per Kalman filter theory. Then, he proposes to compute, experimentally, the expected value and covariance from the forecasted realizations. There is no need to approximate the forecast of neither the mean nor the covariance as in Eqs. (9) and (10); rather, the ensemble of realizations are forecasted in time using the nonlinear transfer function of Eq. (8) and the expected values and covariances are computed afterwards from the set of forecasted realizations.

Consider an ensemble of K realizations from the random function $\mathbf{x}$, at time $t - 1$, $\{\mathbf{x}_1^{t-1}, \mathbf{x}_2^{t-1}, \dots, \mathbf{x}_K^{t-1}\}$, each realization composed of N components corresponding to the N state values associated to the discretization of the study area. They propagate in time according to the nonlinear transfer function Φ

$$\{\mathbf{x}_1^t, \mathbf{x}_2^t, \dots, \mathbf{x}_K^t\} = \Phi\{\mathbf{x}_1^{t-1}, \mathbf{x}_2^{t-1}, \dots, \mathbf{x}_K^{t-1}\}. \quad (12)$$

The expected value and the covariance of the state random function can be approximated from this ensemble of forecasted realizations. Let $\mathbf{X}$ be an array with N rows and K columns containing all states for all realizations. The expected value is given by

$$E\{\mathbf{x}^t\} \approx \frac{1}{K} (\mathbf{X}^t \cdot \mathbf{1}_{K \times 1}) \quad (13)$$

where $\mathbf{1}_{K \times 1}$ is a column vector of K ones, and $\mathbf{X}^t$ contains all the forecasted realizations at time t . Likewise, the covariance is computed as

$$C(\mathbf{x}^t) \approx \frac{1}{K-1} (\mathbf{X}^t - E\{\mathbf{x}^t\} \cdot \mathbf{1}_{1 \times N}) \cdot (\mathbf{X}^t - E\{\mathbf{x}^t\} \cdot \mathbf{1}_{1 \times N})' \quad (14)$$

where $\mathbf{1}_{1 \times N}$ is a row vector of N ones. The Kalman gain is computed using Eq. (6) and all realizations are updated according to

$$\mathbf{X}^t | \mathbf{y}^t = \mathbf{X}^t + \mathbf{K} \cdot ((\mathbf{y}^t + \boldsymbol{\epsilon}^t) \cdot \mathbf{1}_{1 \times K} - \mathbf{H} \cdot \mathbf{X}^t) \quad (15)$$

where $\boldsymbol{\epsilon}^t$ is a vector of random observation errors drawn from the diagonal covariance matrix $\mathbf{R}$. The conditional expectation and the conditional co-variance could be computed, if needed, using $\mathbf{X}^t | \mathbf{y}^t$ in Eqs. (13) and (14) in replacement of $\mathbf{X}^t$. The conditional expectation would be the best estimate, in some least squares sense, of the state of the system at time t , and the conditional covariance a measure of the estimation uncertainty.

The ensemble of updated realizations becomes the new starting ensemble of realizations and the forecast and update steps are repeated starting at Eq. (12).

The algorithm for the ensemble Kalman filter is quite simple and applicable to any dynamic system; its steps would be to:

1. Generate an ensemble of realizations of the system state (these realizations can be generated by randomizing the parameters of the state equation and solving it, or by perturbing with noise a given solution of the state equation)
2. Forecast the state, realization by realization, until the next observation time (Eq. (8))
3. Compute the expected value and covariance of the forecasted realizations (Eqs. (9) and (10))
4. Compute the Kalman gain (Eq. (6))
5. Update the forecasted realizations (Eq. (15))
6. Compute the expected value and covariance of the updated realizations, if necessary (Eqs. (9) and (10))

7. Make the updated realizations the new starting realizations and return to the forecast step (Eq. (8))

The Ensemble Kalman Filter with Augmented State

The EnKF as proposed by Evensen solved the problem of the approximate calculation of the covariance matrices using an alternative (and better) approximation based on the statistical analysis of an ensemble of realizations, but its application had the same purpose as the original filter: the update of the system state each time new observations were obtained. But, it was realized that the method could be used to update the parameters defining the system, too. In modeling, for instance, an aquifer, these parameters could be material parameters such as permeability or porosity, forcing terms such as recharge or contaminant source locations, or boundary conditions such as lateral inflows/outflows or leakages. The idea consists of considering the parameters as part of an augmented state vector (Wen and Chen 2007), in which the parameters are updated each time a new observation is made.

The system state x is now augmented to include the last update of the model parameters α . The system state continues evolving in time according to a state-transition model Φ , but the model parameters do not evolve.

Forecast Eq. (3) becomes

$$\begin{Bmatrix} \alpha^t \\ x^t \end{Bmatrix} = \begin{Bmatrix} \alpha^{t-1} \\ \Phi(x^{t-1}) \end{Bmatrix} \quad (16)$$

Using z to represent the augmented state, a new forecast equation is defined

$$z^t = \Psi(z^{t-1}), \quad (17)$$

and the EnKF algorithm can be applied once a random function model is adopted for z . Each ensemble realization will contain a realization of the parameters and of the states associated to that parameter realization. The forecast will be performed using Eq. (17) on each realization, and the rest of the steps will follow the algorithm described in the previous section. Notice that new parameter observations can be included at the same time as new state observations.

This application of the EnKF becomes a powerful inverse modeling approach since, as observations are assimilated, improved estimates of the model parameters are obtained. The main criticism of this approach is the potential lack of consistency between the updated parameters and the updated state values. After all, the states should abide the principles under which the state-transition model is built, for instance, in the case of an aquifer modeling, the preservation of mass.

This possible inconsistency resulted in the proposal of a variant of the filter, the restart EnKF (Xu and Gómez-Hernández 2018), in which the system forecast is always made from time 0, rather than from the last updated estimate of the system state:

$$\begin{Bmatrix} \alpha^t \\ x^t \end{Bmatrix} = \begin{Bmatrix} \alpha^{t-1} \\ \Phi(x^0) \end{Bmatrix}. \quad (18)$$

The parameters keep being updated after each observation event, but the states are always recalculated from the state at time 0, x^0 , with the latest update of the parameters. This procedure obviously increases the running time of the algorithm but ensures that the state values are always consistent with the latest estimate of the model parameters.

The Normal-Score Ensemble Kalman Filter

There was still a pitfall in the application of the EnKF to nonlinear systems. The fact that all updates are made using covariances imbue a Gaussian flavor to the updated realizations, which, after several updates, have clear Gaussian attributes, starting by their histogram. In geosciences, many of the attributes have non-Gaussian characteristics, like, for instance, the permeabilities of a sand-shale aquifer have a histogram closer to a mixture of two lognormal distributions than to a Gaussian one. Even if the starting parameter realizations display the desired non-Gaussian characteristics, such as a bi-modality of the histograms, the final histograms of the parameter realizations get close to normal. To avoid this problem, the normal-score EnKF (NS-EnKF) was proposed by Zhou et al. (2011).

In the NS-EnKF, all the EnKF steps are performed in variables that are marginally Gaussian. The augmented state vector z is transformed into another one with marginal Gaussian distributions using a normal-score transform $\varphi(\cdot)$, which is a monotonic transformation that will yield a resulting variable with a standardized Gaussian histogram, then, given the normal-score transforms u^t and u^{t-1} of the augmented states z^t and z^{t-1}

$$u^{t-1} = \varphi^{t-1}(z^{t-1})$$

$$u^t = \varphi^t(z^t)$$

and given the state transition Eq. (17), a new forecast equation can be written for the normal-score variates

$$u^t = \Gamma(u^{t-1}), \quad (19)$$

with

$$\Gamma = \varphi^t \cdot \Psi \cdot (\varphi^{t-1})^{-1}. \quad (20)$$

The NS-EnKF proceeds as the EnKF but in normal-score space. At any time, the augmented state values z can be obtained by using the inverse of the normal-score transform $\varphi^{-1}(\cdot)$.

Kalman Filter and Kriging

The observant reader will find striking similarities between the update Eq. (5) and the simple kriging equation, with the Kalman gain being the kriging weights.

Summary and Conclusions

The ensemble Kalman filter is a variant of the Kalman filter capable to deal with dynamic nonlinear systems. It provides optimal estimates of the state of the system as well as of the parameters defining the state- transition model, which are updated sequentially in time as new observations are acquired. The rationale is simple as is its implementation, and could be used with virtually any model that allows predicting the state of the system in the future as a function of the state in the present. Its version using an augmented state vector has proven a very powerful and efficient inverse modeling approach.

Cross-References

- [Forward and Inverse Stratigraphic Models](#)
- [Inversion Theory](#)
- [Inversion Theory in Geoscience](#)
- [Kriging](#)
- [Monte Carlo Method](#)
- [Multivariate Analysis](#)
- [Random Function](#)
- [Random Variable](#)
- [Realizations](#)
- [Simulation](#)
- [Spatial Statistics](#)

References

- Evensen G (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research* 99(C5):10143–10162
- Jazwinski AH (1970) *Stochastic Processes and Filtering Theory*. Academic Press

- Kalman RE (1960) A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* 82:35–45
- Wen XH, Chen WH (2007) Some practical issues on real-time reservoir model updating using ensemble Kalman filter. *SPE Journal* 12(02): 156–166
- Xu T, Gómez-Hernández JJ (2018) Simultaneous identification of a contaminant source and hydraulic conductivity via the restart normal-score ensemble Kalman filter. *Advances in Water Resources* 112: 106–123
- Zhou H, Gómez-Hernández J, Hendricks Franssen H, Li L (2011) An approach to handling non-gaussianity of parameters and state variables in ensemble Kalman filtering. *Advances in Water Resources* 34(7):844–864

Entropy

Julian M. Ortiz¹ and Jorge F. Silva²

¹The Robert M. Buchan Department of Mining, Queen's University, Kingston, ON, Canada

²Information and Decision System Group (IDS), Department of Electrical Engineering, University of Chile, Santiago, Chile

Definition

The *Shannon entropy* is a fundamental measure in information theory (Cover and Thomas 2006). Claude Shannon introduced it in his landmark 1948 paper (Shannon 1948), to quantify the level of randomness or statistical disorder that can be used to determine (encode) a discrete random object. In its basic form, we have a discrete random variable Z with values in a finite alphabet $\mathcal{A} = \{1, \dots, M\}$ equipped with a *probability mass function* (pmf) denoted by $P_Z = (p_Z(i))_{i \in \mathcal{A}}$. Then, the entropy of Z (or the entropy of its distribution P_Z) is given by:

$$\mathcal{H}(Z) = \mathcal{H}(P_Z) \equiv - \sum_{i=1}^M p_Z(i) \log p_Z(i). \quad (1)$$

The term $\log 1/p_Z(i)$ in (Eq. 1) is known as *the self-information* of the event $\{Z = i\}$ (Cover and Thomas 2006), and it can be understood as the level of surprise or information we obtain when Z takes the value i in the execution of this random experiment. The smaller the probability (before the execution of the experiment) of observing $\{Z = i\}$, the bigger the information we gain when measuring this outcome. Consequently, $\mathcal{H}(Z)$ is the average self-information or the information we obtain from observing Z on average. Indeed, when Z is deterministic ($p_Z(i) = 1$ for some $i \in \mathcal{A}$), then $\mathcal{H}(Z) = 0$, and $\mathcal{H}(Z)$ achieves its maximum value $\log M$ if, and only if, P_Z is the uniform distribution in $\mathcal{A}$ (Cover and Thomas 2006).

In communication, the log is based 2, and the *second Shannon coding theorem* (Cover and Thomas 2006; Shannon

1948) shows that $\mathcal{H}(Z)$ determines the minimum number of bits per sample needed to compress Z . Therefore, $\mathcal{H}(Z)$ is not an abstract (inapplicable) concept in digital communication: Shannon shows that $\mathcal{H}(Z)$ precisely indicates how costly (in bits) is the task of compressing Z .

In geosciences, we are interested in random objects with a spatial structure and specific local statistical features. Entropy has been used to measure the complexity or uncertainty of the local features and as a metric of overall spatial complexity (Wellmann and Regenauer-Lieb 2012; Journel and Deutsch 1993). Another interesting application of entropy is in the design of sampling networks and experimental design, intending to define the optimum sampling locations to reduce the overall uncertainty in a random field (Santibañez et al. 2019; Wellmann and Regenauer-Lieb 2012).

Historical Development and Properties

Historical Development

The entropy is one of the most important quantities in information theory introduced by Shannon to characterize the fundamental performance limit for the task of (fixed-rate) almost lossless data compression (Shannon 1948). The entropy has found numerous applications in other problems and fields far beyond its original conception in data compression and communication. For instance, it has been considered in the simulation problem, for experimental design, for the optimal sampling problem, and as a principle for learning, to name some examples.

Basic Properties of $\mathcal{H}(Z)$

The entropy of Z is an indicator of the disorder or uncertainty of Z , and it satisfies the following important properties:

- $\mathcal{H}(Z) = 0$ if, and only if, Z is deterministic¹.
- $\mathcal{H}(Z) \leq \log M$ and $\mathcal{H}(Z) = \log M$ if, and only if, Z has the uniform distribution over $[M] \equiv \{1, \dots, M\}$.
- If we have two probabilities μ and ν on $[M]$ and $w \in (0, 1)$:

$$\mathcal{H}(\underbrace{w\mu + (1-w)\nu}_{\text{convex combination}}) \geq w \cdot \mathcal{H}(\mu) + (1-w)\mathcal{H}(\nu), \quad (2)$$

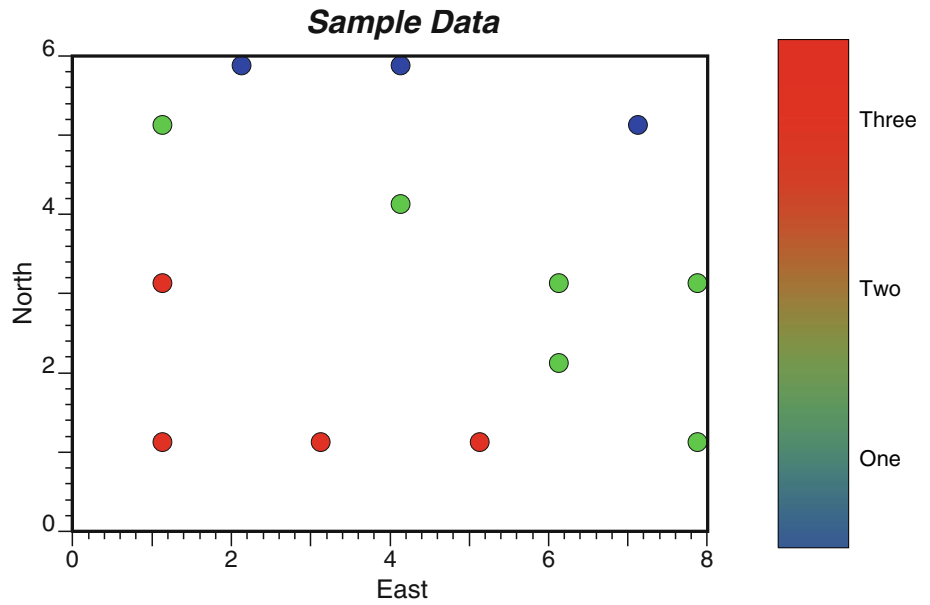
then the entropy is a concave functional of the probabilities.

Conditional Entropy and Mutual Information

Information can be conceptualized as the reduction of uncertainty of an object after measuring a random experiment

¹This means that $p_Z(i) = 1$ for some $i \in [M]$.

Entropy, Fig. 1 Location map of the samples in the domain, coded with the corresponding category code



(Cover and Thomas 2006). Therefore, if we observe (measure) Z , we gain $\mathcal{H}(Z)$ of information about Z , in average. The reason is that before measuring Z , we have $\mathcal{H}(Z)$ of prior uncertainty (about Z). After measuring it, we do not have uncertainty as we ultimately resolved Z . Therefore, measuring Z provides in average $\mathcal{H}(Z)$ information.

If we have two discrete random variables (Z_1, Z_2) with joint distribution P_{Z_1, Z_2} over the space $[M] \times [N]$, we can ask a similar question: how much information in average Z_2 offers about Z_1 . This question is fully determined by the statistical interaction between Z_1 and Z_2 expressed by their joint probability. Following the same rationale, the prior uncertainty of Z_1 is $\mathcal{H}(Z_1)$ and the posterior uncertainty of Z_1 after observing Z_2 in average is given (denoted) by the *conditional entropy*:

$$\mathcal{H}(Z_1|Z_2) \equiv \sum_{z_2 \in [N]} P_{Z_2}(z_2) \cdot \mathcal{H}(P_{Z_1|Z_2}(|z_2)) \quad (3)$$

where $P_{Z_1|Z_2}(|z_2)$ denotes the conditional distribution of Z_1 given the event $\{Z_2 = z_2\}$. Consequently from (Eq. 1), $\mathcal{H}(P_{Z_1|Z_2}(|z_2))$ is the entropy of Z_1 given that $Z_2 = z_2$, and $\mathcal{H}(Z_2|Z_1)$ in (Eq. 3) is the average of those conditional entropies with respect to P_{Z_2} . Therefore, the information we gain about Z_1 from observing (measuring) Z_2 in average is:

$$I(Z_1; Z_2) \equiv \mathcal{H}(Z_1) - \mathcal{H}(Z_1|Z_2), \quad (4)$$

which is known as the *mutual information* (MI) between Z_1 and Z_2 . Important properties of $I(Z_1; Z_2)$ are (Cover and Thomas 2006):

- $\mathcal{H}(Z_1|Z_2) \leq \mathcal{H}(Z_1)$ and, consequently, $I(Z_1; Z_2) \geq 0$.
- $\mathcal{H}(Z_1) - \mathcal{H}(Z_1|Z_2) = \mathcal{H}(Z_2) - \mathcal{H}(Z_2|Z_1) = I(Z_1; Z_2)$.

- $I(Z_1; Z_2) = 0$ if, and only if, Z_1 and Z_2 are independent.
- $\mathcal{H}(Z_1|Z_2) = 0$ if, and only if, Z_1 is a deterministic function of Z_2 .
- $I(Z_1; Z_1) = \mathcal{H}(Z_1)$.

Interestingly, the amount of information we gain about Z_1 by observing (measuring) Z_2 is the same as the amount of information we gain about Z_2 by observing (measuring) Z_1 .

Another worth mentioning characterization of $I(Z_1; Z_2)$ comes from the divergence (or Kullback-Leibler divergence). The divergence between two distributions μ and ν in a finite alphabet space $[M]$ is given by (Cover and Thomas 2006):

$$D(\mu||\nu) = \sum_{i \in [M]} \mu(i) \log \frac{\mu(i)}{\nu(i)} \geq 0 \quad (5)$$

if $\mu \ll \nu$ ² and otherwise $D(\mu||\nu) = \infty$. Interestingly, using (Eq. 4) it is possible to show that:

$$I(Z_1; Z_2) = D(P_{Z_1, Z_2} || P_{Z_1} \times P_{Z_2}) \quad (6)$$

where $P_{Z_1} \times P_{Z_2}$ denotes the product distribution (i.e., $P_{Z_1} \times P_{Z_2}(i, j) = P_{Z_1}(i) \cdot P_{Z_2}(j)$).

Differential Entropy and Multivariate Case

This basic definition in (Eq. 1) can be extended to the case of continuous variables (Renyi 1959) and to the

²If $\nu(A) = 0$ then $\mu(A) = 0$.

multivariate case (Cover and Thomas 2006), which can be used to characterize the bivariate spatial variability of a random field (Journel and Deutsch 1993). For the univariate continuous variable Z equipped with a probability density function (pdf) f_Z , Eq. 1 can be extended as (Cover and Thomas 2006):

$$\begin{aligned} h(Z) &= h(f_Z) \equiv E\{-\log f_Z(z)\} \\ &= -\int_{-\infty}^{+\infty} [\log f_Z(z)] f_Z(z) dz, \end{aligned} \quad (7)$$

which is known as the differential entropy of Z (or f_Z).

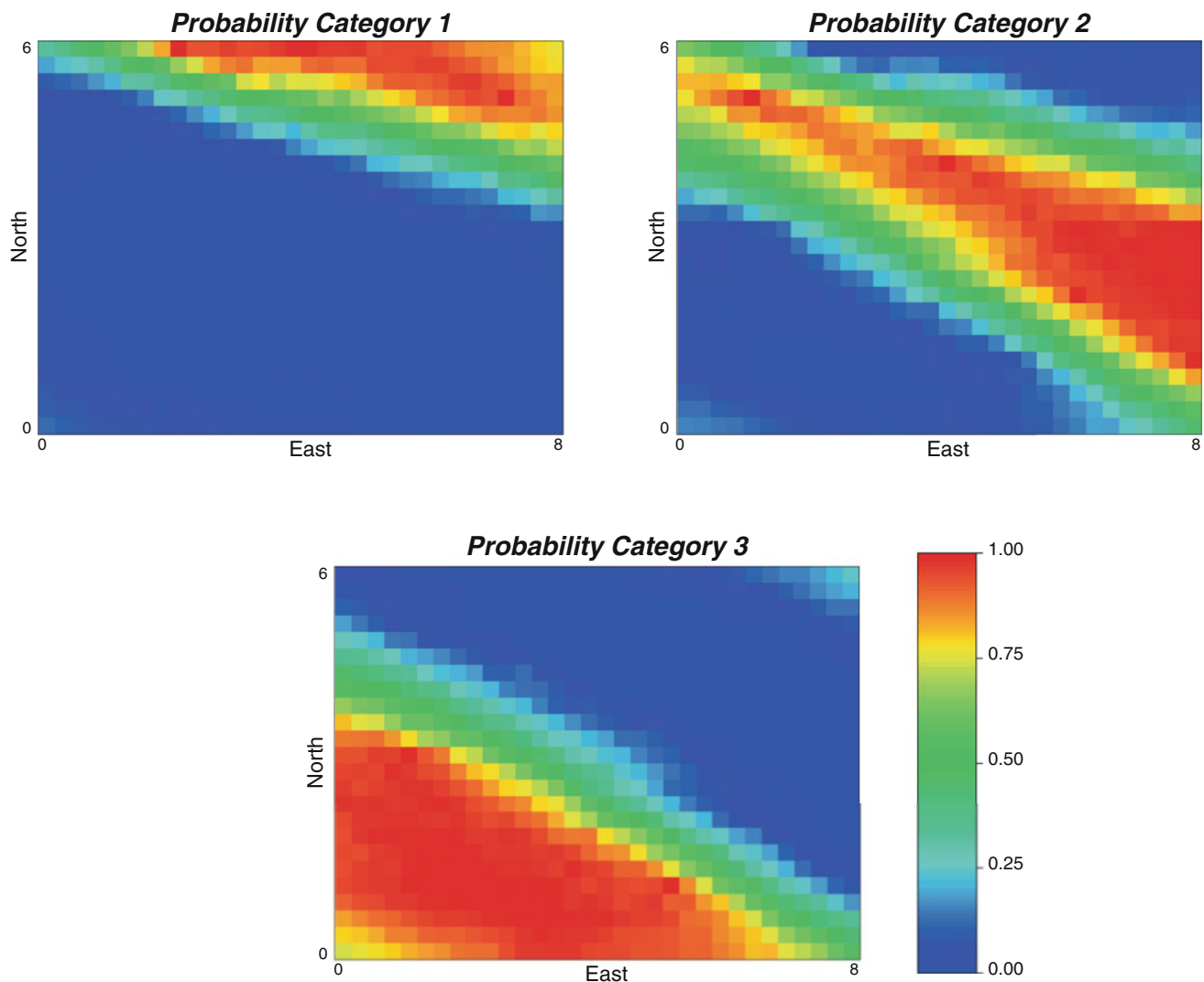
Suppose we have a random object $(Z(u))_{u \in \mathcal{D}}$ distributed over a spatial domain $\mathcal{D}$. In that case, the bivariate (continuous) differential entropy can be computed between

two points with a lag separation $\mathbf{h}$. For instance, we can consider the joint differential entropy between $Z(\mathbf{u})$ and $Z(\mathbf{u} + \mathbf{h})$:

$$\begin{aligned} h(\mathbf{u}, \mathbf{u} + \mathbf{h}) &\equiv h(Z(\mathbf{u}), Z(\mathbf{u} + \mathbf{h})) = \\ &= -\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [\log f_{Z(\mathbf{u}), Z(\mathbf{u} + \mathbf{h})}(z, z')] f_{Z(\mathbf{u}), Z(\mathbf{u} + \mathbf{h})}(z, z') dz dz'. \end{aligned} \quad (8)$$

In a similar way to Eq.(4), the mutual information between $Z(\mathbf{u})$ and $Z(\mathbf{u} + \mathbf{h})$ is (Cover and Thomas 2006):

$$\begin{aligned} I(\mathbf{u}; \mathbf{u} + \mathbf{h}) &\equiv I(Z(\mathbf{u}); Z(\mathbf{u} + \mathbf{h})) \\ &= D\left(P_{Z(\mathbf{u}), Z(\mathbf{u} + \mathbf{h})} \parallel P_{Z(\mathbf{u})} \times P_{Z(\mathbf{u} + \mathbf{h})}\right) \end{aligned} \quad (9)$$



Entropy, Fig. 2 Maps of computed probabilities of prevalence of each category, using indicator simulation

$$= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \log \frac{f_{Z(\mathbf{u}), Z(\mathbf{u}+\mathbf{h})}(z, z')}{f_{Z(\mathbf{u})}(z)f_{Z(\mathbf{u}+\mathbf{h})}(z')} f_{Z(\mathbf{u}), Z(\mathbf{u}+\mathbf{h})}(z, z') dz dz'. \quad (10)$$

Applications

Characterizing Spatial Uncertainty

Entropy can be used as a measure of local uncertainty in a spatial context. In order to illustrate this notion, let us consider a simple two-dimensional domain, where a few locations have been sampled, and a category is assigned after logging or analyzing each sample (Fig. 1). These categories can be facies, rock types, or land use, for example. At unsampled locations, the uncertainty can be characterized in different ways. In this example, we calculate the entropy conditioned to the few logged (sampled) locations.

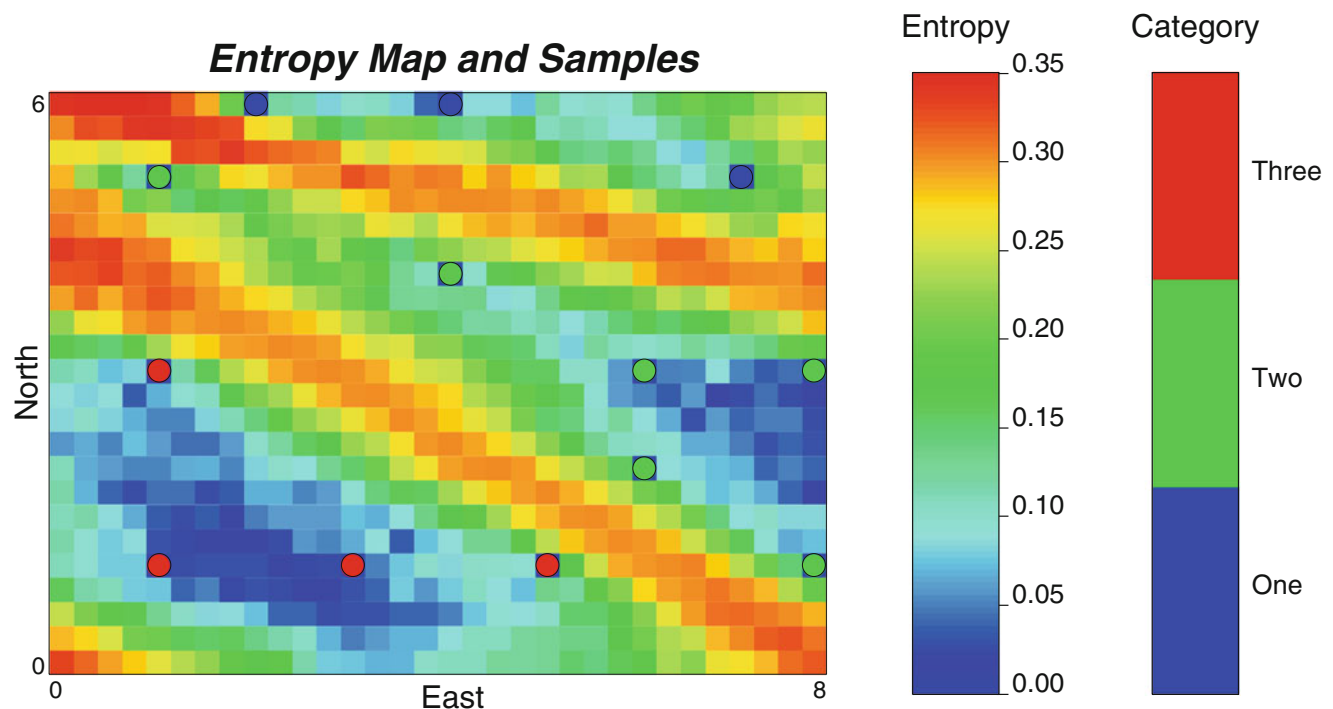
Indicator simulation is used to create multiple realizations of the categorical random function, from which the probability of each category prevailing at locations over the domain is calculated. Figure 2 shows the probability of finding each category at every location in the simulation grid. This result provides the conditional local probability mass distribution of the categorical variable at every unsampled location. We can compute the local entropy by applying Eq. 1. Figure 3 shows the map of local entropies. This map is consistent with the data and their spatial continuity. The high entropy values are located in areas with

higher uncertainty: in the contacts between the different categories and in areas poorly informed.

Maximum Entropy Principle for Sampling Design

From the previous analysis, entropy can be used to determine the sampling location that maximizes the information gain. This can be achieved by computing the Entropy before and after the sample is gathered. In a spatial model, it is important to notice that one sample will affect the uncertainty in neighboring locations. This influence is a function of the spatial correlation of the random function within the domain. Therefore, for sampling design, the decision does not seek the location with the highest entropy but rather the location that globally maximizes the reduction in entropy over the entire domain.

When a sampling network composed of r samples is needed, the problem becomes more complex, as the number of combinations to be assessed increases significantly with the number of samples r . Assuming there are n sample locations possible, the solution space becomes the combination $C(n, r)$, which rapidly becomes very large. Further strategies to deal with the sampling design problem are discussed in (Santibañez et al. 2019).



Entropy, Fig. 3 Map of local entropy computed from the conditional local probability mass function at every location

Summary and Final Remarks

Entropy is a fundamental tool to measure information content, uncertainty, or disorder. Its original formulation, originated in communication problems, has been adapted to be used in geoscience problems and expanded as a tool for spatial characterization.

Entropy is directly linked to information content and can be used as a metric for information gain, becoming a valuable tool to assess the value of new evidence (measurements).

Bibliography

- Cover TM, Thomas JA (2006) Elements of information theory, 2nd edn. Wiley Interscience, New York
- Hansen TM (2020) Entropy and information content of geostatistical models. *Math Geosci* 53:163–184
- Journel AG, Deutsch CV (1993) Entropy and spatial disorder. *Math Geol* 25(3):329–355
- Renyi A (1959) On the dimension and entropy of probability distributions. *Acta Math Acad Sci Hung* 10:193–215
- Santibañez F, Silva JF, Ortiz JM (2019) Sampling strategies for uncertainty reduction in categorical random fields: formulation, mathematical analysis and application to multiple-point simulations. *Math Geosci* 51(5):579–624
- Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27(379–423):623–656
- Wellmann JF, Regenauer-Lieb K (2012) Uncertainties have a meaning: information entropy as a quality measure for 3-d geological models. *Tectonophysics* 526:207–216

Euler Poles of Tectonic Plates

Helmut Schaeben¹, Uwe Kroner¹ and Tobias Stephan²

¹Technische Universität Bergakademie Freiberg, Freiberg, Germany

²Department of Geoscience, University of Calgary, Calgary, AB, Canada

Definition

Since Wegener's hypothesis of continental drift, geoscientists have developed a model of the Earth's evolution in terms of plate tectonics emphasizing horizontal motions of near-surface plates as opposed to vertical motions. Neglecting interactions of plates at their boundaries, their motions are bound to be rotations. A geometrically appealing and instructive parametrization of rotations is provided by their angles and axes, whose intersections with the sphere modeling the shape of the Earth are referred to as Euler poles. It seems tempting and challenging to represent the Earth's evolution in terms of interdependent transient rotations of a few major plates.

Introduction

The uppermost shell of the Earth, that is, the lithosphere, can be subdivided into several rigid, approximately 100-km-thick tectonic plates. These plates move over a weak mantle layer, the asthenosphere. Because of the differential kinematics, adjacent tectonic plates interact at their boundaries causing first-order geologic processes subsumed under the term global plate tectonics. There are three types of plate boundaries. Divergent plate boundaries, also called constructive plate boundaries, are characterized by the production of new lithosphere along a global system of mid-ocean ridges (MOR). Sea floor spreading at divergent plate boundaries is balanced by the consumption of lithosphere at convergent plate boundaries also referred to as destructive plate boundaries. At such a boundary, the leading edge of a tectonic lower plate subducts into deeper parts of the mantle beneath tectonic upper plates. Conservative plate boundaries are the third type of plate boundaries, lithosphere is neither generated nor consumed. Along transform faults, adjacent plates move laterally to each other thereby connecting segments of divergent and/or convergent plate boundaries. The modeling assumption of plate rigidity is not generally valid for its boundaries (Kreemer et al. 2014). Especially, convergent plate boundaries along continent-continent collision zones, that is, orogenic belts, are rather diffuse exhibiting a complex deformation pattern. Nevertheless, the geometry of the rigid parts of tectonic plates can be approximated by spherical polygons, their kinematics can be described in terms of rotations, their angles, and Euler poles, that is, the points of intersection of the axes of rotations and the sphere modeling the shape of the Earth (Bullard et al. 1965). Special attention is required to distinguish finite rotations from infinitesimal rotations. While finite rotations by a finite angle ω apply to represent rotational states, infinitesimal rotations by an infinitesimal small angle $\delta\omega$ refer to a (piecewise) circular motion. Most notably, infinitesimal rotations are commutative, finite rotations are not.

For the analysis of plate motions, an internal reference system and two external reference systems have been applied. Relative plate motions describe the kinematics of a pair of plates, one assumed to be moving and the other one assumed to be fixed. Then the fixed plate constitutes the reference system. In contrast, absolute plate motions reflect the movement of plates with respect to a global reference system. Classically, the deep mantle reference accounts for the absolute motion of the plates (DeMets et al. 1990) assuming a fixed system of hot spots (Morgan 1971) and thus a quasi-stationary mantle beneath the lithosphere. With the development of space geodesy, the International Celestial Reference Frame (ICRF) is available (www.iers.org) employing external quasi-stationary extragalactic sources, that is, quasars. The ICRF constitutes the reference system used by the satellites of the Global Positioning System (GPS) and allows for the

formulation of high precision earth model, that is, the International Terrestrial Reference Frame (ITRF). Analyzing the motion of GPS ground stations relative to the fixed Earth model results in plate kinematic models for recent times (Sella et al. 2002; Kreemer et al. 2014).

In terms of current plate motions, different plate kinematic models exist and thus different Euler pole solutions. Their differences originate from applying different reference systems, considering a different total number of plates and the limitations of the particularly specified data sets. For example, there is a significant deviation of global kinematics of absolute plate motions derived from the deep mantle reference system “HS3-NUVEL 1A” (Gripp and Gordon 2002) compared with global kinematics derived from the analysis of relative plate motions with “NUVEL 1A” (DeMets et al. 1994). The transformation from “HS3-NUVEL 1A” to “NUVEL 1A” requires an additional rotation of the entire lithosphere, that is, the net rotation. Plate kinematic models derived from space geodesy commonly assume no-net rotation, that is, the current motion of all points of the earth sums up to zero (Kreemer et al. 2014; Drewes 2009). In terms of rigid lithospheric plates, three adjacent plates P_1 , P_2 , P_3 meet at one common point, that is, the triple junction. Performing a plate circuit, that is, taking successively the plates P_i , $i = 1, 2, 3$, as fixed and watching each plate P_j , $j = 2, 3, 1$, moving according to an infinitesimal rotation by $\delta\omega$ relative to the fixed one about the axes $\mathbf{e}_{12}$, $\mathbf{e}_{23}$, $\mathbf{e}_{31}$ with rotation vectors (Euler vectors) $\boldsymbol{\omega}_{ij} = (\delta\omega)\mathbf{e}_{ij}$, the rotation vectors must sum up to zero (McKenzie and Morgan 1969; Le Pichon et al. 1973). To avoid any confusion, the plate circuit constraint is a mathematical property of infinitesimal rotations; it is not a geological property of tectonic plates. In general, plate circuits can be applied to several three-plate configurations each with a common triple junction. If the plate circuit constraint is valid for all configurations, the no-net rotation condition is fulfilled (DeMets et al. 1994).

To compare different plate kinematic models, the reader is referred to the plate motion calculator provided by UNAVCO (<https://www.unavco.org/software/geodetic-utilities/geodetic-utilities.html>).

Plate kinematic models of the past are much less precise due to the lack of space geodetic constraints. Moreover, constraints derived from the oceanic lithosphere are only available for Ceno- and Mesozoic times. Since almost all of the pre-Mesozoic oceanic lithosphere has been subducted at convergent plate boundaries, constraints for ancient plate kinematics are exclusively provided from the continental crust. Latitudinal information can be deduced from paleomagnetic, paleobiogeographic, and paleoclimatic studies (Scotese and McKerrow 1990). As with respect to the longitudinal position of the investigated region, information is not available. Hence, such data sets do not allow for an exact positioning of continental plates for various time steps and

thus for the formulation of precise plate kinematic models in pre-Mesozoic times, respectively. Euler poles of relative plate motions can be deduced from the spherical strain pattern of ancient plate boundary zones (Kroner et al. 2016) and have been applied for plate tectonic processes culminating in the late Paleozoic formation of the supercontinent Pangea (Kroner et al. 2016, 2020). Since such paleotectonic reconstructions initially consider plate motions relative to a fixed plate, the latitudinal positions of the continents must be approximated in a second step using appropriate data sets mentioned above.

Modeling the Geometry of Tectonic Plates and the Kinematics of Their Motion

Geometric Modeling of Tectonic Plates

The shape of the Earth is modeled as unit sphere $\mathbb{S}^2 \subset \mathbb{R}^3$ in three-dimensional Euclidean space. The geometry of the interior rigid core of tectonic plates is represented by spherical polygons, that is, closed geometric figures on the sphere, whose vertices are points of the sphere and whose edges are arcs of great circles connecting the vertices. In contrast to the Euclidean real plane, the complement of a spherical polygon is also a spherical polygon. Any motion of a rigid spherical polygon on the sphere is determined by rotations in $\mathbb{R}^3$. Euler’s theorem on rotations (Eulero 1775; Palais and Palais 2007) actually states that in three-dimensional Euclidean space, any rigid motion of a body leaving one of its points fixed is a unique rotation about some axis passing through the fixed point. This theorem implies that the sequential composition, that is, concatenation of two rotations, is a rotation, too. Euler’s theorem is the initializing account of modern mathematics of rotations (Goldstein 1950; Altmann 2013). Several parametrizations of a rotation exist; the parametrization in terms of its angle and axis appears self-evident, geometrically instructive, and practically appealing. In particular, it leads to an algebra to concatenate rotations explicitly in terms of unit quaternions.

Rotations in $\mathbb{R}^3$

In Euclidean geometry, a rotation is a linear transformation that preserves distances, that is, an isometry specified by its properties that it leaves at least one point fixed and that it preserves handedness. A rotation in three-dimensional Euclidean $\mathbb{R}^3$ is specified by an axis of its fixed points and an invariant plane of rotation orthogonal to that axis. Thus, a rotation R in $\mathbb{R}^3$ is uniquely defined in terms of its angle $\omega \in (-\pi, \pi]$ and its axis $\mathbf{e} \in \mathbb{S}^2$ of rotation, $R = R(\omega, \mathbf{e})$, as $R(\omega, \mathbf{e}) = R(-\omega, -\mathbf{e})$. The rotational axis passes through the center of the sphere and intersects the sphere at two antipodal points referred to as Euler poles in geology. Casually, Euler poles and axis of rotation provided by a unit vector are not distinguished here.

The terms are used synonymously and the common symbol is $\mathbf{e} \in S^2 \subset \mathbb{R}^3$. Keeping the angle ω of rotation fixed, the rotation $R(\omega, -\mathbf{e})$ is the inverse rotation of $R(\omega, \mathbf{e})$, that is, $R(\omega, -\mathbf{e})R(\omega, \mathbf{e}) = R(\omega, \mathbf{e})R(\omega, -\mathbf{e}) = \text{id}$, where id denotes the identical rotation by an angle $\omega = 0$.

According to Rodrigues' formula (Rodrigues 1840), applying the rotation $R(\omega, \mathbf{e})$ to the unit vector $\mathbf{u} \in S^2$ results in

$$R(\omega, \mathbf{e})\mathbf{u} = \mathbf{u} \cos \omega + (\mathbf{e} \times \mathbf{u}) \sin \omega + \mathbf{e}(\mathbf{e} \cdot \mathbf{u})(1 - \cos \omega) \quad (1)$$

$$= \mathbf{u} + (\mathbf{e} \times \mathbf{u}) \sin \omega + \mathbf{e} \times (\mathbf{e} \times \mathbf{u}) 2 \sin^2 \frac{\omega}{2}. \quad (2)$$

Otherwise, rotations in $\mathbb{R}^3$ may be expressed as (3×3) matrix of the special orthogonal group $\text{SO}(3)$ or as unit quaternion $q \in \mathbb{S}^3 \subset \mathbb{H}$, the skew field of real quaternions (Hamilton 1844),

$$q = \cos \frac{\omega}{2} + \mathbf{e} \sin \frac{\omega}{2}, \quad (3)$$

which seems to provide the most versatile and useful form close to the angle-axis parametrization, cf. Quaternions and Rotations, this Encyclopedia. Rodrigues formula, Eq. 1, is actually the result of the quaternion multiplication $q\mathbf{u}q^*$, where $\mathbf{u}$ is taken as pure unit quaternion and q^* denotes the conjugate unit quaternion of the unit quaternion q . Let the rotations $R_1 = R_1(\omega_1, \mathbf{e}_1)$ and $R_2 = R_2(\omega_2, \mathbf{e}_2)$ be associated with unit quaternions $q_1 = \cos \frac{\omega_1}{2} + \mathbf{e}_1 \sin \frac{\omega_1}{2}$ and $q_2 = \cos \frac{\omega_2}{2} + \mathbf{e}_2 \sin \frac{\omega_2}{2}$. Then, the concatenation $R = R(\omega, \mathbf{e}) = R_2(\omega_2, \mathbf{e}_2)R_1(\omega_1, \mathbf{e}_1)$ is associated with the quaternion product q_2q_1 .

$$\begin{aligned} q_2q_1 &= \cos \frac{\omega_2}{2} \cos \frac{\omega_1}{2} - \mathbf{e}_2\mathbf{e}_1 \sin \frac{\omega_2}{2} \sin \frac{\omega_1}{2} \\ &+ \mathbf{e}_1 \sin \frac{\omega_1}{2} \cos \frac{\omega_2}{2} + \mathbf{e}_2 \sin \frac{\omega_2}{2} \cos \frac{\omega_1}{2} + \mathbf{e}_2 \times \mathbf{e}_1 \sin \frac{\omega_1}{2} \sin \frac{\omega_2}{2}, \end{aligned} \quad (4)$$

and its scalar and vector part, respectively,

$$\cos \frac{\omega}{2} = \cos \frac{\omega_2}{2} \cos \frac{\omega_1}{2} - \mathbf{e}_2\mathbf{e}_1 \sin \frac{\omega_2}{2} \sin \frac{\omega_1}{2}, \quad (5)$$

$$\begin{aligned} \mathbf{e} \sin \frac{\omega}{2} &= \mathbf{e}_1 \sin \frac{\omega_1}{2} \cos \frac{\omega_2}{2} + \mathbf{e}_2 \sin \frac{\omega_2}{2} \cos \frac{\omega_1}{2} + \\ &\mathbf{e}_2 \times \mathbf{e}_1 \sin \frac{\omega_1}{2} \sin \frac{\omega_2}{2} \end{aligned} \quad (6)$$

provide the angle ω and the axis $\mathbf{e}$ of the rotation $R(\omega, \mathbf{e})$.

Whatever its form, a rotation as defined up to now is a transformation of coordinates. It describes a rotational state. It does not describe a dynamic process like a motion. There is no rationale yet of an angular velocity.

Kinematic Modeling of Plate Motion-Finite Rotations

Rotating a polygon P given by the sequence of its vertices $(\mathbf{u}_1, \dots, \mathbf{u}_n)$ results in the rotated polygon $P' = R(\omega, \mathbf{e})P$ with vertices $\mathbf{w}_i = R(\omega, \mathbf{e})\mathbf{u}_i$, $i = 1, \dots, n$, of the same shape and size as a rotation is first of all an isometry, that is, preserving distances. If a presumably unique rotation is initially unknown and only matched pairs $(\mathbf{u}_i, \mathbf{w}_i) \in S^2 \times S^2$ with $\mathbf{w}_i \approx R(\omega, \mathbf{e})\mathbf{u}_i$, $i = 1, \dots, n$, have been observed, then $R(\omega, \mathbf{e})$ can be estimated by spherical regression (MacKenzie 1957; Chang 1986, 1993; Meister 2001) where only the latter reference gives the solution in terms of quaternions.

Successive concatenation of a sequence of several rotations $R_\ell = R(\omega_\ell, \mathbf{e}_\ell)$, $\ell = 1, 2, \dots$, results in a sequence $(P, R_1P, R_2R_1P, \dots, R_\ell R_{\ell-1} \dots R_1P, \dots)$, of several rotational states. If the axes of rotations remain the same, $\mathbf{e}_\ell = \mathbf{e}$, $\ell = 1, 2, \dots$, then the ℓ -fold concatenation simplifies, that is, $R_\ell R_{\ell-1} \dots R_1 = R\left(\sum_{i=1}^{\ell} \omega_i, \mathbf{e}\right)$, $\ell = 1, 2, \dots$

Assigning increasing temporal marks $t_0 < t_1 < \dots < t_\ell < \dots$ to the rotational states, the association of every temporal mark t_ℓ to exactly one rotational state $R_\ell R_{\ell-1} \dots R_1P$ results in a time series, which in turn provides a first-order discrete model of a proceeding rotational motion of the polygon depending on the finite time lags $\Delta t_\ell = t_\ell - t_{\ell-1}$, $\ell = 1, 2, \dots$, in terms of a sequence of finite rotations describing the progressive change from one rotational state of the polygon to another rotational state. The first-order model immediately allows to calculate the rotation rate, that is, the magnitude $\frac{\omega_\ell}{\Delta t_\ell}$ of the vector $\boldsymbol{\omega}_\ell = \frac{1}{\Delta t_\ell} \boldsymbol{\omega}_\ell = \frac{\omega_\ell}{\Delta t_\ell} \mathbf{e}_\ell$ of the mean angular velocity over the time interval $(t_{\ell-1}, t_\ell]$, $\ell = 1, 2, \dots$. The magnitude of the corresponding mean linear velocity $\mathbf{v}_\ell$ along the small circle path of rotation with radius ρ is $\|\mathbf{v}_\ell\| = \frac{\omega_\ell}{\Delta t_\ell} \rho = \frac{\omega_\ell}{\Delta t_\ell} \sin \angle(\mathbf{e}_\ell, \mathbf{u}) = \|\boldsymbol{\omega}_\ell \times \mathbf{u}\|$. More generally, the rotations resulting from concatenation can be temporally indexed, for example, $R_\ell R_{\ell-1} \dots R_1 = R(\omega(t_\ell), \mathbf{e}(t_\ell))$, where $(\omega(t_\ell), \mathbf{e}(t_\ell))$ are the parameters of the ℓ -fold concatenation. If for all $\ell = 1, 2, \dots$ the axes of rotations are the same, $\mathbf{e}_\ell = \mathbf{e}$, then $\mathbf{e}(t_\ell) = \mathbf{e}$ and $\omega(t_\ell) = \sum_{i=1}^{\ell} \omega_i$. In this case, mean velocities $\frac{\omega(t_j) - \omega(t_i)}{t_j - t_i} = \frac{\sum_{\ell=i}^j \omega_\ell}{t_j - t_i}$ with respect to larger time intervals $(t_i, t_j]$ can be calculated analogously. Otherwise, these velocities are the apparent mean velocities referring to the small circle path of concatenated rotations. The true mean velocity during the period $(t_i, t_j]$ referring to the path composed of finite arcs of small circles of the sequence $(R_\ell, R_{\ell+1}R_\ell, \dots, R_{\ell_j} \dots R_{\ell_i+1}R_{\ell_i})$ of rotations can be calculated as the weighted mean of the individual arc-wise mean velocities.

Kinematic Modeling of Plate Motion-Infinesimal Rotations

A proceeding rotational motion is described by an infinitesimal rotation. If the angle of rotation is infinitesimal, $(\delta\omega) = O(\varepsilon)$, $0 \leq \varepsilon \ll 1$, then $\sin(\delta\omega) = (\delta\omega) + O(\varepsilon^3)$ and $\cos(\delta\omega) = 1 + O(\varepsilon^2)$, and the first-order approximation of $R(\omega, \mathbf{e})\mathbf{u}$ of Eq. 2 is

$$R((\delta\omega), \mathbf{e}) \mathbf{u} = \mathbf{u} + (\mathbf{e} \times \mathbf{u})(\delta\omega). \quad (7)$$

In contrast to finite rotations, infinitesimal rotations commute

$$R((\delta\omega), \mathbf{e}_2)R((\delta\omega), \mathbf{e}_1)\mathbf{u} = \mathbf{u} + (\mathbf{e}_1 \times \mathbf{u})(\delta\omega) + (\mathbf{e}_2 \times \mathbf{u})(\delta\omega) \quad (8)$$

$$= \mathbf{u} + ((\mathbf{e}_1 + \mathbf{e}_2) \times \mathbf{u})(\delta\omega) \quad (9)$$

$$= R((\delta\omega), \mathbf{e})\mathbf{u},$$

and their axes $\mathbf{e}_i$, $i = 1, 2$, sum to the axis $\mathbf{e} = \mathbf{e}_1 + \mathbf{e}_2$ of their concatenation, Eq. 9, that is, concatenation of rotations is turned into summation of their axes for infinitesimal rotations. Moreover, for infinitesimal rotations, the rotation vectors satisfy what is often referred to as closure constraint

$$(\delta\omega)\mathbf{e}_1 + (\delta\omega)\mathbf{e}_2 - (\delta\omega)\mathbf{e} = \boldsymbol{\omega}_1 + \boldsymbol{\omega}_2 - \boldsymbol{\omega} = 0, \quad (10)$$

that is, the three rotation vectors are coplanar.

The notion of an infinitesimal rotation gives rise to the infinitesimal generator of the group $SO(3)$ such that any finite rotation results from integrating the infinitesimal generator. Moreover, the notion of infinitesimal rotations applies to represent the motion of a rigid body driven by transient rotations in terms of a differential equation, where the path of the moving body is composed of infinitesimal small arcs of different small circles. A comprehensive and consistent dynamic model of the movement of tectonic plates throughout geological periods requires a representation in terms of differential equations involving infinitesimal rotations. To the best of our knowledge, such a model has still to be developed. Here it is sufficient to confine ourselves to finite rotations and time series of concatenated finite rotations.

Rotating Several Polygons

Rotating two polygons by a unique rotation, their relative location does not change, the effect of an isometry. That is to say that observed from one polygon, the other polygon is immobile. If two rotations are involved, each referring to one of the two polygons, which differ in their angles or axes of rotation, then the relative location of the two polygons does

no longer remain the same. In fact, if the two rotations differ in their (fixed) axes, then observed from one polygon, the other polygon is rotated by some angle about a migrating axis of rotation. The axes of rotations governing the motion of each polygon are referred to as absolute Euler poles; the migrating axis is referred to as relative Euler pole. If the two absolute Euler poles coincide, then the relative Euler pole coincides with the two. This setting is elaborated as follows.

Path of Migrating Euler Pole

At time t_0 , there are two spherical polygons $P_i(t_0)$, $i = 1, 2$. For the example given below, a single spherical triangle suffices to illustrate the migration. In the most simple case, two fixed Euler poles and corresponding axes of rotations $\mathbf{e}_i \in \mathbb{S}^2$, $i = 1, 2$, are assumed about which the polygons $P_i(t_0)$ are rotated by angles $\omega_i(t_\ell)$, respectively. Here t_ℓ is rather a parameter referring to a sequence $t_0 < t_1 < \dots < t_\ell < \dots$ than a continuous variable to suggest that the assumption of stationary absolute Euler poles could be dropped to get absolute Euler poles $\mathbf{e}_i(t_\ell) \in \mathbb{S}^2$ moving in time in the sense of concatenating finite rotations. In either case, at time $t_\ell > t_0$, there are two polygons $P_i(t_\ell)$, $i = 1, 2$,

$$P_i(t_\ell) = R_i(t_\ell)P_i(t_0) = R(\omega_i(t_\ell), \mathbf{e}_i(t_\ell))P_i(t_0), i = 1, 2. \quad (11)$$

The unique rotation $R(t_\ell) = R(\omega(t_\ell), \mathbf{e}(t_\ell)) = R_2(t_\ell)R_1^{-1}(t_\ell)$ about the migrating relative Euler pole $\mathbf{e}(t_\ell)$ by the angle $\omega(t_\ell)$ takes $P_1(t_\ell)$ to $\tilde{P}_1(t_\ell)$ where

$$\begin{aligned} \tilde{P}_1(t_\ell) &= R(\omega(t_\ell), \mathbf{e}(t_\ell))P_1(t_\ell) \\ &= R(\omega_2(t_\ell), \mathbf{e}_2(t_\ell))R(-\omega_1(t_\ell), \mathbf{e}_1(t_\ell))P_1(t_\ell) \\ &= R(\omega_2(t_\ell), \mathbf{e}_2(t_\ell))P_1(t_0). \end{aligned} \quad (12)$$

Obviously, the relative location of the polygons $P_2(t_\ell)$ and $\tilde{P}_1(t_\ell)$ is the same as of the initial polygons $P_2(t_0)$ and $P_1(t_0)$ as the latter polygons were subject to the same rotation $R(\omega_2(t_\ell), \mathbf{e}_2(t_\ell))$. In turn, observed from polygon $P_2(t_\ell)$ (apparently immobile), polygon $P_1(t_\ell)$ is rotated by $R(t_\ell) = R_2(t_\ell)R_1^{-1}(t_\ell)$ or, equivalently, polygon $\tilde{P}_1(t_\ell)$ is rotated by $R^{-1}(t_\ell) = R_1(t_\ell)R_2^{-1}(t_\ell)$.

Angle and axis of $R(t_\ell)$ are given as follows. If the axes of rotations $\mathbf{e}_1(t_\ell)$ and $\mathbf{e}_2(t_\ell)$ are identical, the solution for $R(t_\ell)$ is trivial, that is, $\mathbf{e}(t_\ell) = \mathbf{e}_1(t_\ell) = \mathbf{e}_2(t_\ell)$ and $\omega(t_\ell) = \omega_2(t_\ell) - \omega_1(t_\ell)$. If $\mathbf{e}_1(t_\ell)$ and $\mathbf{e}_2(t_\ell)$ are stationary, so is $\mathbf{e}(t_\ell)$. If the two absolute Euler poles are assumed to be different, $\mathbf{e}_1(t_\ell) \neq \mathbf{e}_2(t_\ell)$, then the solution is a straightforward application of concatenation in terms of unit quaternions (Altmann 2013, Kuipers 1999, Quaternions and Rotations, this Encyclopedia) with $q_i(t_\ell) =$

$\cos\left(\frac{\omega_i(t_\ell)}{2}\right) + \mathbf{q}_i(t_\ell)$ and $\mathbf{q}_i(t_\ell) = \sin\left(\frac{\omega_i(t_\ell)}{2}\right) \mathbf{e}_i(t_\ell)$, $i = 1, 2$, and $\mathbf{q}_1^*(t_\ell) = \sin\left(\frac{-\omega_1(t_\ell)}{2}\right) \mathbf{e}_1(t_\ell) = -\sin\left(\frac{\omega_1(t_\ell)}{2}\right) \mathbf{e}_1(t_\ell)$

$$\omega(t_\ell) = 2 \arccos \left(\cos \left(\frac{\omega_2(t_\ell)}{2} \right) \cos \left(\frac{\omega_1(t_\ell)}{2} \right) - \mathbf{q}_2(t_\ell) \cdot \mathbf{q}_1^*(t_\ell) \right) \quad (13)$$

$$\mathbf{e}(t_\ell) = \frac{1}{\sin \frac{\omega(t_\ell)}{2}} \left(\mathbf{q}_1^*(t_\ell) \cos \left(\frac{\omega_2(t_\ell)}{2} \right) + \mathbf{q}_2(t_\ell) \cos \left(\frac{\omega_1(t_\ell)}{2} \right) + \mathbf{q}_2(t_\ell) \times \mathbf{q}_1^*(t_\ell) \right). \quad (14)$$

It is emphasized that the relative Euler pole $\mathbf{e}(t_\ell)$ is migrating even if the two absolute Euler poles $\mathbf{e}_1(t_\ell)$ and $\mathbf{e}_2(t_\ell)$ are stationary.

If the two absolute Euler poles $\mathbf{e}_1(t_\ell)$ and $\mathbf{e}_2(t_\ell)$ are assumed to be stationary, that is, $\mathbf{e}_i(t_\ell) = \mathbf{e}_i$, $i = 1, 2$, and if $|\omega_1(t_\ell)| = |\omega_2(t_\ell)|$ is assumed, the former equations, Eqs. (13) and (14), simplify to

$$\omega(t_\ell) = 2 \arccos \left(\cos^2 \left(\frac{\omega_1(t_\ell)}{2} \right) - \mathbf{q}_2(t_\ell) \cdot \mathbf{q}_1^*(t_\ell) \right) \quad (15)$$

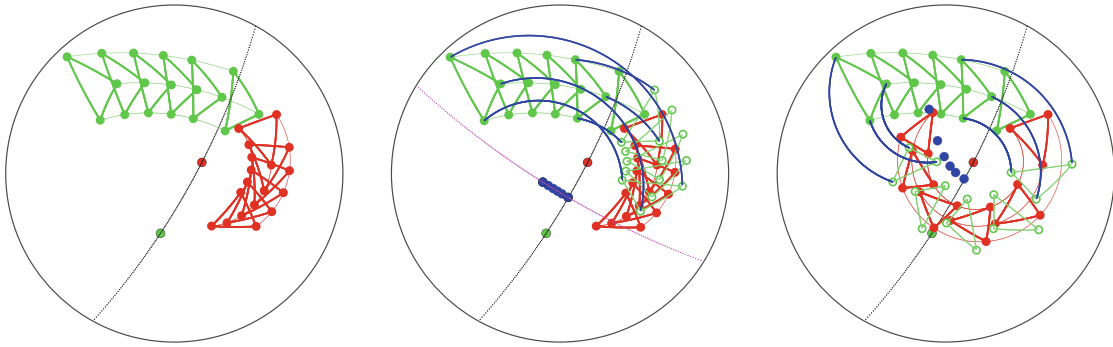
$$\mathbf{e}(t_\ell) = \frac{1}{\sin \frac{\omega(t_\ell)}{2}} \left((\mathbf{q}_1^*(t_\ell) + \mathbf{q}_2(t_\ell)) \cos \left(\frac{\omega_1(t_\ell)}{2} \right) + \mathbf{q}_2(t_\ell) \times \mathbf{q}_1^*(t_\ell) \right). \quad (16)$$

Since the assumption of stationary rotation axes implies that the directions of $\mathbf{q}_1$ and $\mathbf{q}_2$ are stationary, Eq. (16) reveals that if

(i) $\omega_2(t_\ell) = \omega_1(t_\ell)$, the migrating relative Euler pole is an element of the intersection of the plane spanned by $\mathbf{e}_2 - \mathbf{e}_1$ and $\mathbf{e}_2 \times \mathbf{e}_1$ and the unit sphere $\mathbb{S}^2 \subset \mathbb{R}^3$, (ii) $\omega_2(t_\ell) = -\omega_1(t_\ell)$, the migrating relative Euler pole is an element of the intersection of the plane spanned by $\mathbf{e}_1 + \mathbf{e}_2$ and $\mathbf{e}_1 \times \mathbf{e}_2$ and the unit sphere $\mathbb{S}^2$, that is, the relative Euler pole migrates along a great circle in both cases. Thus, it proves what Fig. 1 suggests that for angles $|\omega_1(t_\ell)| = |\omega_2(t_\ell)|$, that is, for equal angular velocities, the relative Euler pole $\mathbf{e}(t_\ell)$ migrates either along the great circle $C\left(\frac{\mathbf{e}_2 - \mathbf{e}_1}{\|\mathbf{e}_2 - \mathbf{e}_1\|}, \frac{\mathbf{e}_2 \times \mathbf{e}_1}{\|\mathbf{e}_2 \times \mathbf{e}_1\|}\right)$ or along the great circle $C\left(\frac{\mathbf{e}_1 + \mathbf{e}_2}{\|\mathbf{e}_1 + \mathbf{e}_2\|}, \frac{\mathbf{e}_1 \times \mathbf{e}_2}{\|\mathbf{e}_1 \times \mathbf{e}_2\|}\right)$, both orthogonal to the great circle $C(\mathbf{e}_1, \mathbf{e}_2)$ spanned by the two fixed absolute Euler poles $\mathbf{e}_1$ and $\mathbf{e}_2$. Figure 1 also confirms that for an infinitesimal angle $\delta\omega$, the relative Euler pole and the two absolute Euler poles are coplanar. Moreover, Eqs. (13) and (15), respectively, indicate that the velocity of the migration is not constant.

Equations (13) and (14) are easily generalized to consider Euler poles $\mathbf{e}_1(t)$, $\mathbf{e}_2(t)$ moving with respect to finite time steps by means of concatenating finite rotations. The path of migration of the Euler pole $\mathbf{e}(t)$ will be more involved than arcs along a unique great circle. To represent the path of the continuously migrating Euler pole corresponding to rotations with both their parameters varying continuously with time requires the solution of a differential equation involving infinitesimal rotations.

The geometric setting can be generalized to N different polygons and $M \leq N$ different Euler poles of their authorized rotations. For N different polygons, there are $\frac{(N-1)N}{2}$ pairs of different polygons. The relative motion of the two polygons of a pair is governed by rotations with a migrating Euler pole if the two corresponding absolute Euler poles are different.



Euler Poles of Tectonic Plates, Fig. 1 Equal angle projection of upper hemisphere

Left: Initially, the green and the red triangle are close to each other with one edge almost adjacent. Then they are successively rotated, in the sense of finite rotations, the green triangle counter-clockwise by $\omega_1, \omega_2, \dots, \omega_5$ about its fixed absolute Euler pole $\mathbf{e}_1$ (green dot) and the red triangle clockwise by $-\omega_1, -\omega_2, \dots, -\omega_5$ about its fixed absolute Euler pole $\mathbf{e}_2$ (red dot). Watching the sphere from outside, the triangles' vertices apparently move along arcs of small circles colored green or red, respectively, centered at their corresponding fixed absolute Euler poles $\mathbf{e}_1$ and $\mathbf{e}_2$, respectively. Both the absolute and the relative locations of the triangles are changing. Center: To preserve the relative location of the rotated green triangles with respect to the rotated red triangles, the vertices of the green triangles

have apparently to move along arcs of small circles colored blue centered at their corresponding relative Euler poles $\mathbf{e}(t_\ell)$ (blue dot) migrating along a great circle orthogonal to the great circle spanned by the two fixed axes $\mathbf{e}_1$ and $\mathbf{e}_2$. That is, when watching the rotated green triangles from the rotated red triangles assumed to be immobile, the green triangles appear to be rotated about the migrating relative Euler poles $\mathbf{e}(t_\ell)$ (blue dot). For an infinitesimal angle $\delta\omega$, the three axes $\mathbf{e}(t_\ell)$, $\mathbf{e}_1$, and $\mathbf{e}_2$ are coplanar and located on the great circle spanned by the two fixed axes $\mathbf{e}_1$ and $\mathbf{e}_2$.

Right: If the red triangle is rotated clockwise by $-3\omega_1, -3\omega_2, \dots, -3\omega_5$ about its fixed absolute Euler pole $\mathbf{e}_2$ (red dot), then the relative Euler poles $\mathbf{e}(t_\ell)$ (blue dot) of the apparent motion of the green triangles as watched from the red triangles do not migrate along a great circle.

Conclusions

Finite and infinitesimal rotations must not be confused. Only the infinitesimal rotations give rise to the global plate circuit closure as applied in space geodesy focusing on the determination of the plates' velocities today. In geology/tectonics, emphasis is put on how plates have moved, that is, on how the locations of the plates have changed over geological times of hundreds of millions years. Since the change of locations can be associated with different rotational states, finite rotations by large angles of Euler poles apply to represent the motion of plates. As shown, for a set of stationary absolute Euler poles, the resulting relative Euler poles migrate in a complex manner with time. Dropping the assumption of stationary absolute Euler poles, the complexity of the plates' motion increases further and requires more involved models than presented here.

Bibliography

- Altmann SL (2013) Rotations, quaternions, and double groups. Dover Publications, Mineola/New York
- Bullard E, Everett JE, Smith AG (1965) The Fit of the Continents around the Atlantic. *Philos Trans R Soc A Math Phys Eng Sci* 258:41
- Chang T (1986) Spherical regression. *Ann Stat* 14:907
- Chang T (1993) Spherical regression and the statistics of tectonic plate reconstructions. *Int Stat Rev* 61:299
- DeMets C, Gordon RG, Argus DF, Stein S (1990) Current plate motions. *Geophys J Int* 101:425
- Drewes H (2009) The actual plate kinematic and crustal deformation model APKIM2005 as basis for a non-rotating ITRF. In: Drewes H (ed) *Geodetic reference frames*. IAG symposium, Springer, Munich/Berlin/Heidelberg, 9–14 Oct 2006, pp 95–99
- Eulero L (Leonhard Euler) (1775) *Novi Comm Acad Sci Petropolitanae* 20:189–207
- Goldstein H (1950) *Classical mechanics*. Addison-Wesley, Reading, Massachusetts, USA
- Gripp AE, Gordon RG (2002) Young tracks of hotspots and current plate velocities. *Geophys J Int* 150:321
- Hamilton WR (1844) On quaternions; or a new system of imaginaries in algebra. *Phil Mag 3rd Ser* 25:489
- Kreemer C, Blewitt G, Klein EC (2014) A geodetic plate motion and Global Strain Rate Model. *Geochem Geophys Geosyst* 15:3849
- Kroner U, Roscher M, Romer RL (2016) Ancient plate kinematics derived from the deformation pattern of continental crust: Paleo- and Neo-Tethys opening coeval with prolonged GondwanaLaurussia convergence. *Tectonophysics* 681:220
- Kroner U, Stephan T, Romer RL, Roscher M (2020) Paleozoic plate kinematics during the PannotiaPangaea supercontinent cycle. *Special publications 503*. Geological Society, London
- Kuipers JB (1999) Quaternions and rotation sequences: a primer with applications to orbits, aerospace, and virtual reality. Princeton University Press, Princeton, New Jersey, USA
- Le Pichon X, Francheteau J, Bonnin J (1973) Plate tectonics: developments in Geotectonics 6. Elsevier, Amsterdam - London - New York
- MacKenzie JK (1957) The estimation of an orientation relationship. *Acta Crystallogr* 10:61
- McKenzie DP, Morgan WJ (1969) Evolution of triple junctions. *Nature* 224:125
- Meister L (2001) Quaternion optimization problems in engineering. In: Bayro Corrochano E, Sobczyk G (eds) *Geometric algebra with applications in science and engineering*, Birkhäuser, pp 387–412
- Morgan WJ (1971) Convection plumes in the lower mantle. *Nature* 230:42
- Palais B, Palais R (2007) Eulers xed point theorem: The axis of a rotation. *J Fixed Point Theory Appl* 2:215
- Rodrigues O (1840) Des lois geometriques qui regissent les déplacements d'un systeme solide dans l'espace, et de la variation des coordonnees provenant de ses déplacements consideres independamment des causes qui peuvent les produire. *J Math Pures Appl Liouville* 5:380
- Scotese CR, McKerrow WS (1990) Revised World maps and introduction. *Geol Soc Lond Mem* 12:1–21
- Sella GF, Dixon TH, Mao A (2002) REVEL: A model for Recent plate velocities from space geodesy. *J Geophys Res* 107:ETG11–ETG11

Expectation-Maximization Algorithm

Jaya Sreevalsan-Nair

Graphics-Visualization-Computing Lab, International
Institute of Information Technology Bangalore, Electronic
City, Bangalore, Karnataka, India

Definition

Expectation-maximization (EM) is a methodology for estimating parameters of statistical models of data from observed samples using an iterative approach (Moon 1996; Do and Batzoglou 2008). As the name suggests, the EM algorithm may include several instances of statistical model parameter estimation using observed data. However, it has come to be used widely in clustering and classification methods, which are predominantly covered in this chapter.

The algorithm was named as such in the seminal paper, also referred to as Dempster–Laird–Rubin method (Dempster et al. 1977), where it has been built on similar existing methods. This paper had generalized the algorithm for several problems and provided the convergence analysis of the iterative method. The paper refers to observations as *incomplete data*, represented using two sample spaces $\mathcal{X}$ and $\mathcal{Y}$, and a many-to-one mapping from $\mathcal{X}$ to $\mathcal{Y}$. The observed data is the set $\mathbf{y} \subset \mathcal{Y}$, and its corresponding data vector to infer $\mathbf{x} \subset \mathcal{X}$ is indirectly observed or realized through $\mathbf{y}$. The complete family of sampling density is given by $f(\mathbf{x}|\boldsymbol{\phi})$, and that for the incomplete data is $g(\mathbf{y}|\boldsymbol{\phi})$, where $\boldsymbol{\phi}$ is the set of parameters for a mixture model that models the sample spaces. The assumption made here is that the observed data is sampled from a larger dataset that can have different subpopulations modeled using a different statistical model. The relationship is given as $g(\mathbf{y}|\boldsymbol{\phi}) = \int_{\mathbf{x}(\mathbf{y})} f(\mathbf{x}|\boldsymbol{\phi})d\mathbf{x}$. The EM algorithm is motivated to estimate $\boldsymbol{\phi}$ that maximizes $g(\mathbf{y}|\boldsymbol{\phi})$.

Overview

The EM algorithm is usually defined with the exponential family of distributions, notably the Gaussian or normal distribution. Thus, often the EM algorithm is described with the use of Gaussian mixture model (GMM). For parameter estimation, the EM algorithm has two steps in each iteration: expectation (E-step) and maximizations (M-step). The complete data is now defined using exponential family:

$$f(\mathbf{x}|\boldsymbol{\phi}) = b(\mathbf{x}) \exp(\boldsymbol{\phi} \cdot \mathbf{t}(\mathbf{x})^T) / a(\boldsymbol{\phi}),$$

where the vectors $\boldsymbol{\phi}$, $\mathbf{t}(\mathbf{x}) \in \mathbb{R}^1 \times r$ and T is the matrix transpose.

The i th iteration is defined as

E-step Estimating the complete data *sufficient statistics* $\mathbf{t}(\mathbf{x})$ using

$$Q(\mathbf{t}|\mathbf{t}^{(i)}) = \mathbf{t}^{(i)} = E(\mathbf{t}(\mathbf{x})|\boldsymbol{\phi}^{(i)}).$$

M-step Determining the parameter

$$\begin{aligned} E(\mathbf{t}(\mathbf{x})|\boldsymbol{\phi}) &= \mathbf{t}^{(i)}, \text{ achieved by maximization of } Q, \boldsymbol{\phi}^{(i+1)} \\ &= \arg \max_{\boldsymbol{\phi}} Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(i)}). \end{aligned}$$

The iterations are terminated when there is no change in the parameters, thus indicating a stable solution.

The function $\mathbf{x}$ can use the maximum likelihood (ML) function $\mathcal{L}(\boldsymbol{\phi}, \mathcal{X}, \mathcal{Y})$, for which log-likelihood function is best suited. The log-likelihood function has the desirable property of its derivative being expressed as the difference of an unconditional and a conditional expectation of the sufficient statistics (Dempster et al. 1977). Since there is no guarantee of a global maximum, local maximum values are determined. Also, there has been relevant work where maximum a posteriori (MAP) estimate has been used instead of ML function.

For mixture models, the mixing proportion value α_i must be determined for $i = 1, \dots, m$ components.

$$f(\mathbf{x}|\boldsymbol{\phi}) = \sum_{i=1}^m \alpha_i f_i(\mathbf{x}|\boldsymbol{\phi}_i), \text{ for } f_i \text{ which are individual model}$$

descriptions. A detailed description of a working toy example of the implementation of the EM algorithm is given by Plasse (2013). The implementation involves an initial assumption for the following values: (i) mixing proportion α_i , (ii) mean or the average μ_i of subpopulation in each component, and (iii) covariance matrix Σ_i to be the estimated and updated values in each step until the separated components are achieved in a

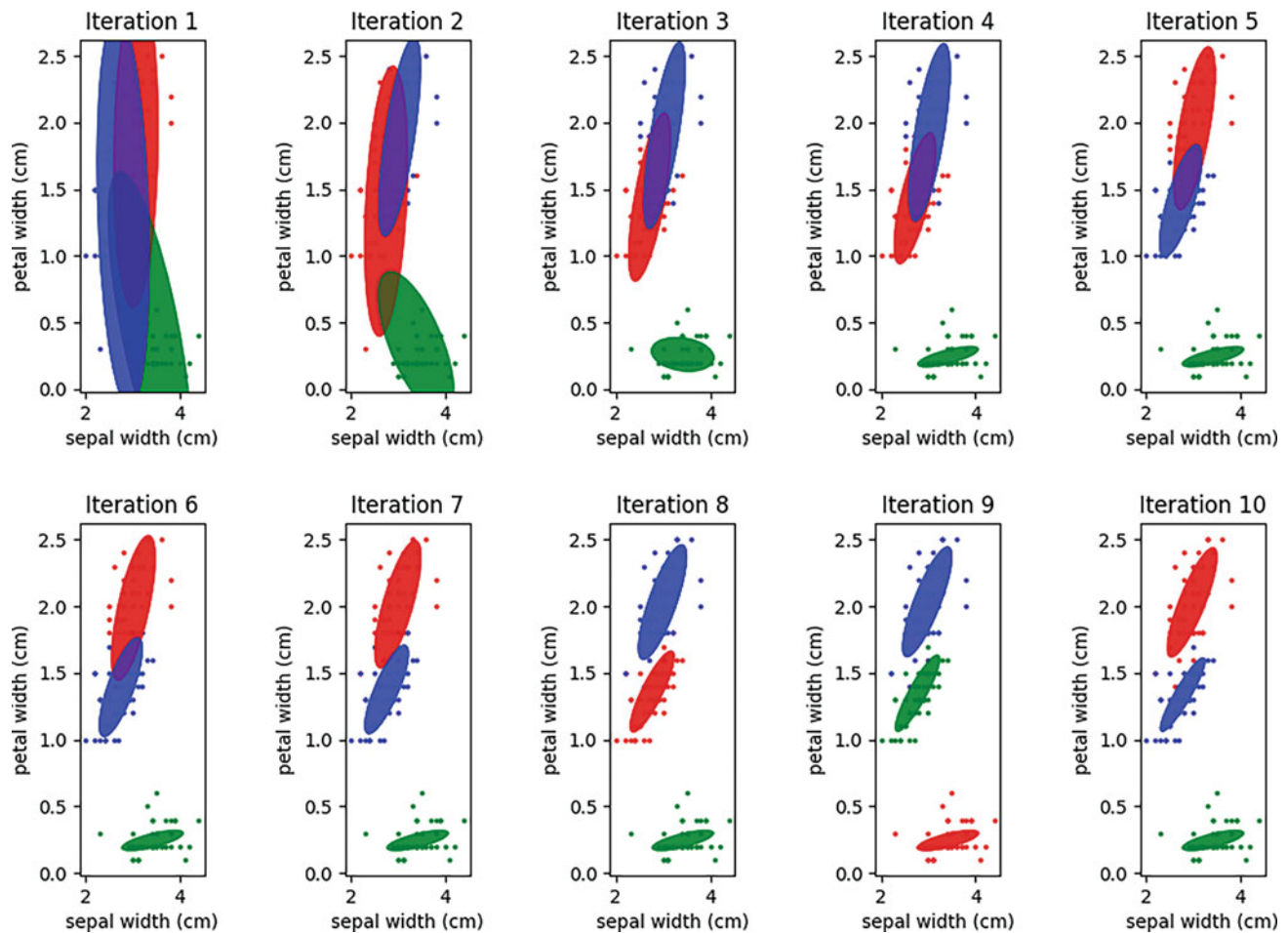
few iterations. An appropriate initialization can help reduce the number of iterations. Initialization can be improved by using conventional clustering algorithms, e.g., K-means clustering, to roughly estimate the subpopulations in components.

Figure 1 demonstrates the implementation of the EM algorithm using the Python library `sklearn` on the Iris dataset, which is a toy example of mostly well-separated clusters. Depending on the number of variables in the observed data, the functions used are usually multivariate Gaussian distributions to improve the model approximation (Do and Batzoglou 2008). The EM algorithm works best with well-separated clusters, without the requirement of circular-shaped clusters, as is the case for K-means clustering, when using Euclidean distance. However, it does not work effectively for ill-separated clusters, as shown in Fig. 2. The degree of class separability is reflected by the cluster separability, and thus, influences the number of iterations for convergence. More iterations are needed for ill-separated clusters than for the well-separated ones. Depending on the degree of class separability, acceleration methods, such as the ones used for fixed-point iteration, enable reducing the number of iterations needed for convergence (Plasse 2013).

Here, the *Iris* and *Forest cover-type* datasets, which are widely used machine learning datasets, are studied in Figs. 1 and 2. The *Iris* dataset includes physical dimensions of samples of different flowers belonging to the three species of the *Iris* family. The *Forest cover-type* dataset includes cartographic variables captured for each observed sample in 30×30 meter cell, obtained in the US Forest Service (USFS) Region 2 Resource Information System. The data for these variables have been procured from both the US Geological Survey (USGS) and USFS data sources for the samples collected from four different wilderness areas. The *Iris* dataset has one class that is linearly separable from the other two, which improves the degree of separation of clusters. In the *Forest cover-type* dataset, the wilderness areas are statistically significant for classification. However, since these areas are represented as categorical variables and not numerical variables, the EM algorithm does not work efficiently in its direct form with the less significant numerical variables. In such cases, the EM algorithm has to use the ML estimation used for mixed continuous and categorical data (Little and Schluchter 1985). The ML estimates used here are for a parameterization of the joint distribution of continuous and categorical variables in incomplete data.

Applications with Focus on Remotely Sensed Images

Apart from identifying the parameters to determine the mixture models of observed data, the EM algorithm also enables



Expectation-Maximization Algorithm, Fig. 1 Using `sklearn` package in Python, the Gaussian mixture model produces clusters as shown, where the clustering changes iteratively before stabilizing at the ninth iteration for the *Iris* dataset. The colors distinguish the clusters

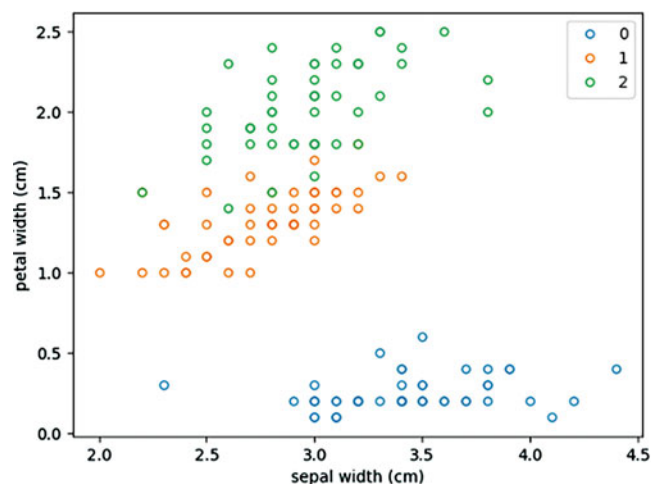
alone and are not tagged to specific clusters. The ellipses indicate the Gaussian functions, their mean, and variance. (Image source: author's own; dataset source: <https://archive.ics.uci.edu/ml/datasets/iris>)

knowledge discovery. Several processes tend to have latent variables, which can be discovered by explicitly including them in the model without the knowledge of their statistical significance. In addition to the tabular or multivariate data in the toy examples in Figs. 1 and 2, the EM algorithm is widely used with image data. Image processing applications use the algorithm to determine class probabilities and classify the pixels in the images for semantic segmentation, or the images themselves. A set of representative diverse applications in the use of EM algorithm for remotely sensed images is reported in this chapter.

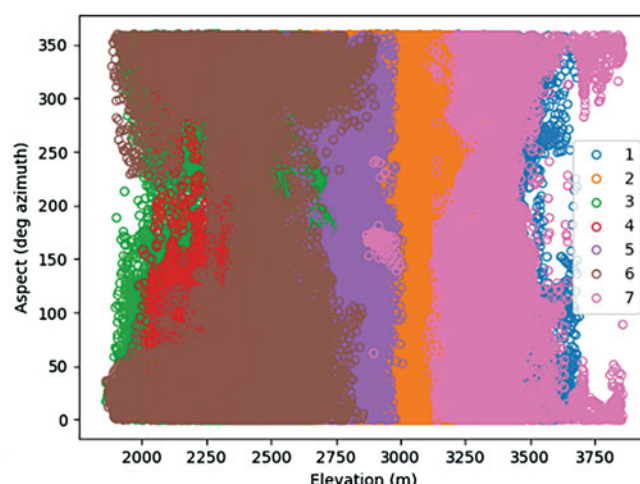
For data fusion needed for classification of multitemporal multisource remote sensing images, the temporal correlation between images is used for computing prior joint probabilities of classes, further estimated using a modified variant of the EM algorithm (Bruzzone et al. 1999). The EM algorithm, thus, has been used to alleviate the need for a human expert to compute these prior joint probabilities. The guarantee of convergence of the EM algorithm to a local maximum of the

ML function enables getting accurate estimates of the probabilities. In this work, the probabilities are used in a classifier, specifically, the multilayer perceptron neural network.

In a similar way, the EM algorithm can be used for change detection applications, where the pixel-wise discrimination of change requires the statistical distribution of “changed” and “unchanged” pixels in a difference image between two time stamps (Bruzzone and Prieto 2000). The use of the EM algorithm for this application has been implemented under the assumption that the conditional density functions of the classes in the difference image can be approximately modeled by GMMs. In this work, the semantic segmentation of changed pixels in the difference images has been further used for thresholding and change detection estimation error computation. Extending this work, the change detection modeled using the EM algorithm in the difference image from multitemporal remote sensing images is then subjected to a level-set method for unsupervised classification (Hao et al. 2013).



(i) Iris Dataset
(3 classes, 150 instances, 4 attributes)



(ii) Forest Covertype dataset
(7 classes, 581012 instances, 54 attributes)

Expectation-Maximization Algorithm, Fig. 2 Comparison of well- and ill-separated classes demonstrated in the two-dimensional projection of the (i) Iris and (ii) Forest cover-type datasets, respectively. The conventional EM algorithm is well- and ill-suited for clustering in these cases, respectively. The visualizations were generated using

matplotlib in Python using two of the numerical attributes that have been found statistically significant in clustering using exploratory data analysis. (Image source: author's own; dataset source: <https://archive.ics.uci.edu/ml/datasets/iris> and <https://archive.ics.uci.edu/ml/datasets/Covertype>)

Summary

The EM algorithm is an effective method for estimating statistical distributions that enable data mining processes, such as classification and clustering. This has widespread applications in geoscience. Owing to the stability and convergence of its iterative implementation, the EM algorithm continues to be widely used in the geoscientific data processing.

Cross-References

- [K-means Clustering](#)
- [Maximum Likelihood](#)
- [Normal Distribution](#)

References

- Bruzzone L, Prieto DF (2000) Automatic analysis of the difference image for unsupervised change detection. *IEEE Trans Geosci Remote Sens* 38(3):1171–1182
- Bruzzone L, Prieto DF, Serpico SB (1999) A neural-statistical approach to multitemporal and multisource remote-sensing image classification. *IEEE Trans Geosci Remote Sens* 37(3):1350–1359
- Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. *J R Stat Soc Series B Stat Methodol* 39(1):1–22
- Do CB, Batzoglou S (2008) What is the expectation maximization algorithm? *Nat Biotechnol* 26(8):897–899
- Hao M, Shi W, Zhang H, Li C (2013) Unsupervised change detection with expectation-maximization-based level set. *IEEE Geosci Remote Sens Lett* 11(1):210–214

- Little RJ, Schluchter MD (1985) Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* 72(3):497–512
- Moon TK (1996) The expectation-maximization algorithm. *IEEE Signal Process Mag* 13(6):47–60
- Plasse JH (2013) The EM algorithm in multivariate Gaussian mixture models using Anderson acceleration. PhD thesis, Worcester Polytechnic Institute

Exploration Geochemistry

David R. Cohen
School of Biological, Earth and Environmental Sciences, The University of New South Wales, Sydney, NSW, Australia

Synonyms

Geochemical prospecting

Definition

The application of geochemical techniques and data analysis methods to the discovery of new mineral deposits and energy resources.

Introduction

Exploration geochemistry involves the assessment of chemical and mineralogical data obtained from a variety of sampling media to locate new mineral deposits and energy resources. In the case of mineral deposits this includes characterizing and modeling geochemical patterns and processes within the primary environment (alteration of the lithological hosts to mineral deposits and the mineralization itself) and the secondary environment (the weathered zone into which elements may disperse or in which secondary mineral deposits may form). A wide range of statistical and spatial analysis techniques may be applied to geochemical data to help identify, isolate, and map geochemical patterns related to the influence of mineralization.

The process of mineralization and associated alteration or the lateral geochemical dispersion of elements in either environment generates atypical geochemistry and mineralogy. The design of geochemical exploration program must therefore encompass the selection, measurement, mapping, and modeling of geochemical features that most effectively separate such atypical or anomalous geochemical patterns from those characteristic of unmineralized (background) zones. Ideally, the analysis or modeling of the geochemical data does not just create a binary categorization of samples or areas into background or anomalous clusters (with associated probabilities of membership), but also provides indicators of the proximity to mineralization via geochemical gradients from which vectors toward mineralization can be established.

Most geochemical datasets currently used in mineral exploration are multivariate. There is increasing use of statistical methods that address multidimensional space to detect and isolate patterns that reflect various geochemical processes (including mineralization) that produce concentration changes in groups of elements. These generally prove more effective than univariate methods in detecting subtle mineralization signatures and providing more reliable probabilistic predictive maps (Grunsky and Caritat 2020).

Historical Developments

Beyond simple visual assessment of spatial geochemical patterns, early statistical approaches focused on simply detecting individual element enrichment or depletion. Univariate anomalies were defined as observations with a low probability of being part of the expected background population – including the assumption that such populations were parametric. Threshold values were set and different populations separated and characterized using graphical techniques such as probability or quantile-quantile plots.

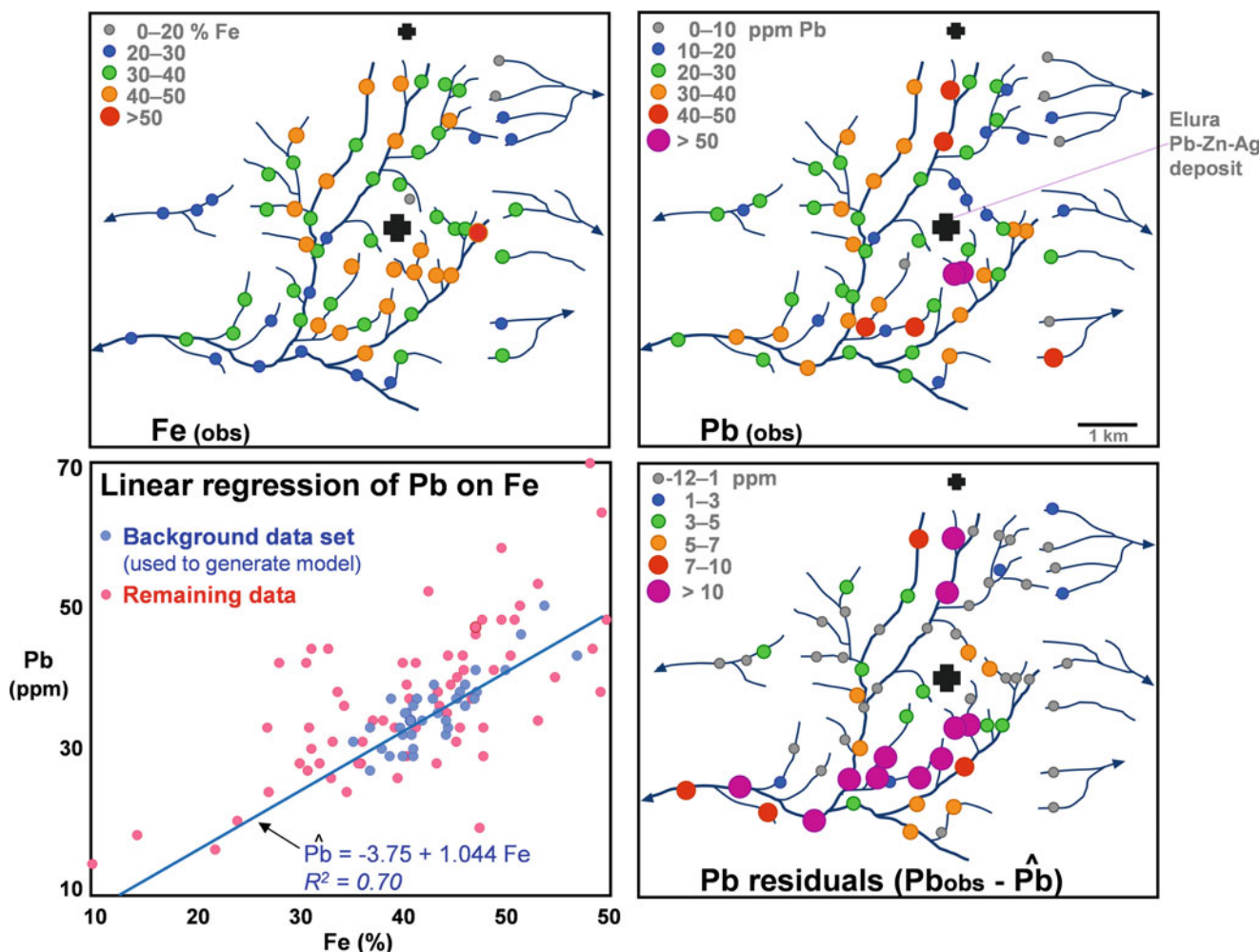
Advances in analytical methods that provide lower detection limits (resulting in a reduction in the proportion of “left-censored” data) provided greater understanding on the factors controlling variability in background patterns, especially for the precious and platinum group elements. “Anomalism” would need to be defined against a highly variable background (Reimann and Garrett 2005) reflecting variation in parent lithology, regolith-landform formation processes, controls on element mobility, and other factors that can exert significant influence on trace and major elements in the primary or secondary environment and spatial discontinuities at various scales.

From the early 1980s onward, the general availability of hardware and software capable of undertaking complex multivariate modeling, simulations, and visualization of data, along with developments of new parametric and nonparametric statistical methods, invited a revisitation to the fundamental question for exploration geochemists – “what is a geochemical anomaly?”. This has resulted in an evolution in the analysis of geochemical data from finding “satellites” (outlying values distant from the main cluster of observations) to detecting more subtle (multivariate) patterns related to the geochemical effects of mineralization, and associated clustering of samples or associations between variables within the main body of observations.

Univariate and Bivariate Methods

There are many approaches to identification of anomalous observations and the separation and characterization of different overlapping populations. This has included the setting of arbitrary boundaries to assumed probability distributions, of which the normal distribution (or distribution transformed to normality) has dominated and linked to tests of significance, outside of which an observation is assigned a high chance of being anomalous. Various rank-order criteria and boundaries have also been used, such the inner or outer fences of Tukey boxplots, as they are less influenced by extreme values in datasets.

There are a number of controls on the physical (mechanical) and chemical (hydromorphic) dispersion of elements and minerals from deposits in the secondary environment. These vary substantially between elements and environments, but can be quantified using appropriate geochemical models. Examples include the effects of stream energy on the relative abundance of heavy minerals (e.g., gold grains or cassiterite) and the incorporation of various chalcophile or siderophile metals within secondary Fe(III) oxides in soils and sediments. Geochemical patterns related to the influence of mineralization must be separated from such



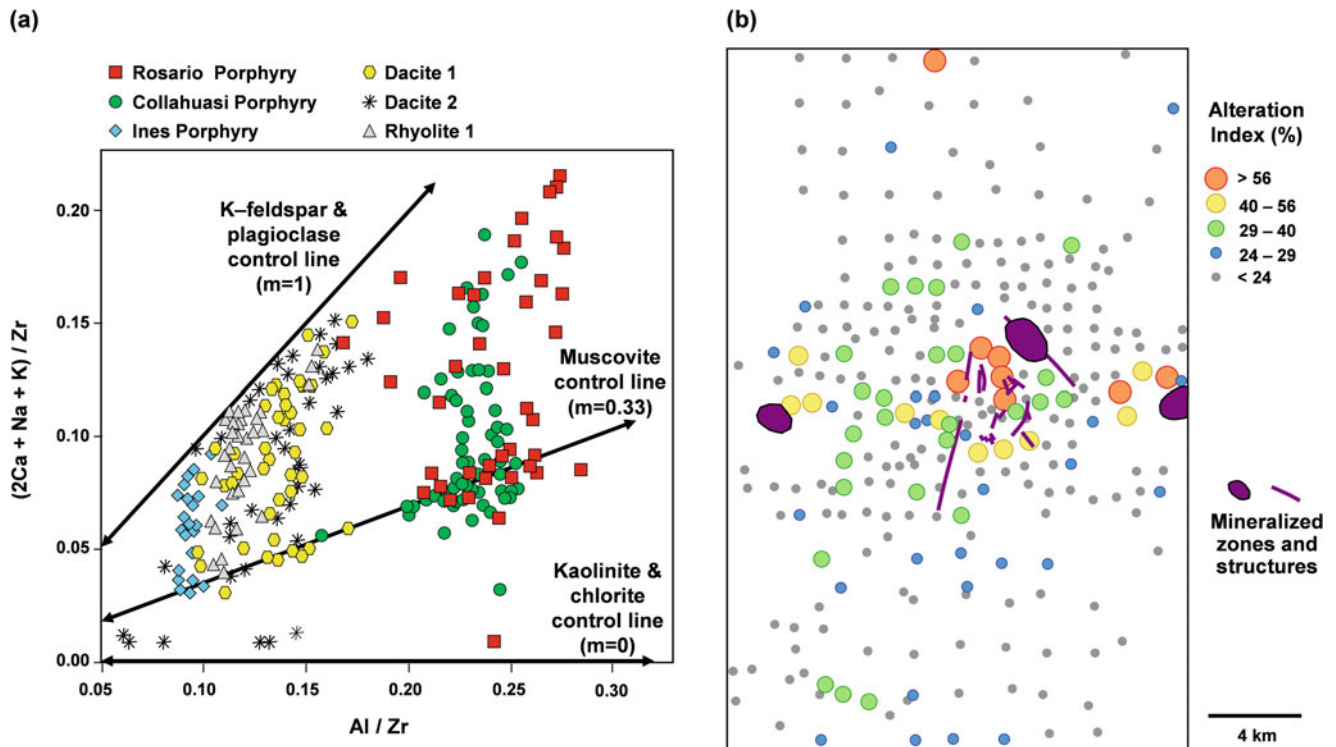
Exploration Geochemistry, Fig. 1 Enhanced anomalous dispersion patterns of Pb in lag by regressing out the effects of variation in Fe. (Data from Dunlop et al. 1983)

local environmental factors. The use of various forms of regression analysis has been common. An example is the effect “regressing out” the influence of variation in Fe in the concentration of Pb incorporated into clasts collected from drainages surrounding the Elura Ag-Pb-Zn sulfide deposit. The resulting Pb residuals display a more distinct and coherent set of anomalies downstream of the deposit and subdued response in other sub-catchments (Fig. 1).

Geochemical data commonly display fractal or multifractal behavior in response to changes in metal sources or variations in the geochemical landscape at various scales. Fractal analysis has progressively overtaken other methods to split populations and identify anomalous populations within geochemical data.

Multivariate Methods

A wide range of multivariate methods have been applied to exploration geochemical data with the object of pattern recognition, clustering of variables and samples (R-mode and Q-mode) and anomaly detection through the isolation and enhancement of patterns associated with the effects of mineralization and identifying samples exhibiting such patterns. Most mineral deposits styles involve enrichment (or depletion) of suites of elements resulting in distortion or clustering of element relationships in the resultant multidimensional space, or separation of populations due to the effects of mineralization and the resulting highly atypical geochemistry of most sampling media. This yields many advantages of multivariate over univariate statistical and data visualization methods, not the least being a greater capacity to detect subtle influences of mineralization on regolith in areas of deep weathering or where surface materials have been transported.



Exploration Geochemistry, Fig. 2 (a) PER for samples from various intermediate to felsic intrusive units in the Collahuasi District of northern Chile ratioed against a conserved element (Zr) with selected mineral control lines and (b) Plot of an alteration index reflecting mass transfer

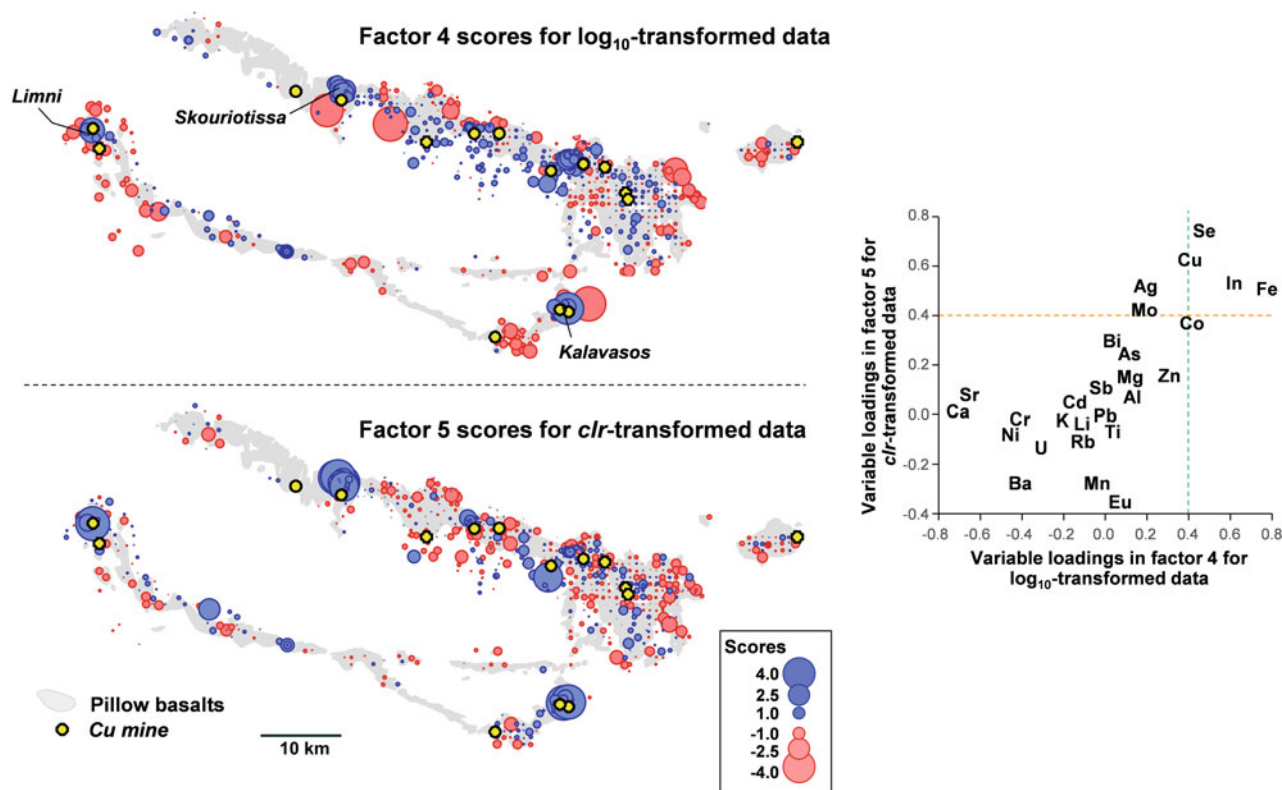
Geochemical data can also be projected into mineralogical space by conversion of mass abundances to molar abundances (with or without normalization against conserved elements) and the selection of suitable axes that reflect mineral compositions. The use of Pearce element ratio (PER) diagrams is one example that has been used to measure the extent and nature of rock alteration, and the relationship to the abundance of elements of economic interest (Stanley 2018). In a case study on the Collahuasi District of northern Chile (Urqueta et al. 2009), the degree of hydrolytic alteration in intrusive units (loss of Ca and Na and/or gain of K) was quantitatively related to the distance of observations from the feldspar control line along which samples with no alteration would plot (Fig. 2). In this and other studied the extent and nature of such alteration is more effectively and efficiently represented in the geochemical data than petrological examination in the field or laboratory and assists vectoring to mineralization based on varying alteration intensity.

Most geochemical datasets used in mineral exploration involve total or selectively extracted elemental abundances. Statistical methods applied to such compositional data are subject to the potential biasing effects of data closure. A number of corrections can be applied to reduce closure effects, for which the use of component ratios to open the data into real number space are commonly applied, such as

under hydrolytic alteration for samples based on a distance metric of observations from the feldspar control line (unaltered samples). (After Urqueta et al. 2009)

the additive, centered, and isometric log-ratios log-ratio (*alr*, *clr*, and *ilr*) (Buccianti and Grunsky 2014). This pre-processing of data can also reduce the effects of skewed distributions or outliers on classical multivariate methods such as principal components analysis, factor analysis, and other methods involving decomposing of variance-covariance or correlation matrices.

As an example, factor analysis was applied to multi-element data in soils collected from a grid across the basaltic pillow and basal lavas that host a number of VHMS Cu deposits in Cyprus. A six-factor model (with varimax rotation) was applied to both $\log_{10}$ -transformed data to reduce the effects of outliers and non-normal distribution on the model and *clr*-transformed data that also reduce the effects of data closure. Both approaches delivered factors displaying moderate to strong factor loadings for some chalcophile elements, although the *clr*-transformed data displayed more of these with a loading >0.4). However, the *clr*-transformed data displayed stronger separation between areas with known mineralization (high sample factor scores) in terms of both geochemical contrast (numerical differences between mineralized and background areas) and spatial coherence of the patterns (Fig. 3).



Exploration Geochemistry, Fig. 3 Comparison of mineralization-related factor score patterns and known Cyprus-style VMS Cu deposits in factor analysis of $\log_{10}$ -transformed and centered-log ratio

transformed multivariate geochemical data in soils overlying basaltic pillow lavas of the Troodos Ophiolite. (Data from Cohen et al. 2012)

Spatial Methods

Two features common to most exploration geochemical datasets are that (i) samples are spatially related in two or three dimensions and (ii) sampling is very limited or even skeletal compared with the complexity and scale of the entities that such data is being used to characterize (Cheng 1999). Geostatistical techniques have long been used to quantify the relationship between distance between samples and the associated variance of geochemical attributes, with applications in ore reserve estimation and in conversion of point data to spatially continuous variables (e.g., rasterizing). It remains an essential step in the systematic analysis of geochemical mapping data to spatially correlate observed or extracted patterns with other relevant spatial datasets including regional and local geology, regolith-landform maps, and other relevant spatial information. This can be accomplished within various GIS platforms (Carranza 2008).

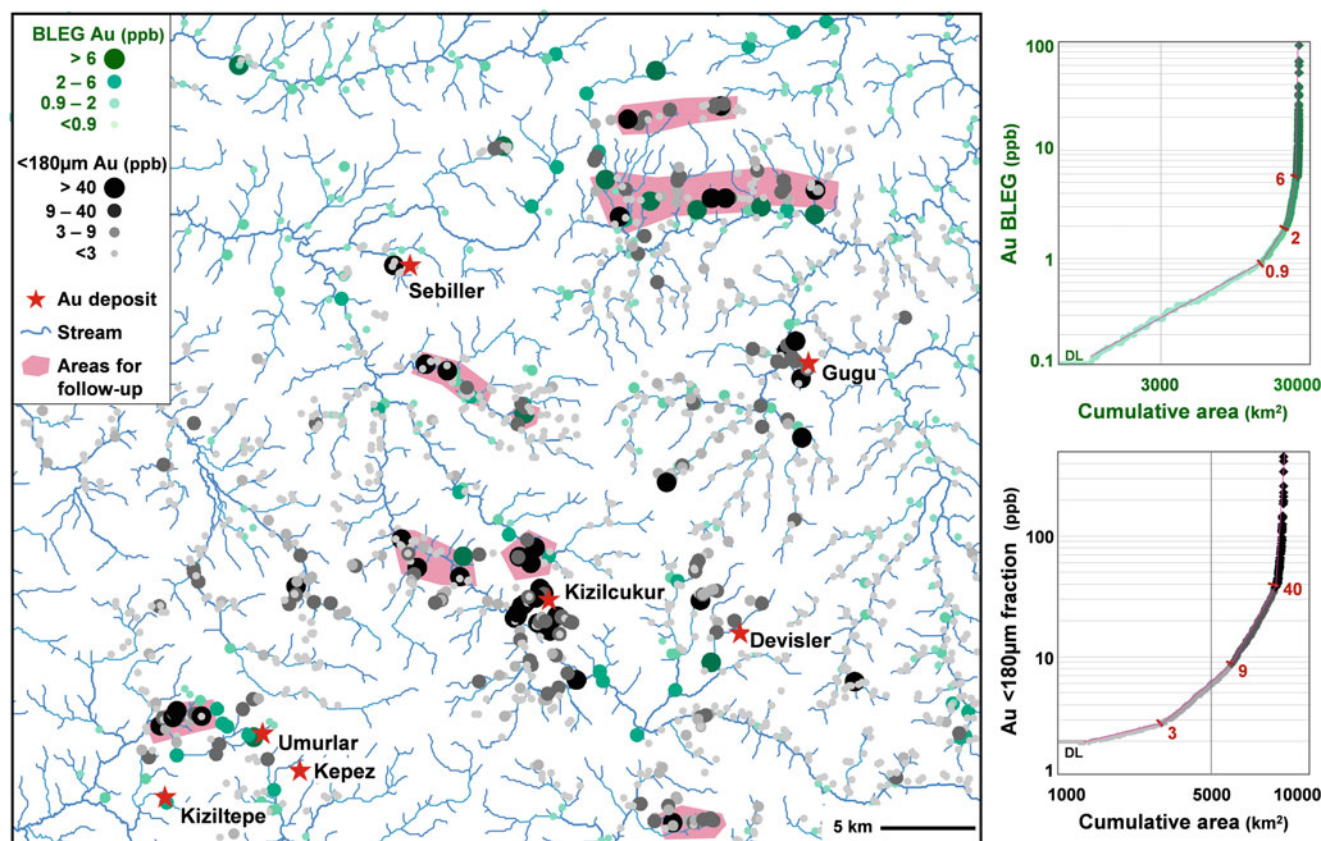
Univariate or multivariate geochemical anomaly delineation techniques that incorporate spatial variability and correlation among samples, along with frequency distributions, are defined as structural approaches (Reimann et al. 2011). They assist the setting of criteria for defining anomalous geochemical signatures that better reflect variations in local geology,

regolith-landform setting, and other features that can exert substantial control on the geochemistry of media such as soils and stream sediments. Structural methods include a variety of fractal techniques (e.g., concentration-area), U-statistics, and singularity analysis among others (Geranian et al. 2013; Liu et al. 2017; Ghezelbash et al. 2019).

An example is drawn from bulk leach-extractable and $< 180 \mu\text{m}$ fraction Au contents in stream sediments from a large region of western Turkey (Yilmaz et al. 2017). Using a concentration-area fractal model $A(c \geq c_i) \propto c_i^{-D}$, where $A(c \geq c_i)$ represents the area (A) with concentrations $\geq$ to contour value c_i and D is the fractal dimension, both geochemical variables displaying fractal behavior. Known mineralized catchments are associated with the upper two fractal populations and a number of new areas with mineral potential were identified (Fig. 4).

Directions

On-going developments in sampling and analytical methods, including progressive lowering of detection limits, access to methods for routine analysis of a wider range of isotopes or chemical species, and understanding geochemical processes



Exploration Geochemistry, Fig. 4 Comparison of concentration-area fractal populations in BLEG and < 180 μm fraction Au in stream sediments from central-western Turkey. (After Yilmaz et al. 2017)

controlling dispersion in the primary and secondary environments are providing new dimensions to geochemical data and capacity to predict the form that geochemical anomalies might take. The evolution of statistical methods is towards techniques that are able to separate complex multivariate patterns within geochemical data and isolate the signal(s) that relate to the effects of mineralization based on the inherent characteristics of the data rather than preconceived models or assumptions. This will be facilitated by advances in the broad field of supervised and unsupervised machine learning approaches capable of modeling complex multivariate geochemical distributions and identifying latent associations between variables and samples (Zuo 2017) and other feature extraction and selection techniques (Zekri et al. 2019). Specific techniques include artificial neural networks, support vector machines, and random forests.

Deep learning algorithms are proving more effective in dealing with complex datasets that could enhance the geochemical anomaly detection and help in identifying hidden geochemical patterns in data. Deep machine learning architectures include stacked auto-encoding, convolutional and recurrent neural networks, and generative adversarial networks (Zuo et al. 2019). Exploration geochemistry utilized

by various machine learning methods have growing roles in mineral prospectivity mapping, especially areas where mineralization displays very weak surface expression.

Summary

The principal objective of exploration geochemistry is to define areas, at regional to mineral deposit scales, that have high probability of containing mineral deposits of the style and size required. This is largely dependent on the selection and application of numerical criteria to the distribution characteristics of individual variables, patterns of correlation between variables, spatial patterns within geochemical (and associated) datasets, capable of isolating patterns and signals associated with the effects of mineralization (anomaly identification) from other factors (background characterization).

Cross-References

- [Compositional Data](#)
- [Deep Learning in Geoscience](#)

- Machine Learning
- Neural Networks
- Principal Component Analysis
- Regression
- Singularity Analysis

Bibliography

- Buccianti A, Grunsky EC (2014) Compositional data analysis in geochemistry: are we sure to see what really occurs during natural processes? *J Geochem Explor* 141:1–5
- Carranza EJM (2008) Geochemical anomaly and mineral prospectivity mapping in GIS. In: *Handbook of exploration and environmental geochemistry*, vol 11. Elsevier, Amsterdam
- Cheng Q (1999) Spatial and scaling modelling for geochemical anomaly separation. *J Geochem Explor* 65:175–194
- Cohen DR, Rutherford NF, Morisseau E, Christoforou E, Zissimos AM (2012) Anthropogenic versus lithological influences on soil geochemical patterns in Cyprus. *Geochem: Explorat Environ Anal* 12:349–360
- Dunlop AC, Atherden PR, Govett GJS (1983) Lead distribution in drainage channels about the Elura zinc-lead-silver deposit, Cobar, New South Wales, Australia. *J Geochem Explor* 18:195–204
- Geranian H, Mokhtari AR, Cohen DR (2013) A comparison of fractal methods and probability plots in identifying and mapping soil metal contamination near an active mining area, Iran. *Sci Total Environ* 463:845–854
- Ghezelbash R, Maghsoudi A, Carranza EJM (2019) Mapping of single- and multi-element geochemical indicators based on catchment basin analysis: application of fractal method and unsupervised clustering models. *J Geochem Explor* 199:90–104
- Grunsky EC, de Caritat P (2020) State-of-the-art analysis of geochemical data for mineral exploration. *Geochem: Explorat Environ Anal* 20:217–232
- Liu Y, Zhou K, Cheng Q (2017) A new method for geochemical anomaly separation based on the distribution patterns of singularity indices. *Comput Geosci* 105:139–147
- Reimann C, Garrett RG (2005) Geochemical background – concept and reality. *Sci Total Environ* 350:12–27
- Reimann C, Filzmoser P, Garrett R, Dutter R (2011) *Statistical data analysis explained: applied environmental statistics with R*. Wiley, Hoboken
- Stanley CR (2018) Molar element ratio analysis of lithogeochemical data: a toolbox for use in mineral exploration and mining. *Geochem: Explorat Environ Anal* 20:233–256
- Urqueta E, Kyser TK, Clark AH, Stanley CR, Oates CJ (2009) Lithogeochemistry of the Collahuasi porphyry Cu-Mo and epithermal Cu-Ag (-Au) cluster, northern Chile: Pearce element ratio vectors to ore. *Geochem: Explorat Environ Anal* 9:9–17
- Yilmaz H, Cohen DR, Sonmez FN (2017) Comparison between the effectiveness of regional BLEG and <80# stream sediment geochemistry in detection of precious and base metal mineral deposits in Western Turkey. *J Geochem Explor* 181:69–80
- Zekri H, Cohen DR, Mokhtari AR, Esmaeili A (2019) Geochemical prospectivity mapping through a feature extraction-selection classification scheme. *Nat Resour Res* 28:849–865
- Zuo R (2017) Machine learning of mineralization-related geochemical anomalies: a review of potential methods. *Nat Resour Res* 26:457–464
- Zuo R, Xiong Y, Wang J, Carranza EJM (2019) Deep learning and its application in geochemical mapping. *Earth Sci Rev* 192:1–14

Exploratory Data Analysis

Emmanuel John M. Carranza

Department of Geology, Faculty of Natural and Agricultural Sciences, University of the Free State, Bloemfontein, Republic of South Africa

Definition

Exploratory data analysis (hereafter denoted as EDA) is not a technique but it is a paradigm or strategy for robust analysis of data (Tukey 1977). It comprises descriptive, statistical, and, largely, graphical investigations meant for (a) obtaining the greatest understanding of a set of data, (b) revealing structure of data, (c) determining important data variables, (d) determining outlying and anomalous values in the data, (e) proposing and examining hypotheses, (f) creating judicious models, and (g) identifying optimum way(s) to manage and analyze data. It is a vigorous initial phase of data analysis prior to their treatment, for the purpose the data were to be used, by applying other methods of statistical analysis.

Introduction

The progression of EDA is from problem to data to analysis to model to conclusion. This contrasts with the progression of data analysis in classical statistics, which is from problem to data to model to analysis to conclusion. The progression of EDA also differs with the progression of data analysis in probabilistic framework, which is from problem to data to model to analysis of a priori data distribution to conclusion. Therefore, whereas data analyses in classical statistical and probabilistic frameworks are confirmatory in nature because they require models of prior assumptions of data distribution, EDA is, by virtue of its name, exploratory in nature.

EDA has four main aspects (cf. Tukey 1977; Howarth 1984): (a) graphical analysis to unravel data structure; (b) data transformation to simplify data behavior; (c) application of resilient techniques that are not overly affected by a few outlying values; and (d) focus on residual values, i.e., some portion of the data not accounted for by a certain model. The objective of EDA is, therefore, to identify “theoretically explainable” but often not obvious patterns in data (Good 1983) by way of using resilient and robust tools for descriptive and graphical statistical analyses that are different from those used for classical statistical analysis. A statistic is resilient and robust if it is only faintly influenced (a) by either a few gross errors or several small errors (resilience) and (b) by outlying values (robustness). In EDA, the tools employed for descriptive statistical and graphical

analyses are not according to a model of data frequency distribution (e.g., normality) but based on data itself. However, the tools for EDA offer resilient characterizations of statistics and outliers in univariate data as well as anomalous relationships in multivariate data.

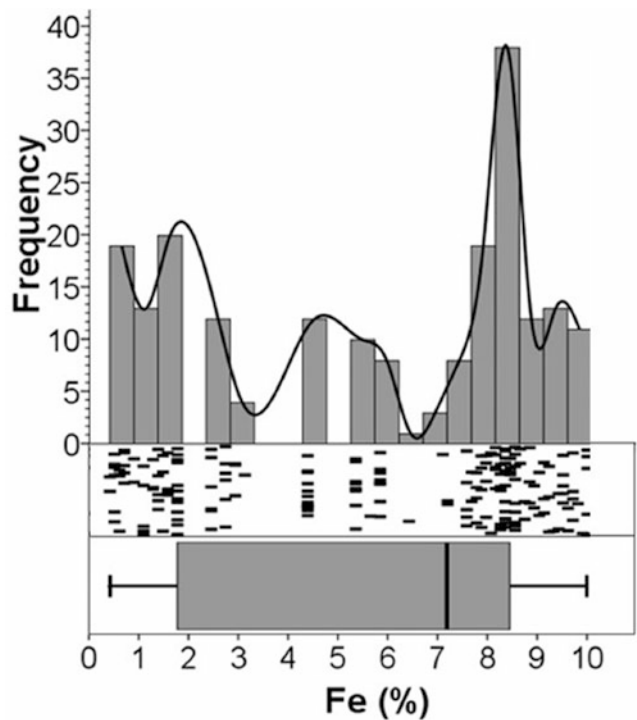
Graphical Analysis

EDA stresses on the use of statistical graphics as interfaces between computation and human cognition; these graphics enable users to realize the structure and behavior of data. The most commonly used graphics for EDA, which can be easily stacked on top of other, are (Tukey 1977) density trace, jittered one-dimensional scatterplot, and boxplot (Fig. 1). These are used frequently and jointly with a histogram, as the pictorial intuition one achieves regarding data structure and behavior from just a histogram is prejudiced by user-specified number of classes for its creation. The joint application of such graphics for EDA with a histogram offers superior perception of univariate data structure and behavior compared to one offered by a histogram alone. These three graphics for EDA, unlike a histogram, can easily reveal any “oddities” in a set of data.

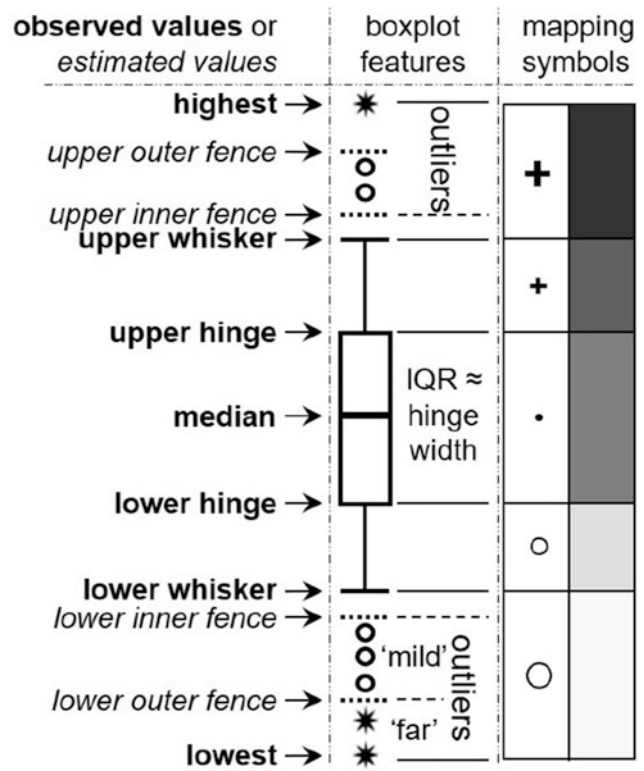
A density trace is akin to a histogram although it allows for better realistic perception of the empirical frequency

distribution of data, as its shape is resilient to variation in number of classes. The suitable number of histogram classes can be constrained by creating a jittered one-dimensional scatterplot, which charts data at random locations in a narrow band (typically [0,1] range) perpendicular to data axis. As creation of jittered one-dimensional scatterplot does not rely on classes of data, it delivers supplementary statistics about data (i.e., local densities, structure, behavior, gaps, outliers) that ought to be depicted by a density trace and a histogram. A boxplot portrays descriptive statistics of the empirical frequency distribution of data, and so it is perhaps the most valuable graphics for EDA.

To create a boxplot, data are first arranged from the lowest value to the highest value or the other way around (Fig. 2). The median is defined by counting midway from the lowest value to the highest value, or the other way around, thereby splitting a set of data into two nearly equal subsets. Values of the lower and upper hinges are identified, correspondingly, by counting midway from the lowest value to the median and counting midway from the highest value to the median. The lower hinge, median, and upper hinge split a data set into four nearly equal subsets called quartiles. For a set of data, the first quartile includes values from the lowest to the lower hinge,



Exploratory Data Analysis, Fig. 1 EDA graphics: histogram with density trace (top), jittered one-dimensional scatterplot (middle), and boxplot (bottom)



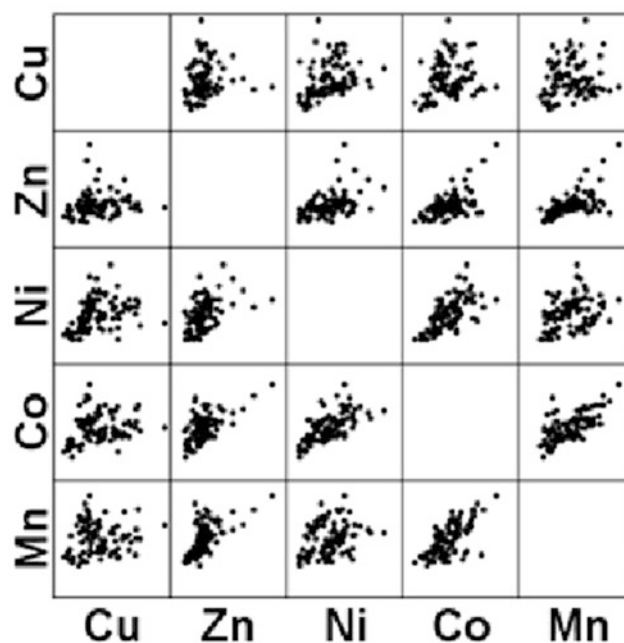
Exploratory Data Analysis, Fig. 2 Features of a boxplot. Features for values estimated based on the hinge width or the inter-quartile range are labeled in *italics*. Features for observed or measured values at which a univariate data set may be divided into five robust classes are labeled in **bold**. EDA mapping symbols or gray scale colors can be assigned to each class for visualization of data on a graph or a map

the second quartile includes values from the lower hinge to the median, the third quartile includes values from the median to the upper hinge, and the fourth quartile includes values from the upper hinge to the highest. The box, which is defined by the lower and upper hinges, is split often by a line representing the median. The absolute difference of the two hinge values defines the hinge width or inter-quartile range (IQR). A lower inner fence is marked at $1.5 \times$ hinge width and a lower outer fence is marked at $3 \times$ hinge width away from the lower hinge to the lowest value. An upper inner fence is marked at $1.5 \times$ hinge width and an upper outer fence is marked $3 \times$ hinge width away from the upper hinge to the highest value. A lower whisker is marked from the lower hinge and an upper whisker is marked from the upper hinge, toward extreme values within the inner fences. Any value past either the lower or the upper inner fence is regarded as an outlier. Outliers are “mild” if they fall within the inner and outer fences and “far” if they fall beyond either the lower or the upper outer fence. Mild and far outliers are symbolized differently (Fig. 2).

Therefore, a boxplot expresses the five-number summary statistics (lowest value, lower hinge, median, upper hinge, and highest value), and it illustrates the most vital features of a data set, namely (Tukey 1977), (a) central tendency; (b) spread; (c) skewness; (d) lengths of tails; and (e) outliers. As the box embodies roughly two quartiles of a set of data, it implies that $\leq 25\%$ of data may be outlying values; however, outliers radically influence neither the median nor the hinges. Accordingly, the inner fences are not critically influenced by outliers because they are outlined by the hinge width. Thus, a boxplot is a resilient and robust tool for descriptive statistical and graphical analyses of data.

To explore relationships between two variables, say, X and Y , in a set of data, the commonly used graphic for EDA is a scatterplot. In this plot, values of Y , a variable thought to be a response of another variable, are plotted on the y -axis, whereas values of X , a variable suspected to be related to the response variable, are plotted on the x -axis. The relationship of Y to X is revealed by a nonrandom pattern in the plot. Therefore, scatterplots can manifest whether Y is positively or negatively related to X , whether Y is linearly or nonlinearly related to X , and whether or not there are outliers (i.e., pairs of X and Y values that do not conform to the general pattern in a scatterplot). To aid understand higher-level pattern in a set of data with more than two variables, all pairs of scatterplots can be displayed together in scatterplot matrix (Fig. 3); this helps to visualize a correlation coefficient matrix.

For visualization of multivariate data, a typical EDA graphic that can be used is star plot (Chambers et al. 1983). It consists of a series of equiangular spokes, denoted as radii, and every spoke represents a variable. A spoke's length is proportional to the value of a variable for an observation relative to the maximum value of the variable across all

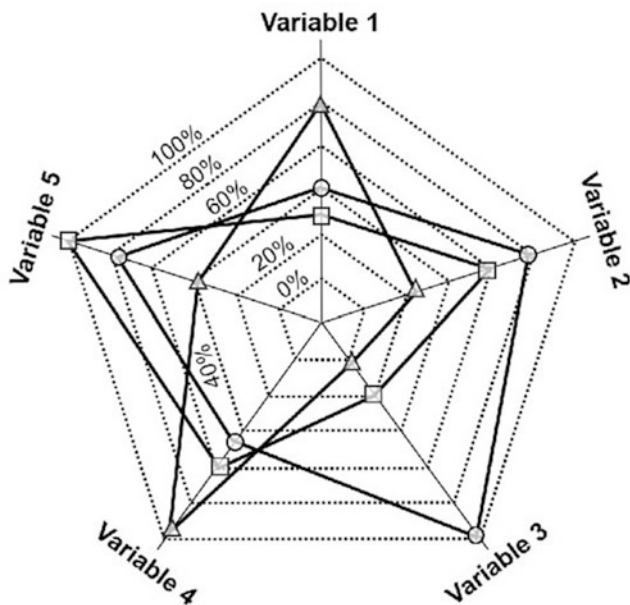


Exploratory Data Analysis, Fig. 3 Example of a scatterplot matrix

observations. A line connects data values per spoke, giving the plot a star-like form and, thus, its name (Fig. 4). Star plots are usually created in a multi-plot format with several stars per page and every star portrays a single observation or sample. A star plot helps to visualize (a) which variables are dominant per observation, (b) presence of clusters of observations, and (c) presence of outliers. However, star plots are useful mainly for multivariate data sets with at most a few hundred points; for larger multivariate data sets, visual analysis of star plots can be cumbersome.

Robust Classification of Univariate Data

A typical research may involve a classification problem. According to a boxplot, a set of univariate data may be split into five robust classes: (1) lowest value, lower whisker; (2) lower whisker, lower hinge; (3) lower hinge, upper hinge; (4) upper hinge, upper whisker; and (5) upper whisker, highest value. For classification of geochemical data for mineral exploration, the first class may be labeled as extremely low background, the second class (comprising $\leq 25\%$ of data set) as low background, the third class (comprising $\leq 50\%$ of data set) as background, the fourth class (comprising $\leq 25\%$ of data set) as high background, and the fifth class as anomaly. Each of these classes can be given suitable symbols (Fig. 2), which can be used to symbolize and visualize the data on a map. Some researchers use the upper inner fence as threshold for distinguishing background values from anomalous values (e.g., Reimann et al. 2005), whereas other research use the upper outer fence as threshold for the same purpose (e.g.,



Exploratory Data Analysis, Fig. 4 Example of multiple star plots for three samples (symbolized by box, circle and triangle) with data for five variables

Yusta et al. 1998). However, because the upper inner fence (i.e., $1.5 \times$ hinge width) or the upper outer fence (i.e., $3 \times$ hinge width) is an estimated value that may not be among the measured values in set of univariate data, outliers past the upper whisker may be regarded as anomalous value. Besides a threshold defined from a boxplot (e.g., upper inner fence or upper whisker), a threshold for distinguishing background from anomalous values can be calculated using EDA statistics as median plus a multiple of median absolute deviation (or MAD). The MAD, which is a measure of data spread, is calculated as median of absolute differences of individual values from their median (Tukey 1977), thus:

$$\text{MAD} = \text{median}[|X_i - \text{median}(X_i)|]$$

where the term $|X_i - \text{median}(X_i)|$ represents absolute deviations of individual values (X_i) from their median (i.e., median (X_i)) in a data set. The MAD is to EDA as the standard deviation is to classical statistics, and so an EDA-calculated threshold of median plus $2 \times$ MAD is equivalent to a threshold calculated as mean plus $2 \times$ standard deviation, which was used traditionally to map geochemical anomalies for mineral exploration (Hawkes and Webb 1962). Certainly, other boxplot- or EDA-derived classes and different class label are possible depending on purpose/scope of research, which dictates what “anomalous values” represent (e.g., anomalous values in exploration geochemistry and in environmental geochemistry are not the same).

In case of univariate data that exhibit presence of several populations, the analysis of only the whole data set is insufficient or even inappropriate for identification of anomalous

values that may be linked with each or some population in the data. It is, thus, crucial to split such data into subsets that represent the different populations present. Univariate data empirical frequency distribution portrayed in a boxplot or in a Q-Q (quantile-quantile) or cumulative probability plot is valuable in graphical analysis to identify gaps or inflection points for splitting data into subsets that represent data populations. Instead, if populations in univariate data are regarded linked to specific geological variables (e.g., rock units), which may also have been logged during the field sampling/observations, then such data may be split into populations according to such variables. Each data population is then subjected to EDA (e.g., boxplot per data subset).

Data Standardization and Transformation

For proper comparison of classes in data populations (e.g., levels of geochemical anomalies in different rock units), a suitable standardization is required. A typical standardization formula from classical statistics is:

$$Z_{ij} = \frac{X_{ij} - \bar{X}_{ij}}{\text{SDEV}_{ij}}$$

where Z_{ij} and X_{ij} represent standardized and original values, respectively, of variable i in population j , and $\bar{X}_{ij}$ and SDEV_{ij} represent mean and standard deviation, respectively, of X_{ij} . As both mean and standard deviation are not resilient to outlying values, data should not be standardized using the above formula. Instead, the following standardization formula from EDA statistics can be used (Yusta et al. 1998).

$$Z_{ij} = \frac{X_{ij} - \text{median}_{ij}}{\text{MAD}_{ij}}.$$

The above standardization formula allows leveling of every population j to one another in a univariate data set, and, thus, classes of different populations in the data set become comparable to one another. The IQR can be used instead of MAD in the above formula. However, using the above formula results in standardized values that can be negative, which cannot be log-transformed if needed. To make all the standardized values positive, say values greater or equal to 1, the following formula may be used (Camizuli and Carranza 2018):

$$z_{ij} = 1 + (Z_{ij} - \min(Z_{ij}))$$

where $\min(Z_{ij})$ is the minimum standardized value. Such EDA-based standardization can be important in the compilation and analysis of, for example, geochemical data pertaining to different map sheets or to different periods of data collection.

Exploratory Analysis of Multivariate Relationships

Any nonrandom relationships among at least two variables revealed from a scatterplot matrix or from star plots can be explored further for detailed understanding of their behaviors by using methods of multivariate data analysis, namely, cluster analysis (Templ et al. 2008) and principal component analysis (Camizuli and Carranza 2018). Cluster analysis can be employed as a tool for EDA to better recognize the multivariate structure and behavior of a set of data. It yields clusters of the data, each cluster consisting of variables that are strongly similar to each other and are strongly dissimilar to the variables in each of the other clusters. Most methods of clustering employ a measure of similarity, which is typically based on distances between values in the data space. Principal component analysis, on the other hand, employs correlations among variables to yield components of data that are uncorrelated. It is useful in determining which variables are redundant in the measure of providing information, and it is useful in recognizing multivariate outliers; it is typically employed as an EDA tool prior to cluster analysis in order to obtain strongly coherent clusters. Details of cluster analysis can be found in the entry on ► [“Clustering and Classification”](#) and those of principal component analysis in the entry on ► [“Principal Component Analysis.”](#)

For exploratory analysis of compositional data (e.g., geochemical data shown in Fig. 3), which are typically affected by the so-called closure problem (see entry on ► [“Compositional Data”](#)), a suitable transformation needs to be applied (see entry on ► [“Compositional Data”](#)) before visualizing multivariate relationships by using EDA graphical tools such as the scatterplot matrix or star plots and before analyzing multivariate relationships through either cluster analysis or principal component analysis.

Concluding Remarks

Exploratory data analysis is widely used in the broad field of geosciences for summarizing and graphically displaying earth

science data collected for a variety of purposes. Exploratory data analysis and confirmatory data analysis are complementary to each other. Exploratory data analysis typically results in information that provide stimulus for further confirmatory data analysis. It is more instructive to employ, first, both exploratory data analysis and, then, confirmatory data analysis than to apply each approach exclusively in order to preclude erroneous interpretation of data.

Cross-References

- [Cluster Analysis and Classification](#)
- [Compositional Data](#)
- [Cumulative Probability Plot](#)
- [Multivariate Analysis](#)
- [Principal Component Analysis](#)
- [Univariate](#)

Bibliography

- Camizuli E, Carranza EJ (2018) Exploratory data analysis (EDA). The encyclopedia of archaeological sciences. Wiley, pp 664–670
- Chambers J, Cleveland W, Kleiner B, Tukey P (1983) Graphical methods for data analysis. Wadsworth Press, Belmont
- Good IJ (1983) The philosophy of exploratory data analysis. *Philos Sci* 50:283–295
- Hawkes HE, Webb JS (1962) Geochemistry in mineral exploration. Harper, New York
- Howarth RJ (1984) Statistical approach in geochemical prospecting: a survey of recent developments. *J Geochem Explor* 21(1):41–61
- Reimann C, Filzmoser P, Garrett RG (2005) Background and threshold: critical comparison of methods of determination. *Sci Total Environ* 346:1–16
- Templ M, Filzmoser P, Reimann C (2008) Cluster analysis applied to regional geochemical data: problems and possibilities. *Appl Geochem* 23:2198–2213
- Tukey JW (1977) Exploratory data analysis. Addison-Wesley, Reading
- Yusta I, Velasco F, Herrero J-M (1998) Anomaly threshold estimation and data normalization using EDA statistics: application to litho-geochemical exploration in lower cretaceous Zn±Pb carbonate-hosted deposits, northern Spain. *Appl Geochem* 13:421–439

F

FAIR Data Principles

Abdullah Alowairdhi and Xiaogang Ma
Department of Computer Science, University of Idaho,
Moscow, ID, USA

Synonyms

FAIR guiding principles

Definition

FAIR refers to the four fundamental pillars: 1) findability; 2) accessibility; 3) interoperability; and 4) reusability. The aim is to improve data management and stewardship by empowering computers to find, access, interoperate, and reuse data. FAIR data principles, in brief, are: 1) findable, discoverable with metadata, locatable, and identifiable through a standard identification mechanism; 2) accessible, obtainable, and available all the time; even if the data are restricted, the metadata is open; 3) interoperable, both syntactically explainable and semantically understandable, enabling data exchange and allowing data reuse among researchers, institutions, organizations, and countries; and 4) reusable, adequately described, and shared with the permitted licenses, supported by provenance empowering the broadest reuse possible and minimal tedious integration with other data sources.

Introduction

FAIR data principles are a set of 15 guiding principles aimed at publishing data in a format that enables findability, accessibility, interoperability, and reusability. For better data management and stewardship, data must be findable, accessible,

interoperable, and reusable by human and computer agents. Findable requires that each dataset has a specific global identifier and is combined with accurate, searchable (meta)data. Furthermore, to be accessible, those data and metadata should all be addressed using an accessible, standard protocol; further, they should use a structured, widely understood descriptive language and use popular and widely used frameworks of related vocabulary and ontologies to make the data and metadata interoperable. Finally, excess cross-references and a simple, well-defined procedure for obtaining licensing and provenance data should be richly represented by the data owner to address the data reusability (Wilkinson et al. 2016).

Findable F

“Principle F1: (meta)data are assigned a globally unique and persistent identifier”

Principle F1 specifies that digital resources allocated globally unique and permanent identifiers (i.e., data, metadata, software code, and workflow) to be found and resolved by humans and machines. This principle is an essential principle of FAIR since globally unique and permanent identifiers are core for all the 15 FAIR data principles. Globally unique ensures the identifier corresponds to only one resource in the world unequivocally. Persistence means the condition that even though the digital resource no longer remains or moved, this globally unique and persistent identifier is not ever used anywhere in any sense and perseveres to represent the same digital resource. Practically, this also implies using third-party agencies, such as DataCite that provides a digital object identifier (DOI), to devise an identifier that has assured consistency and is autonomous (Juty et al. 2020).

“Principle F2: data are described with rich metadata”

Principle F2 refers to the capability of locating the digital resource of interest by searching or filtering. Rich metadata should describe the digital resource, the contents of that

digital resource referred to with that descriptor. It is challenging to describe the insufficient richness of these metadata, but the more abundant it is, the more accurately it can be found by both humans and computers in a specific search. Although the other FAIR data principles correspond to certain sets of metadata that must be used, principle F2 perceives it is not possible to discover insufficiently described digital resources. Therefore, this F2 principle inspires data owners to predict the different aspects of a search performed by users, which support the discovery of these digital resources. Moreover, data owners must give general and domain-specific descriptions to permit search engines to locate digital resources (Jacobsen et al. 2020).

“Principle F3: (meta)data clearly and explicitly include the identifier of the data it describes”

Principle F3 stipulates the (meta)data of every digital resource clearly and explicitly includes the identifier of the defined digital resource. For example, in an unambiguous way, the dataset description should specifically include the dataset identifier. This inclusion of the identifier is especially important when the metadata and the digital resource are reserved separately but consistently connected, which is considered by FAIR as a good practice. Besides, by necessitating the metadata to include the identifier of its mentioned digital resource, the identifier can then be effectively used as a search reference for the discovery of its data record (Jacobsen et al. 2020).

“Principle F4: (meta)data are registered or indexed in a searchable resource”

For satisfying this principle, searchable catalogs such as data.gov must register or index the digital resource. The searchable catalog includes the platform from which it is likely to detect the (meta)data identifier, using either the properties of that (meta)data or the digital resource identifier itself (Weigel et al. 2020).

Accessible A

“Principle A1: (meta)data are retrievable by their identifier using a standardized communications protocol”

Principle A1 notes that the retrieval of FAIR data should be mediated without advanced instruments or registered methods of communication. This principle focuses on how this is possible to extract data and metadata from their identifiers. The standardized communication protocol like HTTP and FTP are intended to enforce a predictable method of accessing a digital resource, irrespective of whether open access is given to the content of the digital resource. HTTP (s) and FTP are based on TCP and make it considerably simpler to seek and provide digital resources than other

communications protocols, forming the core of the modern Internet (Jacobsen et al. 2020).

“Principle A1.1: the protocol is open, free, and universally implementable”

The worldwide network protocols, such as HTTP, FTP, and SMTP, are a model for a public protocol, gratis, and globally deployable. Because these worldwide network protocols are well defined and accessible, they minimize the expense of accessing digital resources and permit agents to apply their standards compliance. The free usage of these worldwide network protocols guarantees the digital resource is evenly available to those with less financial resources. It is broadly applicable means the technology is open and not controlled by any country or a specific community (Bailo et al. 2020).

“Principle A1.2: the protocol allows for an authentication and authorization procedure, where necessary”

Principle A in FAIR does not imply “open” and/or “free”; it means that one should determine the exact circumstances under which the data are accessible. Even highly secured and confidential data could also be FAIR. Communication protocols define accessibility in such a manner that they will automatically recognize the credentials and then either perform the requests automatically or alert the user for the requirement. It also prospers practice to allow the data owner to set up a user account to authenticate the owner or contributor of each dataset, and if necessary, to set user-special privileges. These criteria will, however, influence your opinion of where you would store your data (Brewster et al. 2020).

“Principle A2: (meta)data are accessible, even when the data are no longer available”

Digital resources tend to decline or vanish over time because there is a price to retain web presence to data resources. When this occurs, linkages become void, and users will spend hours searching for data that may no longer exist. In practice, maintaining metadata is much cheaper and more convenient than retaining data. Thus, the A2 principle specifies that metadata should be preserved by repositories even though data are no longer available. A2 refers to the issues of registration and indexing listed in F4 (Martone 2014).

Interoperable I

“Principle I1: (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation”

Principle I1 means that computers can interpret data despite the need for advanced or functional software, interpreters, or configurations. Usually, interoperability requires that each

computing system at least knows the data integration formats of the other system. To guarantee automated interoperability of digital sources, it is essential to use widely adopted standardized vocabulary, ontology, thesauri with globally specific and permanent identifiers (i.e., principle F1), and well-defined data model structure for representing and structuring (meta) data. A means of defining and structuring the digital sources is the resource description framework (RDF), an extendable knowledge representation model (Vita et al. 2018).

“Principle I2: (meta)data use vocabularies that follow FAIR principles”

Principle I2 utilizes vocabulary to point to approaches, which explicitly describe terms that occur in a specific field. Moreover, the usage of mutual and properly organized terms is an integral portion of FAIR. Furthermore, constructing community standards require structured terminology containing fine-grained knowledge specifications such as data structures and ontologies and structured thesauri. The vocabulary used to describe (meta)data, though, must be searchable, reachable, interoperable, and reusable by itself so that people, as well as computers, can understand the context of the metadata terms used. Furthermore, computers should effectively differentiate the used vocabulary terms in RDF, the language for knowledge representation. Therefore, principle F2 demands that the vocabulary used must adhere to FAIR data principles (Lamprecht et al. 2020).

“Principle I3: (meta)data include qualified references to other (meta)data”

The objective of Principle I3 is to establish as many concrete ties as necessary across (meta)data in (meta)data repositories to improve contextual data knowledge. A qualified reference is a cross-reference defining its purpose. To be quite specific, if one dataset depends on another dataset, you must determine if more datasets are required to fulfill the whole data or if supplementary content is deposited in diverse datasets. In particular, the scientific ties between datasets must be specified by dataset owners. Furthermore, all datasets, and their globally unique and persistent identifiers, must be referenced (Lamprecht et al. 2020).

Reusable R

“Principle R1: (meta)data are richly described with a plurality of accurate and relevant attributes”

When many annotation tags are assigned to the data, it would be much faster and easier to discover and reuse data. Principle R1 is same as F2, but R1 depends on a user’s (machine or human) capacity to determine if the data in a specific context is useful. To make this determination, the data owner should

include not only metadata that facilitates exploration but also metadata that richly represents the context under which the data was created. These activities could involve the research method, the producer, and the model of the instrument or device that provided the data, the organisms that used the drug regime, and any other source of data. Moreover, R1 declares that the owner of data does not intend to anticipate the identity and desires of the data user. To suggest that the metadata creator should be as lavish as possible in supplying metadata, even though with details that might seem unrelated, we used the phrase “plurality” (Jacobsen et al. 2020).

“Principle R1.1: (meta)data are released with a clear and accessible data usage license”

Principle R1.1 is a matter of legal interoperability. What rights of access apply to data? These rights of access must be explicitly defined. Uncertainties can dramatically impact the reuse of your data by companies struggling to conform to licensing restrictions. In addition, the clarity of licensing status would become more substantial with digital searches requiring more licensing criteria. Without exception, digital resources and their (meta)data must always contain a license defining the state of its use. Moreover, it must be apparent to humans and computers the circumstances under which the digital resources will be used. Furthermore, the license specified must be as open as possible to enable reuse, therefore licenses widely used, such as MIT and Creative Commons, can be connected to your digital resource (Labastida and Margoni 2020).

“Principle R1.2: (meta)data are associated with detailed provenance”

Principle R1.2 indicates that a precise history of provenance should be known in order to reuse the digital resource. Furthermore, provenance addresses detailed information covers factors like: 1) How the digital resource was created? 2) Why it was produced? 3) Under what circumstances, based on which original data or digital resource of origin and what grant is used? 4) Who manages the data that should be attributed? and 5) What are the modifications and cleaning methods used since creation? Both humans and computers need provenance information and data manipulation techniques to determine whether a digital resource meets their requirements for intended reuse (Jauer and Deserno 2020).

“Principle R1.3: (meta)data meet domain-relevant community standards”

Principle R1.3 declares that (meta)data ought to comply with community standards or community best practices if they exist. Moreover, minimal information standards (e.g., MIAME, MIAPE, and ISO 19115-2) have been established by many organizational communities, defining the minimum

collection of metadata items needed to determine data processing and the data extraction standard to promote reusability. Furthermore, the FAIR data demand to at least conform to these standards; other community criteria are less standardized, but the purpose of FAIRness is to present (meta)data in a way that improves its reusability for the community. Eventually, these standards are a decent start point, considering that real reusability would usually involve additional richer metadata (Jacobsen et al. 2020).

Summary and Conclusions

We have defined the FAIR data principles and explained the enclosed 15 FAIR data principles. Toward better data management and stewardship, digital resources ought to be findable, accessible, interoperable, and reusable by human and machine. Furthermore, several aspects are explained to help improve data management and stewardship for digital resources.

Finally, scientific communities are confronted with several challenges such as lack of metadata standard, struggle to locate needed datasets, ruinous metadata authoring process, and longtime spent on data curation. Therefore, the solution to rectify these issues is to have satisfying data management and stewardship that relies on adopting FAIR data principles.

Acknowledgments This book chapter is based upon work supported by the National Science Foundation under Grant No. 2019609.

Bibliography

- Bailo D et al (2020) Perspectives on the implementation of FAIR principles in solid earth research infrastructures. *FrEaS* 8:3
- Brewster C et al (2020) Ontology-based access control for FAIR data. *Data Intell* 2(1–2):66–77
- Jacobsen A et al (2020) FAIR principles: Interpretations and implementation considerations. (2020):10–29
- Jauer M, Deserno T (2020) Data provenance standards and recommendations for FAIR data. *Stud Health Technol Inform* 270:1237–1238
- Juty N et al (2020) Unique, persistent, resolvable: identifiers as the foundation of FAIR. *Data Intell* 2(1–2):30–39
- Labastida I, Margoni T (2020) Licensing FAIR data for reuse. *Data Intell* 2(1–2):199–207
- Lamprecht A et al (2020) Towards FAIR principles for research software. *Data Sci* 3(1):37–59
- Martone M (2014) Joint declaration of data citation principles. *FORCE11*
- Vita R et al (2018) FAIR principles and the IEDB: short-term improvements and a long-term vision of OBO-foundry mediated machine-actionable interoperability. *Database* 2018
- Weigel T et al (2020) Making data and workflows findable for machines. *Data Intell* 2(12):40–46
- Wilkinson M et al (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci data* 3(1):1–9

Fast Fourier Transform

Frits Agterberg

Central Canada Division, Geomathematics Section,
Geological Survey of Canada, Ottawa, ON, Canada

Definition

The Fast Fourier transform (FFT) is an algorithm that computes the discrete Fourier transform of a 1-dimensional sequence or a 2- or 3-dimensional array. In general, Fourier analysis converts a signal from its original domain (usually time or space) to a representation in the frequency domain (and vice versa). The discrete Fourier transform of complex numbers $x_0, \dots, x_{N-1}$ can be written as:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-2\pi i k n / N} \quad k = 0, 1, \dots, N-1$$

where $e^{2\pi i / N}$ is the primitive N th root of 1. In terms of big O notation, direct evaluation of this definition requires $O(N^2)$ operations but the FFT requires only $O(N \log_2 N)$ operations. Other formulations and algorithms of the FFT can be found in Press et al. (2007). The preceding definition applies to 1-dimensional series. Savings in computing time are even greater in 2- and 3-dimensional applications of the FFT.

Introduction

The most commonly used FFT algorithm is the one devised by Cooley and Tukey (1965). Historically it is of interest that Heideman et al. (1984) discovered much later that the earliest version of the FFT can be traced to unpublished work by Gauss in 1805 who used it in an unpublished manuscript for interpolation to describe the orbits of two asteroids. A brief history of early FFT inventions and applications can be found in Cooley et al. (1967).

Many features in the Earth's crust tend to be periodic in that they repeat themselves at more or less the same intervals. Examples are anticlines and synclines as well as fault systems. In these situations, harmonic trend analysis can produce much better results than other 2-D interpolation techniques including polynomial trend surface analysis. The 2-D power spectrum and autocorrelation function are important tools even when there are no spatial periodicities. A 2-dimensional example of harmonic trend surface analysis using the Cooley-Tukey algorithm will be given later in this entry. The relationship between harmonic analysis and other 2-D

geostatistical methods also will be considered. The 2-D theory of FFT-based 2-D harmonic analysis can be summarized as follows (cf. Agterberg 1974, 2014).

The inverse Fourier transform $A(p, q)$ of a 2-D array of gridded $(m \times n)$ data, $X_A(i, j)$ satisfies:

$$A(p, q) = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} X_A(i, j) \cdot \exp \left[-2\pi I \left(\frac{ip}{m} + \frac{jq}{n} \right) \right]$$

where $I = \sqrt{-1}$.

Consequently, $X_A(i, j) = \frac{1}{mn} \sum_{p=0}^{m-1} \sum_{q=0}^{n-1} A(p, q) \cdot \exp \left[-2\pi I \left(\frac{ip}{m} + \frac{jq}{n} \right) \right]$.

Without loss of generality, it can be assumed that $\bar{X}_A$ is zero. Any value $A(p, q)$ consists of a real part $\text{Re}(A_{p,q})$ and an imaginary part $\text{Im}(A_{p,q})$. Any wave can be represented by the continuous function:

$$X(p, q; u, v) = \text{Re}(A_{p,q}) \cos \left[2\pi \left(\frac{ip}{m} + \frac{jq}{n} \right) \right] - \text{Im}(A_{p,q}) \sin \left[2\pi \left(\frac{ip}{m} + \frac{jq}{n} \right) \right],$$

where u and v are geographical coordinates. The square of amplitude is given by:

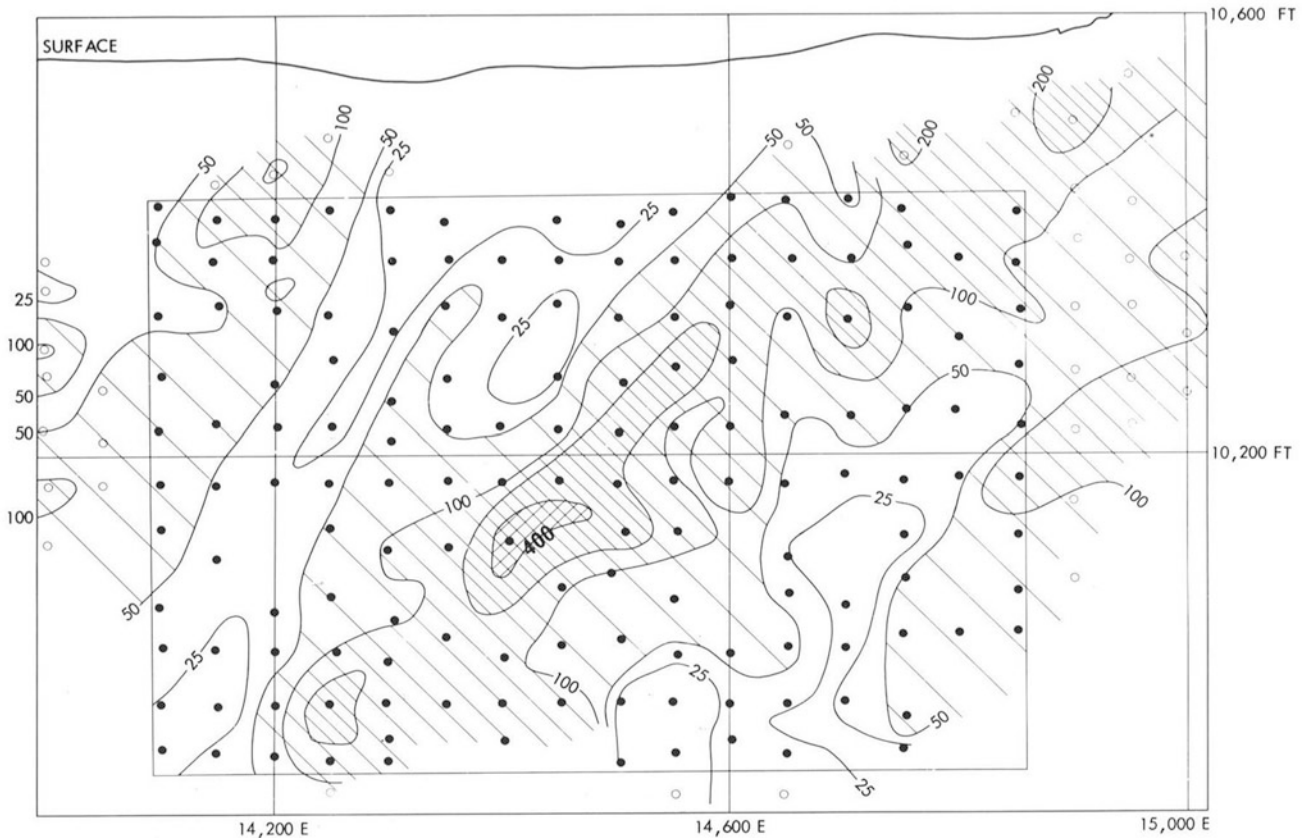
$P(p, q) = \text{Re}^2(A_{p,q}) + \text{Im}^2(A_{p,q})$ and the 2-D autocovariance satisfies:

$$C(r, s) = \sum_{p=0}^{m-1} \sum_{q=0}^{n-1} P(p, q) \cdot \exp \left[2\pi I \left(\frac{ip}{m} + \frac{jq}{n} \right) \right].$$

The $X(p, q; u, v)$ functions for a block of values (p, q) then form a so-called harmonic trend surface with: $Y(u, v) = \sum_p \sum_q X(p, q; u, v)$.

Application to Drill-Hole Data from the Whalesback Copper Deposit

Figure 1 shows a longitudinal cross-section of the Whalesback copper deposit, Newfoundland. It contains the intersection points of 188 underground drill-holes with the central part of the ore zone. The average copper value was calculated for each hole over the width of the orebody. This



Fast Fourier Transform, Fig. 1 Manually contoured map of percent-foot values for copper in longitudinal section across Whalesback deposit, Newfoundland. Each dot represents the approximate intersection point

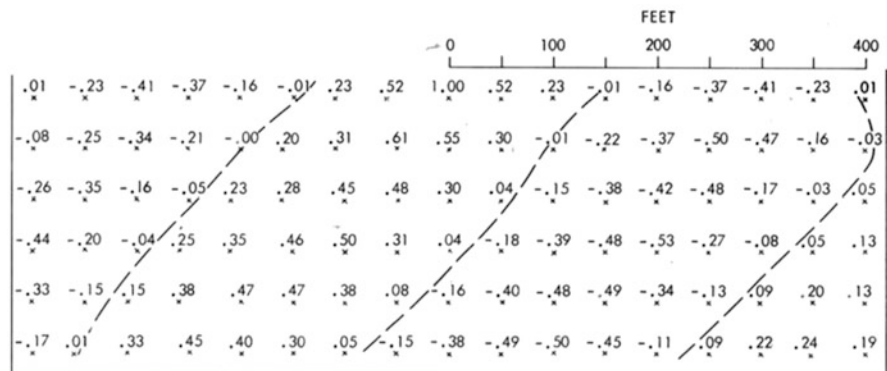
of an underground borehole with the center of the mineralized zone. (Source: Agterberg 1974, Fig. 47)

Fast Fourier Transform, Table 1 Matrix of underground drill-hole data for area depicted in Fig. 1. Values represent products of horizontal width of orebody (in feet) and average concentration value of copper (in percent). (Source: Agterberg 1974, Table XXVIII)

29	146	116	76	22	6	(23)	11	18	48	60	94	9	32	137	128
51	47	91	(60)	12	21	43	32	40	52	158	139	94	77	118	77
43	60	94	34	19	67	31	22	65	112	126	78	258	80	(86)	110
91	(70)	60	34	66	29	(41)	32	164	286	160	(126)	(97)	(97)	74	70
91	73	68	21	75	44	35	73	297	89	140	52	66	49	52	43
105	43	36	42	40	77	83	272	209	107	177	40	29	42	52	199
66	32	(47)	58	128	114	354	(194)	284	169	(91)	25	(34)	9	(61)	27
47	(38)	31	86	76	(159)	(150)	113	43	78	(69)	40	21	64	(51)	35
41	8	41	125	196	128	118	74	76	33	70	65	29	64	100	57
19	19	62	260	135	174	135	163	23	18	33	24	36	95	(67)	(78)
20	35	59	168	199	(146)	91	(85)	16	13	58	30	(40)	23	(59)	(68)

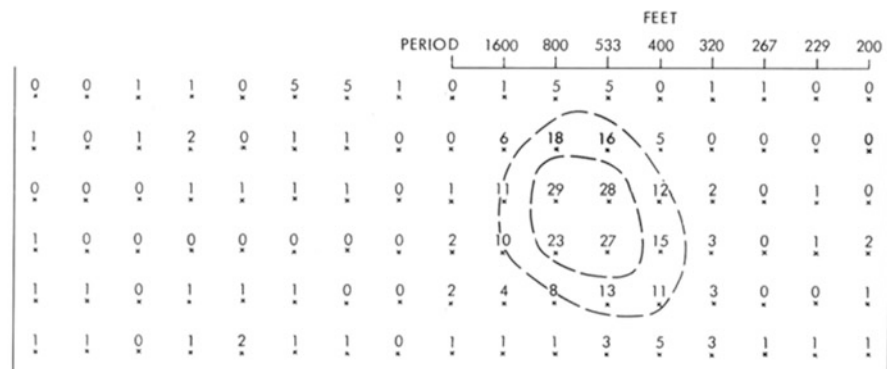
Fast Fourier Transform,

Fig. 2 Two-dimensional autocorrelation function for data from Table 1, Whalesback copper deposit. (Source: Agterberg 1974, Fig. 61)



Fast Fourier Transform,

Fig. 3 Part of 2-dimensional power spectrum for data in Table 1; sharply defined peak reflects zoning of copper percent-foot values shown in Fig. 1. (Source: Agterberg 1974, Fig. 77)



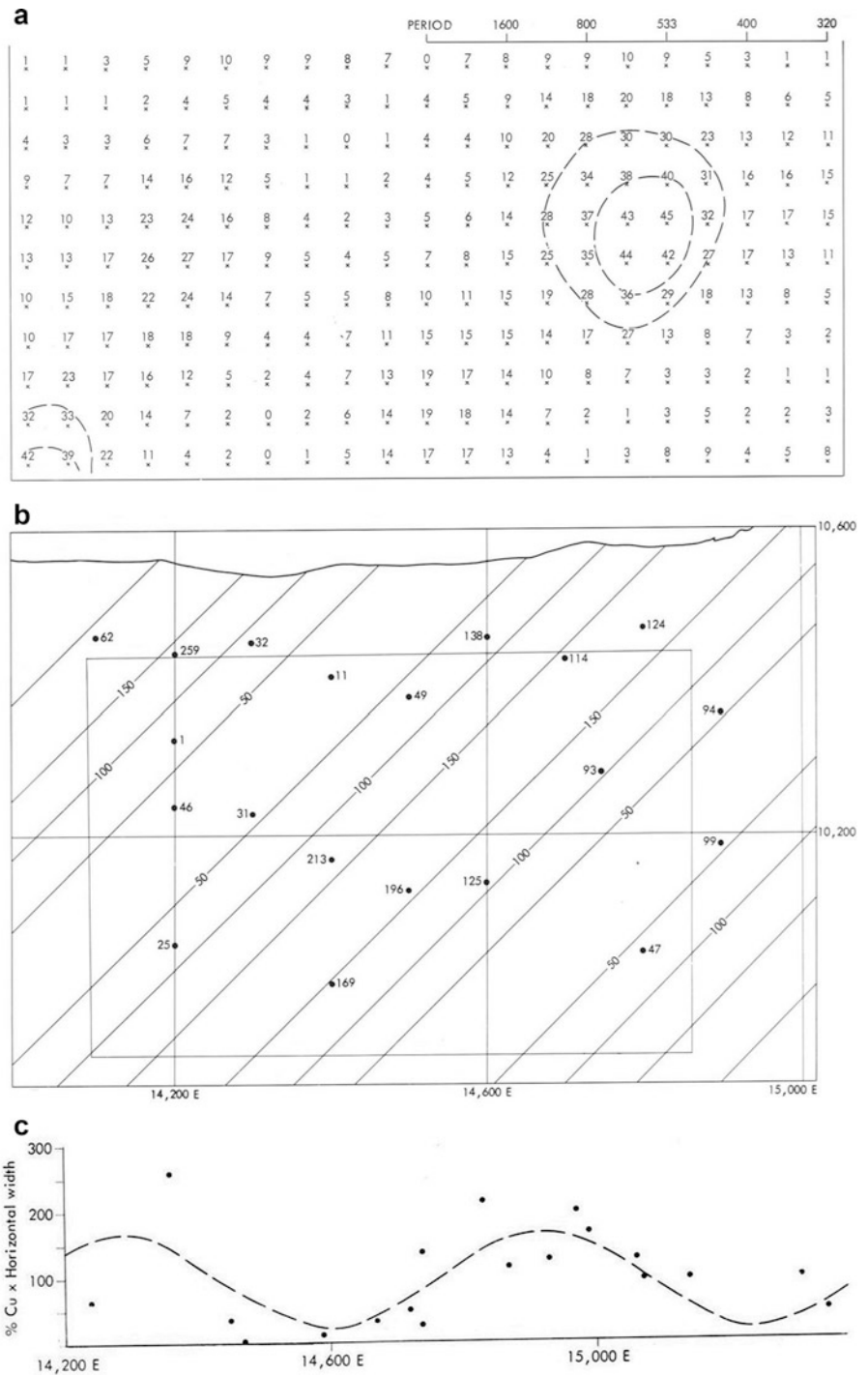
value was multiplied by horizontal width yielding so-called percent-foot values contoured by mining staff. The central part of this diagram shows a relatively rich copper zone that dips about 45° downward to the west.

For the FFT application, the array of Table 1, which is based on Fig. 1, was enlarged to size (32 × 32). The resulting

2-D autocorrelation function and power spectrum are shown in Figs. 2 and 3, both of which demonstrate the strong zoning in this ore deposit. The central copper zone is flanked by two elongated minima and at about 460 ft. from it there occur two other relatively copper-rich zones, although their maxima are not as high as that for the central zone. The following

Fast Fourier Transform,

Fig. 4 Harmonic analysis of irregularly spaced data (after Agterberg 1974). Intersection points of 20 exploratory boreholes from topographic surface with ore-zone shown in (b). Two-dimensional sine-waves were fitted to 20 values by least squares method; percentage explained sum of squares for each fitted wave is shown in (a) with period scale in feet. Wave for value 44 on peak in (a) is shown by straight-line contours in (b), and also in cross-section for horizontal line ($V = 10,600$) near topographic surface in (c). (Source: Agterberg 1974, Fig. 78)



experiment illustrates further that harmonic analysis can be an excellent exploration tool in situations of this type.

Development of the Whalesback copper deposit was performed at two different stages. Before the deposit was mined, it was explored by drilling relatively few bore-holes downward from the surface of the Earth. Statistical analysis applied subsequently to these original data is shown in Fig. 4.

Copper percent-foot values from these surface boreholes are shown in Fig. 4b. There were only 20 exploration values, which are irregularly distributed. The question can be asked of how the pattern of Fig. 1, which is based on 188 more precise underground drill-holes, which are more evenly spaced and were sampled in much more detail as well, could have been predicted from these 20 values. This kind of

problem is analogous to that of constructing a topographical contour map of a mountain chain when only a few elevations would be provided for interpolation. Simple polynomial trend surface analysis applied to the 20 development values produces patterns that show no resemblance to the pattern of Fig. 1. In general, trend surface results can be evaluated by using squared multiple correlation coefficients R^2 that provide a convenient measure to assess degree of fit of a trend surface. In its application to the 20 values of Fig. 4, both the linear ($R^2 = 0.16$) and quadratic ($R^2 = 0.32$) fits were found to be not statistically significant.

On the other hand, Agterberg (1974) fitted the function $X(p, q; u, v)$ (see previous section) by least squares to estimate the coefficients $b_1 = \text{Re}(A_{p,q})$ and $b_2 = \text{Im}(A_{p,q})$ for many possible directions of axes and periods for the waves. Ideally, a least-squares model should have been used by which the four parameters (direction of axis, period, amplitude, and phase) are estimated simultaneously but then a nonlinear model should have been developed and applied. The linear model can only be used when p/m and q/n are assumed to be known. For each separate sine wave, R^2 was plotted as a percentage value in Fig. 4a. The result can be regarded as a 2-D power spectrum for irregularly spaced data. Since three coefficients were fitted for each sine wave (including the constant term that is related to the mean value), its R^2 value could be converted into an F -value at the upper tail of a theoretical F -distribution with 3 and 17 degrees of freedom. The $F_{0.95}$ - and $F_{0.99}$ -values correspond to 27% and 37% values for R^2 and these were contoured. There are two contoured peaks in Fig. 4a and one of these corresponds to the peak in Fig. 1. The contour map of the sine-wave whose amplitude falls on the well-developed peak of Fig. 4a is shown in Fig. 4b, and its profile for a line along the surface ($V = 10,600$) in Fig. 4c. The 20 original observations were projected onto Fig. 4c along lines parallel to the contours. Obviously, the contours in Fig. 4b are an oversimplification of those shown in Fig. 1 but existence of zones with alternately high and low copper could already have been correctly predicted from as few as 20 bore-holes drilled from the surface.

Summary and Conclusions

The Fast Fourier transform (FFT) is an algorithm that computes the discrete Fourier transform of a sequence or a 2- or 3-dimensional array. It can be used to convert a signal from its original domain (usually time or space) to a representation in the frequency domain (and vice versa). In this entry, the 2-dimensional version of the Cooley-Tukey algorithm was used to illustrate that in mathematical geoscience this approach is particularly useful to analyze 2-dimensional map patterns that are spatially repetitive in nature. Copper

analyses from drill-cores from the Whalesback Mine in Newfoundland were taken for example. The FFT method was used to obtain 2-dimensional autocorrelation function and power spectrum for the spatial distribution of copper in this copper deposit. It was also shown that a relatively small number of bore-holes drilled from the surface during early exploration would already have yielded the intrinsic pattern of copper mineralization obtained by harmonic analysis of the later underground mining data.

Cross-References

- [Autocorrelation](#)
- [Power Spectral Density](#)
- [Predictive Geologic Mapping and Mineral Exploration](#)
- [Trend Surface Analysis](#)

Bibliography

- Agterberg FP (1974) *Geomathematics*. Elsevier, Amsterdam
- Agterberg FP (2014) *Geomathematics: theoretical foundations, applications and future developments*. Springer, Dordrecht
- Cooley JW, Tukey JW (1965) An algorithm for the machine calculation of complex Fourier series. *Math Comput* 19(90):297–301
- Cooley JW, Lewis PAW, Welch PD (1967) Historical notes on the fast Fourier transform. *IEEE Trans Audio and Electroacoust* 15(2): 76–79
- Gauss CF (1866) Theory regarding a new method of interpolation. *Werke* (in Latin and German). Königlichen Gesellschaft der Wissenschaften zu Göttingen 3, pp 265–303
- Heideman MT, Johnson DH, Burrus CS (1984) Gauss and the history of the fast Fourier transform. *IEEE ASSP Mag* 1(4):14–21
- Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2007) *The art of scientific computing*, 3rd edn. Cambridge University Press, New York

Fast Wavelet Transform

Indu Solomon and Uttam Kumar

Spatial Computing Laboratory, Center for Data Sciences,
International Institute of Information Technology Bangalore
(IIITB), Bangalore, Karnataka, India

Definition

A signal is said to be stationary if its frequency content is invariant with time. These signals can be analyzed with techniques such as Fourier transform. Nonstationary signals, on the other hand, are those whose frequency content changes with time. The nonstationary signals are analyzed using wavelet transform as it helps to decompose the complicated

signal into simpler components. Wavelet transform is constituted by wavelets that are small oscillations or wave-like signals. Scale and location are two main properties of wavelets. Scale of the wavelet defines the stretch or squeeze of the wavelet, and location property gives the time or space positioning. The faster version of wavelet transform known as Fast wavelet transform was introduced by Mallat (1989).

Introduction

A brief history of wavelet transform (Chun-Lin 2010) needs a mention here. Haar wavelet was proposed in 1909 by the mathematician Alfrd Haar, but the concept of wavelets did not exist at that time. Later, wavelet was proposed in 1981 by the geophysicist Jean Morlet. Eventually, the term wavelet was introduced by Morlet and Alex Grossman in 1984. The second orthogonal wavelet after Haar wavelet was introduced in 1985 by mathematician Yves Meyer and is called Meyer wavelet. Multiresolution analysis concept was introduced by Mallat and Meyer. Compact support orthogonal wavelet was introduced by Daubechies. Fast wavelet transform was proposed in 1989 by Mallat, and later on various applications of wavelet transform in the field of signal processing were identified.

A wavelet is a mathematical function which is used to decompose a given signal into different scale components. The wavelets are scaled and translated copies of fast decaying waveform called mother wavelet. The scaled and translated copies are called daughter wavelets. There are numerous applications for wavelet transforms in the field of signal processing viz. data compression, signal denoising, image compression, signal analysis, etc. Wavelet transforms are widely used in geoscience applications as well as for example, spatial analysis tool for modeling complex multivariate geographic relationships. Wavelets are also used in spatial statistics and spatial analysis of landscape patterns.

Wavelet Transform

The continuous wavelet transform (CWT) is defined as follows:

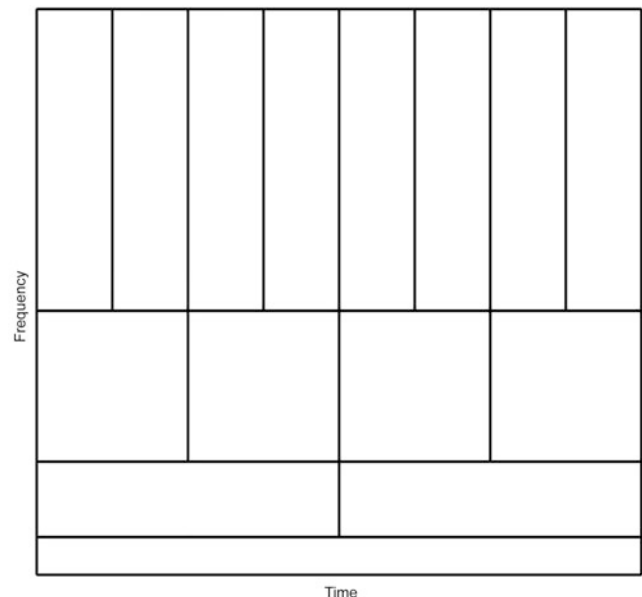
$$\Psi_x^\psi(\tau, s) = \frac{1}{\sqrt{|s|}} \int x(t) \psi^*\left(\frac{t-\tau}{s}\right) dt \quad (1)$$

The transformed signal is a function of two variables τ and s , where τ represents translation and s represents the scale. The transforming function $\psi(t)$ with which the signal is convoluted is called the mother wavelet. The mother wavelet is the main function from which other windowing functions called daughter wavelets are generated. The term translation

refers to the location of the window function, and translation captures the time information of the signal (Polikar 1996). The scale parameter is inverse of the frequency of wavelet. Low scales of wavelet give very fine details, and high scales give lesser details. The scale $s > 1$ elongates the signal and $s < 1$ compresses the signal. The wavelet function is convoluted with the given signal by translating through the time axis at multiple scales and the transformed values are stored for further processing. Unlike other transformations, the main advantage of wavelet transform is that it can simultaneously extract spectral and temporal information. Multiresolution analysis is used to analyze a signal at different frequencies with different resolutions. Multiresolution analysis is useful in giving good frequency resolution and poor time resolution at low frequencies (Morehart et al. 1999), and good time resolution and poor frequency resolution at high frequencies. This approach is helpful for analyzing most naturally occurring signals, as they have slow varying low-frequency components and fast varying high-frequency components. Multiresolution interpretation diagram for wavelet transforms is shown in Fig. 1. Even though the height and width of each box is different in Fig. 1, their area is same. At low frequencies, the height of the boxes are shorter and width are longer representing better frequency resolution and poor time resolution.

CWT Versus DWT

Continuous wavelet transform (CWT) decomposes a signal into wavelets over the whole real axis and hence over infinite

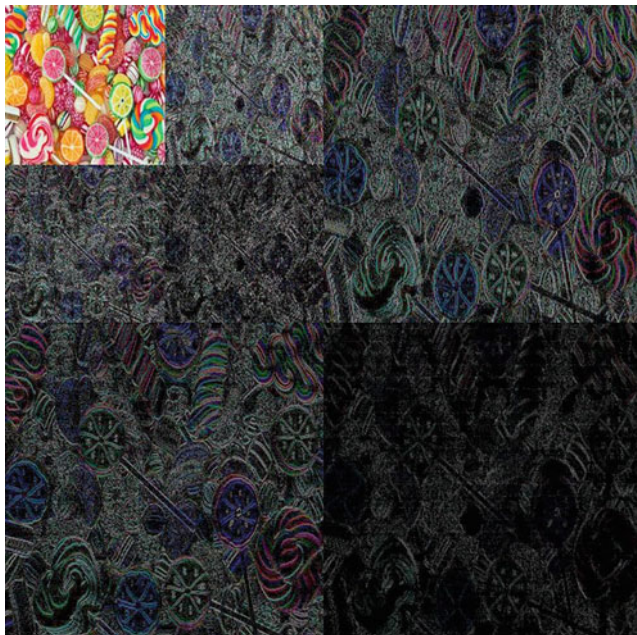


Fast Wavelet Transform, Fig. 1 Time and frequency resolution interpretation

scales and locations while the discrete wavelet transform (DWT) samples the signal over a range of integers and uses a finite set of wavelets. Since CWT is computationally intensive, DWT is more suited for real-time applications. The Mallat algorithm for DWT is known as fast wavelet transform. It is known as a two-channel subband coder using conjugate quadrature filters or quadrature mirror filters (QMFs).

Fast Wavelet Transform (FWT)

In case of FWT, the signal is passed through a set of filters to extract the details of the signal. Low-pass filters and high-pass filters are used to extract low-frequency and high-frequency information, respectively. DWT decomposes the signal into coarse approximation and fine detail information. The decomposition halves the time resolution and doubles the frequency resolution and can be repeated multiple times. The process is called subband coding which reduces the uncertainty in frequency. Figure 2 shows an example of subband decomposition of an image. The decomposition algorithm starts with signal S . S is decomposed into approximation coefficients A_1 and detail coefficient D_1 , and then A_1 is further decomposed into approximation A_2 and detail D_2 . The inverse transform starts with approximation coefficients A_n , detail coefficients D_n , and computes A_{n-1} , and so on. FWT is a power-of-two algorithm. Each successive decomposition of the FWT operates on a vector which is half the size of the previous operation. This is achieved by decimation in time of the multiresolution analysis. Lossless and lossy compression are



Fast Wavelet Transform, Fig. 2 Subband decomposition of an image

two main applications of FWT in image compression applications (Boyer 1995). FWT was introduced to remove the border effects caused in DWT. FWT algorithm is suitable for computationally intense geospatial applications too.

Wavelet Families

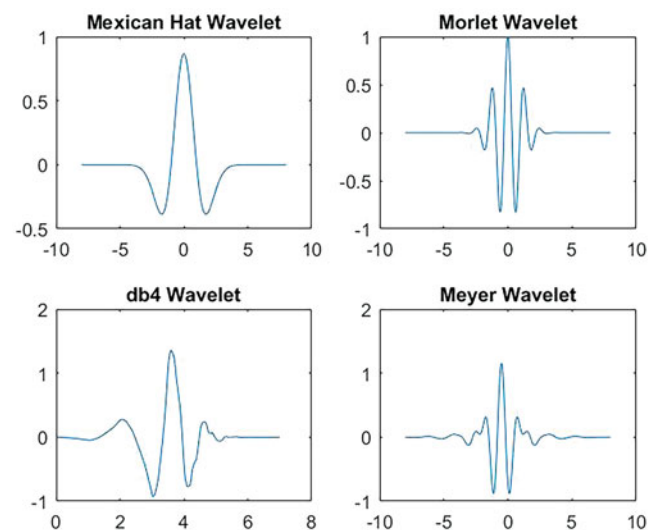
There are various wavelet families and the selection of best family is based on the shape of the signal to be processed. Few of the wavelet families are Haar, Daubechies, Biorthogonal, Coiflets, Morlet, Mexican Hat, and Meyer. Figure 3 shows examples of the shape of wavelet function for selected wavelet families.

DWT: Case Studies in Geospatial Applications

There are many geospatial domains in which wavelet transforms are extensively used, namely tree detection (Stupariu et al. 2020), ecology (Dale and Mah 1998), crop yield estimation in agriculture, cloud detection, spatial traffic analysis, soil variation study, and so on. This section will discuss some important applications of DWT in geospatial domain.

Classification of Hyperspectral Images Based on 3D: DWT

Hyperspectral image consists of a large number of spectral bands of narrow bandwidth and the number of spectral bands (channels) often depends on the sensor. Hyperspectral images are useful in a wide variety of applications including geology, mining, agriculture, mineralogy, environment, surveillance, and so on. Wavelet transforms are very useful for non-stationary signals and can extract location (spatial) and frequency (spectral) information from the signals. 3D discrete wavelet transformation is generally used for extracting the spatial and spectral features from the hyperspectral images and these extracted features are used as input to a trained



Fast Wavelet Transform, Fig. 3 Example – wavelet families

classifier for classification (Anand et al. 2021). Three-dimensional wavelet transform was implemented with the help of three-stage 1D wavelet transform using Haar, Fejér-Korovkin, and Coiflet wavelet. First step of the algorithm uses principal component analysis (PCA) for reducing the dimensionality of the data to get compact representation. The second step is to apply 1D - DWT to the rows of the compact data and feed the outputs of 1D - DWT along with the columns to generate the wavelet transform in different image dimensions. This creates four subbands with spatial and spectral features that are then fed to the classifier. The classification experiments were performed on the Salinas and Indian Pines hyperspectral dataset.

Remotely Sensed Image Data Fusion: A Wavelet Transform Approach for Urban Analysis

There are several existing techniques for fusing panchromatic and multispectral data to obtain high spatial and spectral resolution images. Compared to the traditional techniques, multiresolution wavelet decomposition based techniques give a better trade-off between spatial and spectral information (Fanelli et al. 2001). Prerequisite for fusion of two images is that both the images must be of the same spatial resolution. If this is not the case, either we upsample or downsample one of the images to match the resolution. Upsampling is preferred over downsampling because downsampling a high resolution image takes away important details from the image. In Fanelli et al. (2001), bicubic resampling technique was used for resampling the IKONOS multispectral images to the resolution of IKONOS panchromatic image. Brovey, HSV, and PCA transformations are a few classical techniques used in literature for fusion of satellite images. These classical techniques are simple and easy to implement, but they distort the spectral (thematic) details in the fused images.

Multiresolution wavelet decomposition based fusion technique such as Dyadic Wavelet Substitution (DWS) solves the issues of classical fusion techniques. DWS is as summarized below:

- Select the wavelet basis family and the required wavelet decomposition level.
- Apply histogram matching between panchromatic image and selected multispectral band.
- Perform wavelet decomposition of panchromatic image and selected multispectral band to obtain low-frequency approximation and high-frequency detail images.
- Replace the low frequency approximation image of panchromatic wavelet decomposition with the high spectral content (low-frequency) approximation image of the multispectral data.
- Perform inverse wavelet decomposition to obtain high spectral and spatial content multispectral fused images.

The abovementioned wavelet decomposition based fusion technique is very simple and effective in retaining the spectral details compared to classical fusion techniques.

Evaluation of Spatiotemporal Characteristics of Precipitation Using Discrete Maximal Overlap Wavelet Transform and Spatial Clustering Tools

The monthly precipitation characteristics of 30 stations in southeastern United States during the time period of 1968–2018 were studied in a recent research (Roushangar et al. 2021) where maximal overlap discrete wavelet transform (MODWT) was used for data preprocessing. Daubechies (db) wavelet family was selected for processing and the data were decomposed into various subbands using MODWT. Finally, K-means clustering was used for grouping the data.

MODWT is a modified version of DWT. It can perform a multiresolution analysis of the time series signal. MODWT projects the time series data onto a collection of non-orthogonal basis wavelet functions. For a time series P with N samples, MODWT is given by

$$P = \sum_{j=1}^N W_j + V_N \quad (2)$$

where W_j is the wavelet coefficients, V_N is scaling coefficients, and N is the decomposition level. The output of MODWT was fed as input to the clustering model. Energy E values for each of the wavelet subband for each wavelet and scaling factor were calculated from the precipitation values.

The MODWT-K-means framework was used for investigating spatiotemporal analysis of precipitation. MODWT was used to extract multiresolution features of nonstationary data, and K-means could locate spatially identical clusters on wavelet-transformed feature space.

Summary and Conclusion

Wavelet transforms are very useful for analyzing non-stationary or time series signals. They are capable of analyzing a signal at different scales and different times known as multiresolution analysis. There are different varieties of wavelet families with different basis functions, and we can select a suitable wavelet family depending on the signal to be analyzed. Wavelet transforms are extensively used for signal-processing and image-processing applications. FWTs are capable of removing the border artifacts caused due to DWT and are extensively used in image compression applications. Wavelet transforms are robust and useful for geospatial

applications as well. Few case studies of geospatial applications which used wavelet transforms were discussed in this chapter.

Cross-References

- [Fast Fourier Transform](#)
- [Hyperspectral Remote Sensing](#)
- [Wavelets in Geosciences](#)

Bibliography

- Anand R, Veni S, Aravinth J (2021) Robust classification technique for hyperspectral images based on 3D-discrete wavelet transform. *Remote Sens* 13(7):1255
- Boyer KG (1995) The fast wavelet transform(FWT) (Master's thesis, in partial fulfillment of the requirements for the degree, Master of Science, Applied Mathematics. University of Colorado at Denver)
- Chun-Lin L (2010) A tutorial of the wavelet transform. NTUEE, Taiwan, pp 21–22
- Dale MRT, Mah M (1998) The use of wavelets for spatial pattern analysis in ecology. *Journal of Vegetation Science* 9(6):805–814
- Fanelli A, Leo A, Ferri M (2001) Remote sensing images data fusion: a wavelet transform approach for urban analysis. In *IEEE/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas* (Cat. No. 01EX482) (pp. 112–116). IEEE
- Mallat SG (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans Pattern Anal Mach Intell* 11(7): 674–693
- Morehart M, Murtagh F, Starck JL, Bi Y (1999) Multiresolution spatial analysis. *Proc. Geo Computation*
- Polikar R (1996) Fundamental concepts and an overview of the wavelet theory. *The Wavelet Tutorial Part I*, Rowan University, College of Engineering Web Servers, p 15
- Roushangar K, Moghaddas M, Ghasempour R, Alizadeh F (2021) Evaluation of spatial–temporal characteristics of precipitation using discrete maximal overlap wavelet transform and spatial clustering tools. *Hydrology Research* 52(2):414–430
- Stupariu MS, Pleş oianu AI, Pătru-Stupariu I, Fürst C (2020) A method for tree detection based on similarity with geometric shapes of 3D geospatial data. *ISPRS International Journal of Geo-Information* 9(5):298

Favorability Analysis

Guocheng Pan

Hanking Industrial Group Limited, Shenyang, Liaoning, China

Definition

Mineral target selection has been an important step for search of mineral deposits in mineral exploration. Significant progress has been made so far in development of mathematical

techniques and estimation methodologies for mineral target selection. Integration of multiple data sets, either by experts or statistical methods, has become a common practice in estimation of mineral potentials. Geoscience map patterns are commonly used as primary inputs for decision-making on mineral target selection. Favorability analysis as a broadly adopted methodology for data integration and target selection can be effectively applied by using modern GIS technologies. The outputs of the analysis are prognostic maps that indicate where hidden ore bodies or mineralized objects may occur. Favorability analysis is a whole class of mathematical tools that are capable of quantifying and integrating diverse geoscience information for mineral potential mapping. This chapter will introduce the concepts of geoscience information synthesis and discuss rudimentary favorability modeling techniques, including characteristic analysis and canonical favorability analysis.

Introduction

The activities of mineral exploration have been motivated chiefly by economic and social pursuits (Pan and Harris 2000). Constantly growing economic and social demands require greater amounts of raw material, including non-renewable mineral commodities. The conduct of mineral resource exploration is predicated upon the economic rent expected from the discovery of new deposits. An increase in the price of a mineral product, which is equivalent to the sum of the marginal rent and marginal extraction cost, indicates that the mineral resource has become scarce. A basic perspective of both geologists and economists is that mineral resources are scarce materials in the crust as they occupy only an insignificant portion of crustal material.

Any major ore deposit may be regarded in principle as an anomalous or rare phenomenon commonly characterized by one or more geological, geochemical, and geophysical features. Consequently, signatures of significant endogenic mineralization are anomalous and exceptional geologic settings created by earth processes (Pan 2010, 2018). In particular, the formation of a giant deposit is an extremely rare event created by an exceptional combination of earth processes. Rareness of the giant deposits is reflected in both spatial and temporal dimensions. Significant concentrations of a metal usually have a strong affinity or correlation with special geologic formations and epochs, as well as metallogenic environments. The genesis of giant deposits may be controlled by special regularities that differ from those controlling the formation of medium and small-size deposits of the same composition.

Simply stated, mineral exploration is the process of searching for undiscovered mineral deposits. While such a statement is correct, it belies the complexity and diversity of tasks and decisions that constitute modern exploration. In

private-enterprise economies, mineral exploration is both a process of scientific detection and a business investment. To be successful on a continuing basis, exploration must find not just an ore deposit, but a deposit that can be developed technically and produced economically. This compounds difficulties and risk, because such deposits generally constitute only a small percentage of those deposits that occur in a region. Nevertheless, successful discovery of such a deposit initiates profitable economic activity that may endure for many years, providing cash flow for future exploration and development ventures.

Exploration bears a high risk. The risk occurs in the exploration investment that may not lead to an economic discovery. Hence, the key for a successful exploration program is to minimize the uncertainty of target selection. Reduction of the uncertainty of drilling targets helps increase the chance of new discoveries and on average shorten the cycle of a successful exploration process. This can be achieved by full use of available data by means of mathematical integration technologies, such as Favorability Analysis.

Basic Concepts

A major task in modern exploration programs aims at the delineation of targets for the search of blind ore bodies as mineral deposits outcropped or near the earth surface are quickly depleting. Due to the advent of advanced information technologies, various multivariate techniques have been introduced to interpret and integrate complex and multiple geoscience maps for target selection. Favorability analysis is one of the most powerful tools for such a task (Pan and Harris 1992a; Pan 1993a). The idea of favorability analysis is to construct a linear combination of a set of selected geoscience variables relevant to the mineral deposit or occurrence of interest.

Classification of Spatial Variables

Mineralization is always associated with some special physical and chemical properties within a favorable mineralization environment. For example, an epithermal gold occurrence is often relevant to a heat source or an event of igneous intrusion or volcanic eruptions, which provided heat and fluids necessary for metal movement and concentration. The enriching process of a metal is generally related to multiple geological, geochemical, and geophysical processes in the spatial domain that contains the mineralization. This most often involves structural movements, faulting, and hydrothermal alterations as signatures of metal concentration.

Descriptors of mineral deposit or occurrence can be direct or indirect variables, which can take continuous or discrete (including categorical) values. It is customary that continuously valued attributes like magnetic measurements are considered as quantitative variables, while discrete or categorical attributes such as rock type are treated as qualitative variables. Soil geochemical measurements are typically quantitative, while presence or absence of mineral occurrence is regarded as qualitative binary data. Most quantitative or qualitative variables that define unique components of mineralization belongs the category of direct descriptor, such as gold grade of a gold deposit, copper grade in a porphyry copper deposit, and presence of mineral occurrence.

Each geoscience attribute can be classified as either “target” variable or “explanatory” variable (Pan and Harris 1992a; Pan 1993a). The attributes that provide direct evidence on the mineral occurrence of interest are referred to as target variables, such as mineral resource descriptors, metal grades, and key hydrothermal alteration assemblage. These features are usually known only in the best explored parts of the study region and are unevaluated in the rest of the region. Target variables characterize the major objective of a study and serve as identification criteria for target selection.

The attributes that contain indirect information on the mineral occurrence of interest are referred to as explanatory variables, which consist of various geoscience measurements, such as chemical and physical descriptors of favorable geological objects. Explanatory variables are generally observable throughout the entire study region by means of conventional sensing and sampling technologies. Stream sediment geochemical measurements, gravity observations, and intensity of magnetic flux are a few examples of explanatory variables.

Geophysical surveys generate a broad range of physical property measurements, including magnetic intensity, gravity density, IP conductivity, electro-magnetic resistivity, etc. Most of these are usually treated as explanatory variables that provide indirect information. Geochemical prospecting provides information on both explanatory and target variables. Anomalous gold assays from rock samples exhibit direct evidence on the gold mineralization, which can be treated as target variables. Concentrates of associated elements, such as As and Hg, reveal indirect signatures, which are used as explanatory variables. Table 1 contains an example of the two types of geological variable identified for an epithermal gold-silver district.

Lithology carries both target and explanatory information. For example, presence of well-defined mineralized host rocks on a map area can be treated as a useful target variable that provides firm or direct evidence for the presence of mineralization, whereas post-mineralization units laying over the host rock may be viewed as explanatory characteristics.

Favorability Analysis, Table 1 Example of explanatory and target variables

Variable Class	Label	Descriptions of the Information Fields	Data Type
Explanatory Variables	Z1	Preprocessed geochemical fields	Continuous
	Z2	Synthesized structural fields	Categorical
	Z3	Band-pass gravity fields	Continuous
	Z4	Band-pass magnetic fields	Continuous
	Z5	Ratio of EM fields	Continuous
	Z6	Correlated geophysical fields	Continuous
	Z7	Host rocks of gold-silver deposits	Ternary
	Z8	Depth estimates to hidden intrusives	Continuous
Target Variables	Y1	Hydrothermal alteration assemblages	Binary
	Y2	Mineralized Tertiary intrusives	Binary
	Y3	Gold-silver mines or occurrences	Discrete

Similarly, local ore-controlling shear zones are usually recognized as powerful spatial target variables, but regional contemporary structures like faulting and folding may be employed as explanatory variables.

Quantification of Geoscience Maps

Geological bodies like ore deposits, intrusive, and fault zones are typical 3D objects. Most geoscience attributes are considered as spatial random variables, each carrying spatial location and geologic typicality in a 3D space. For example, a magnetic observation is viewed as a random variable realization at the sampled location and the measurement physically exhibits magnetic intensity of geological objects at and around that location. However, most data integration techniques for target selection are carried out in 2D space due to restriction of information acquisition in the third dimension (like drilling). In practice, data integration is usually performed on geoscience maps or “layers,” each of which contains a geoscience attribute of interest. Combination of these multilayers is facilitated greatly with assistance of modern GIS technologies.

Typical geoscience layers are geological, geochemical, and geophysical maps. Geological maps include rock formations, structures, alterations, and mineral occurrences. Geochemical prospecting generates maps of rock samples, stream sediment, and soil samples. Geophysical surveys create various maps of physical properties, such as gravity, magnetic flux, and EM fields. Spatial target variables provide direct and firm evidence for mineral anomalies or deposits on a map, whereas spatial explanatory variables are expressive of indirect evidence or relevant signatures for mineralization.

Spatial variables are usually observed or sampled on a map or study region which is divided into many equal-sized cells, also referred to as inter-grid cells (see Fig. 1). Each cell

on the map is considered as a sampling point (1D) or area (2D) or block (3D), where information on the value and location is recorded for each spatial variable involved in the analysis. In many cases, geoscience variables are not sampled in a regular grid and they must be preprocessed into cell-based observations. For instance, a stream sediment sampling generally follows stream basin patterns instead of a regular grid. The sediment anomalies should then be restored to their sources and interpolated into a regular grid prior to further modeling.

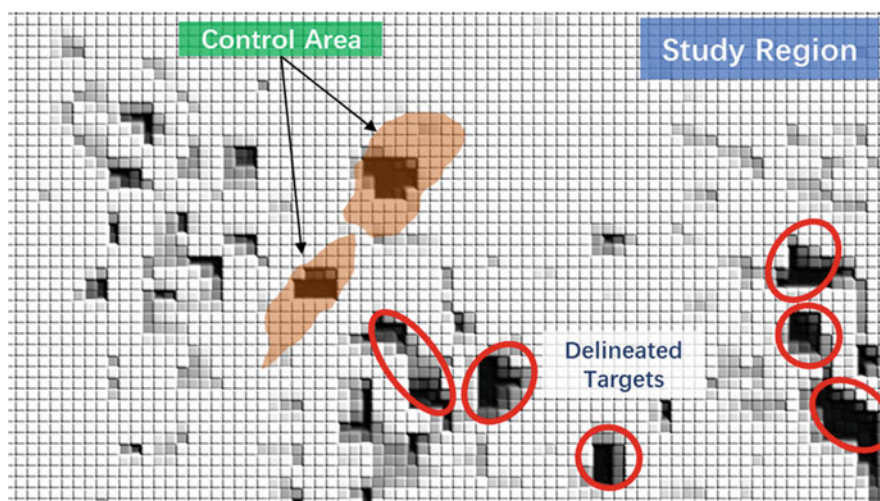
It is often that original maps are preprocessed before being used in data integration. For instance, magnetic and gravity maps preprocessed for noise removal and signal enhancement prior to integration are more informative in revealing anomalous basement structures and heat sources at depth. Geochemical maps after removal of regional background by trend analysis reveal subtle anomalous signatures for the mineralization of interest.

A fundamental principle in quantitative target evaluation is to assume that the data indeed provide reasonable descriptions on the nature of mineralization and their variability in space. The descriptions can be either quantitative or qualitative with the reference of evidential nature of spatial mineral distributions. The accuracy of target prediction largely depends on the amount of information contained in the data sets and captured in quantitative modeling.

Prior to data integration, the map variables must be assigned with values in each cell. The value assignment to the j^{th} geoscience variable at cell i (z_{ij}) is relevant of both science and art. Appropriate assignment requires scientific knowledge and ad hoc judgment as well. Geophysical and geochemical data are usually continuous and the cells can be assigned with actual measurements or processed values. Most geological features on a map are discrete or categorical and their cells are usually assigned with qualitative values in binary and ternary forms (Table 1).

Favorability Analysis,

Fig. 1 Inter-grid based favorability analysis (CFM) for mineral potential mapping



Control Area and Statistical Analogy

Most favorability analysis for mineral potential mapping is constructed on some pre-selected well-explored subregions on a map, which are collectively called the Control Area (Fig. 1). The favorability model established in the control area can then be extrapolated into unknown regions for favorability estimation using the principle of analogy. A control area can be one map region or a collection of several disconnected subregions. Target variables are usually known or observable within the control areas, whereas explanatory variables are generally available or measurable in the entire study map. An area that qualifies as control area must satisfy the following conditions:

- The area must be well-explored and all target variables and explanatory variables involved can be recorded or observed.
- The mineralization style of interest in the area is typical and representative of the entire study region.
- The characteristics of association between target and explanatory variables established in the control area are representative or roughly homogeneous throughout the entire study region.
- Major deterministic conditions for ore-controlling geological formations and structures in the control area are relatively uniform in other areas of the study region.

Data integration using the control area assumes the fundamental principle of statistical analogy, which implies that similar geological environments should have similar mineral occurrences. In the same token, it assumes that similar mineral deposits occur in similar geological and tectonic environments. Clearly, this assumption is conceptually reasonable but practically difficult to verify.

In general, the principle of analogy is used in a statistical sense for some given conditions. For example, within the study area, similar controlling geological factors are considered as “necessary conditions” for the same type of deposit to occur. This relaxed assumption makes it possible to predict mineral targets based on the information collected from known deposits within the same study area.

Quality of the Favorability estimates are heavily reflective of the “quality” of control areas, that is, how good of a geological analogue the control area is of the unexplored region. When analogue and desired mineral potential estimate is for geological conditions similar to those that induced the exploration of control area, the estimates produced by a favorability model established on the control area may be unbiased and reliable. However, when geological references for the estimates differ or when the control area is not a good geologic analogue, favorability estimates tend to be biased and even totally meaningless (Pan 2018).

Characteristic Analysis

Characteristic analysis (CA), proposed by Botbol (1978), was demonstrated as a tool for quantitatively describing how typical each of a number of geological characteristics, such as mineralogy, geochemistry, and geophysics, is of a set of mineral deposits (Harris 1984; Pan and Harris 1992b). This statistical technique was later used as a tool for target selection in mineral exploration (McCammon et al. 1983). CA is perhaps the most primitive form of favorability analysis.

Like most other statistical models in mineral exploration, characteristic analysis is primarily constructed on observations or measurements taken in and around known ore deposits. Use of the model has been predicated upon the assumption that favorability to the mineral potential of a

region is a positive function of the similarity of its characteristics with those of the known deposits.

Let $\mathbf{Z} = (z_{ij})$ denote the matrix containing n observations on m characteristics (explanatory variables), where z_{ij} is either binary or ternary. In the binary form, $z_{ij} = 1$ if characteristic j exists in cell i and 0 otherwise. In the ternary form $z_{ij} = 1$ if its presence is favorable; $z_{ij} = -1$ if its presence is unfavorable; $z_{ij} = 0$ when it is absent or unevaluated. The characteristic values at the cell i and at the n cells are defined as

$$c_i = a_1 z_{i1} + a_2 z_{i2} + \cdots + a_m z_{im} \quad \text{or} \quad \mathbf{c} = \mathbf{Za} \quad (1)$$

where $\mathbf{a} = (a_1, \cdots, a_m)'$ is the unknown coefficient vector; $\mathbf{c} = (c_1, \cdots, c_n)'$ is the estimated characteristic vector of n cells.

The question now focuses on estimation of the weight parameters $\mathbf{a}$. Let us consider the matching (similarity) coefficient matrix $\mathbf{S}$ between pairs of m characteristics

$$\mathbf{S} = (s_{jk}) = \mathbf{n}^{-1} \mathbf{Z}' \mathbf{Z} \quad (2)$$

where $s_{jk} = \mathbf{n}^{-1} \sum_{i=1}^n z_{ij} z_{ik}$, $j, k = 1, 2, \cdots, m$.

Beginning with the similarity matrix $\mathbf{S}$ in Eq. (2), the weight for each of the m characteristics can be estimated by either square root method (SRM) or principal component method (PCM), which is discussed here for convenience. According to the theorem of spectrum decomposition for a matrix (Johnson and Wichern 1988; Li 2008; Pan and Harris 1992b), the matrix $\mathbf{S}$ can be decomposed as

$$\mathbf{S} = \lambda_1 \mathbf{a}_1 \mathbf{a}_1' + \cdots + \lambda_m \mathbf{a}_m \mathbf{a}_m' \quad (3)$$

where λ_j is the j^{th} eigenvalue of $\mathbf{S}$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$; $\mathbf{a}_j$ is the j^{th} eigenvector of $\mathbf{S}$ associated with λ_j .

Define $\phi_j = \lambda_j / \sum_{k=1}^m \lambda_k$, $j = 1, 2, \cdots, m$. Clearly, $1 \geq \phi_1 \geq \cdots \geq \phi_m \geq 0$, since matrix $\mathbf{S}$ is generally a non-negative definite matrix. If ϕ_1 is large enough (e.g., 0.8), matrix $\mathbf{S}$ can be approximately represented by its first eigenvector without losing much information about variability, that is,

$$\mathbf{S} \approx \lambda_1 \mathbf{a}_1 \mathbf{a}_1' \quad (4)$$

Therefore, it is reasonable to define coefficients in the characteristic model (1) as the first principal component: $\hat{\mathbf{a}} = \mathbf{a}_1 = (a_{11}, a_{21}, \cdots, a_{m1})'$. Consequently, the characteristic model is estimated by

$$\hat{\mathbf{c}} = \mathbf{Z} \hat{\mathbf{a}} \quad (5)$$

where $\mathbf{Z}$ is the $[n \times m]$ matrix containing observations of the m characteristics on n cells. The characteristic vector $\hat{\mathbf{c}}$ is often

interpreted as favorability indexes of the n cells to the mineral occurrence of interest.

A major advantage of CA is its simplicity, rendering it easily understood by geologists. Another appealing feature of the method is that it consists of quantitative procedures by which different modes of geological data pertinent to mineralization can be quantified, integrated, and analyzed. However, CA using the first principal component (FPC) as the weights of multiple characteristics is subject to ambiguity of its estimates, since FPC accounting for the largest variability among the characteristics does not necessarily have anything to do with the mineral occurrence of interest. Another major limitation is lack of statistical means that can be used to validate statistical significance of the characteristic estimates.

Favorability Analysis

As discussed early, characteristic analysis presents possible ambiguities in the meaning of favorability function, namely, favorability estimates may not carry sufficient information about the mineral occurrence of interest. This ambiguity is significantly reduced by incorporating target variables in the favorability modeling (Pan 1989; Pan and Harris 1992a). Canonical favorability model (CFM) constructed by maximizing the total covariance between favorability function and preselected target variables has explicitly linked the favorability estimates to the mineral potentials of interest (Pan 1993a).

Let $\mathbf{Z} = (Z_1, Z_2, \cdots, Z_m)$ be the vector of m explanatory variables. The generic form of favorability model is defined as

$$\mathbf{F} = \mathbf{Za} = a_1 \mathbf{Z}_1 + a_2 \mathbf{Z}_2 + \cdots + a_m \mathbf{Z}_m \quad (6)$$

where $\mathbf{F}$ is called the favorability function of explanatory variables $\mathbf{Z} = (Z_1, \cdots, Z_m)$; $\mathbf{a} = (a_1, \cdots, a_m)'$ is an unknown coefficient vector. A more generalized form is

$$\mathbf{F}^w = \mathbf{ZWa} = a_1 w_1 \mathbf{Z}_1 + a_2 w_2 \mathbf{Z}_2 + \cdots + a_m w_m \mathbf{Z}_m \quad (7)$$

where $\mathbf{F}^w$ is called the weighted favorability function; $\mathbf{W} = \text{diag}(w_1, \cdots, w_m)$ with w_j being a priori selected weight for $\mathbf{Z}_j$.

A priori information which should be independent of the information contained in the explanatory data is not always available. An appropriate description of such information requires a thorough understanding of geological features and their intrinsic relations to the mineralization processes.

The favorability functions in Eqs. (6) and (7) are meaningful only with respect to certain stated objectives pertinent to the target variables $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_t)$, which will be incorporated in the estimation later. In mineral exploration, $\mathbf{F}$ is interpreted as the favorable level to the mineralization

represented by the target variables given a collection of observations on explanatory variables.

Favorability function F in CFM is estimated such that it characterizes the maximum correlation between explanatory and target variables. In order to characterize the largest variability and removal of multicollinearity, the target variables are converted to their principal components using the covariance matrix of target variables.

Let $Y = (Y_1, Y_2, \dots, Y_t)$ be the vector of t target variables and its covariance matrix $\Sigma_{YY} = [\text{Cov}(Y_j, Y_k)]$. Let P_k be the k^{th} principal component of the target variables

$$P_k = Y u_k = u_{1k} Y_1 + u_{2k} Y_2 + \dots + u_{tk} Y_t, \quad k = 1, 2, \dots, t \quad (8)$$

where $u_k = (u_{1k}, u_{2k}, \dots, u_{tk})$ is the coefficients of the k^{th} principal component; $\text{Var}(P_k) = \lambda_k^2$ (the k^{th} largest eigenvalue of Σ_{YY}) and $\text{Cov}(P_j, P_k) = 0$ ($j \neq k$).

Suppose that the first q ($\leq t$) principal components account for a sufficiently large proportion of the total variability in the t target variables ($>80\%$). Then, the total variability in t target variables can be approximated by its q ($\leq t$) principal components corresponding to the first q largest eigenvalues, that is

$$P = (P_1, P_2, \dots, P_q) = YU \quad (9)$$

where $U = (u_1, u_2, \dots, u_q)$ be the $t \times q$ matrix containing the q eigenvectors associated with the first q largest eigenvalues of Σ_{YY} .

Consider the criterion to determine the coefficient vector a in Eq. (6) by maximizing the sum of the squared covariance between favorability function F and each of the q principal components P_k . Let

$$\phi^2(a) = \sum_{k=1}^q \text{Cov}^2(P_k, F) \quad (10)$$

with covariance function

$$\text{Cov}(P_k, F) = \text{Cov}(Y u_k, Z a) = u_k' \Sigma_{YZ} a, \quad k = 1, 2, \dots, q \quad (11)$$

where Σ_{YZ} is the covariance matrix between Z and Y . Thus, Eq. (10) is rewritten as

$$\phi^2(a) = \sum_{k=1}^q a' \Sigma_{YZ}' u_k u_k' \Sigma_{YZ} a = a' H a \quad (12)$$

where

$$H = \Sigma_{YZ}' U U' \Sigma_{YZ} \quad (13)$$

The variance of F is given by

$$\alpha^2(a) = \text{Var}(F) = a' \Sigma_{ZZ} a \quad (14)$$

where Σ_{ZZ} is the covariance matrix of explanatory variables Z . Using Eqs. (12) and (13) and Lagrange multiplier μ , construct the following objective function

$$L(a, \mu) = a' H a - \mu (a' \Sigma_{ZZ} a - 1) \quad (15)$$

The coefficients a are selected such that L is maximized. Taking the first-order partial derivatives of function L with respect to a and μ , and setting them to zero, we derive

$$H a = \mu \Sigma_{ZZ} a, \quad a' \Sigma_{ZZ} a = 1 \quad (16)$$

Obviously Σ_{ZZ} is required to be nonnegative definite such that $\Sigma_{ZZ} = \Sigma_{ZZ}^{1/2} \Sigma_{ZZ}^{1/2}$. Hence, Eq. (16) can be rewritten as

$$\Phi e = \mu e, \quad e' e = 1 \quad (17)$$

where $e = \Sigma_{ZZ}^{-1/2} a$ is the unit eigenvector associated with the largest eigenvalue of Φ

$$\Phi = \Sigma_{ZZ}^{-1/2} H \Sigma_{ZZ}^{-1/2} \quad (18)$$

Finally, the estimate of favorability function in Eq. (6) is given by

$$\hat{F} = Z \Sigma_{ZZ}^{-1/2} e \quad (19)$$

Similarly a priori weighted favorability function in Eq. (7) is estimated by

$$\hat{F}^w = Z W \Sigma_{ZZ}^{-1/2} e \quad (20)$$

Equations (6) and (7) are called population estimates of the favorability function. The sample version for a given sample of n observations on all explanatory and target variables can be estimated using the same canonical correlation approach above. Details can be found in Pan (1993a).

Interpretation of the favorability estimates can be aided by estimating various correlation coefficients between favorability function and explanatory and target variables as well as the principal components of target variables. Given the optimum estimate of a , correlation vectors between $\langle F, Z \rangle$, $\langle F, Y \rangle$, and $\langle Z, Y \rangle$ all can be estimated and used for validation of the favorability relations between target and explanatory vari-

ables (Pan 1993a). These correlation estimates provide a powerful tool to gain insights into intrinsic relationships between target and explanatory variables.

Several steps are required to implement the CFM in mineral potential mapping. The foremost important step is to identify and construct target variables and explanatory variables. The data sets or map layers are preprocessed, if necessary, to remove noise and to enhance important signals for the deposits of interest. A complete database for both target and explanatory variables can be constructed in the control area consisting of complete information. Such data will serve as the “training” information in the estimation of favorability function.

The favorability function established in the control area is then extrapolated into the entire study region outside the control area by using an equal-area grid cell approach. For convenience of interpretation, the favorability estimates can be regularized into the range $[0, 1]$ with 1 to be the highest favorability to the mineralization of interest and 0 the least favorable (Pan 1993a, b, c). The map of showing favorability patterns is superimposed by the maps of known deposits or mineral occurrences in the control area for cross validation. The targets can be delineated in the areas where favorability values are greater than some preselected cutoff. Using the estimates of correlation coefficients, each selected target can be evaluated and interpreted in terms of different combinations of explanatory and target variables. This can help eliminate spurious targets created by the observations of less accuracy. Fig. 1 shows an example of the targets selected by CFM in an epithermal gold-silver mineralization district.

Canonical favorability model (CFM) creates a unique, optimal, and flexible linear function for mineral target definition and delineation based on explanatory and target variables. It is optimal in the sense that favorability estimates are derived from maximization of the total correlation between favorability function and target variables. Note that the optimization is valid only when variables are not correlated in spatial locations. In reality, however, most geoscience variables are spatial and their valuations are relevant to geographical locations. Variables are not only correlated each other but also correlated themselves in different spatial locations. CFM, which only captures information on variable values, is not capable of capturing the spatial information of variables associated with geographical locations. As such, more sophisticated favorability methods, such as Indicator Favorability Analysis and Regionalized Favorability Analysis, were proposed by Pan (1993b, 1993c) to capture the information of spatial association in addition to the nonspatial correlations.

Concluding Remarks

An advantage of quantitative integration by favorability analysis comparing to subjective approach is to create unique measures in selecting exploration targets. A favorability map by CFM, for example, shows clear statistical patterns and values that highlight favorable degrees for mineral occurrences. Areas containing the highest favorability estimates are retained for further exploration. Conversely, areas with the lowest favorability values are removed from further consideration. The outcomes of favorability analysis serve as useful gauge for ranking of promising areas, from which top targets will receive the top priority in further exploration efforts.

In a broader sense, successful modern mineral exploration brings together multiple disciplines and bodies of knowledge. The discovery and development of an economic ore deposit require objective and scientific decisions in each of the development stages. The following are considered the primary components necessary for a successful exploration program:

- Geoscience theories to identify favorable areas and to guide the acquisition of incremental information, for example, geology, geochemistry, geophysics, and remote sensing.
- Spatial database management systems to optimally organize the exploration information of diverse forms and to efficiently support daily exploration activities.
- Mathematical tools that calculate or estimate favorability or probability values to describe the degrees of mineral potential which aids the selection of top targets.
- Further exploration or drilling investments to verify the target areas selected in the previous step and to work toward new discoveries.
- Economic decision analysis to perform financial and risk analysis and to ensure appropriate resource development, if a discovery is made.

The quality of favorability analysis largely depends on the quality of the input information. Garbage-in will always result in garbage-out. The early quantitative models, although important demonstrations, were less than impressive geologically because of their limited content of geoscience information. Although these models contained many quantified geological variables, their collective information about the genesis, concentration, and preservation of mineral deposits was meager at best. That objective could not be achieved solely by using more sophisticated and powerful “number crunching” methods. The quantified geology was anemic in information about the formation and preservation of mineral deposits (Harris and Pan 1990; Pan and Harris 1993). The notion of Information Synthesis was initially developed to improve the power of quantitative data integration (Pan and Harris, 2000). The favorability analysis like CFM was an

effort toward the goal of increasing information in mineral potential mapping.

Traditional favorability modeling has been conducted using regular inter-grids or cells as the sampling scheme and estimation unit. The “cell” approach is associated with several drawbacks. The most significant problem is that geological processes can be reconstructed through observable geoscience features, which are measurable in geological units, not artificial cells. The cell-based measurements tend to distort the intrinsic relations between geological features and mineralization descriptors. The cell-based quantification of geological features, spatially correlated and even connected, is difficult to capture essential genetic factors that played key roles of metal enrichment. Moreover, the cell-approach easily ignores exceptional conditions for formation of large deposits, which cannot be readily quantified through grids.

As such, the notion of Intrinsic Geological Unit (IGU) or Intrinsic Sample (IS) was proposed as a sampling framework on which favorability analysis for mineral potential mapping is conducted (Pan 1989, 2010, 2018). This domain-based sample reference ensures the use of consistent geological areas as the units of quantification and interpretation. Thus, instead of estimating statistical relations on the quantified geology and discovered resources of cells, they would be estimated on the characteristics of intrinsic samples. Furthermore, the IGU methodology also makes it convenient to incorporate more geoscience information from the third dimension (depth) in mineral potential mapping.

Bibliography

- Botbol JM, Sinding-Larsen R, McCammon RB, Gott GB (1978) A regionalized multivariate approach to target selection in geochemical exploration. *Econ Geol* 73:534–546
- Harris DP (1984) Mineral resources appraisal—mineral endowment, resources, and potential supply: concepts, methods, and cases. Oxford University Press, New York, 455p
- Harris DP, Pan GC (1990) Updated concepts of intrinsic samples and methodology for simultaneous estimation of discovered resources and endowments: report on research sponsored by U.S. Geol Survey, 300p
- Johnson RA, Wichern, DW (1988) Applied multivariate statistical analysis. Prentice Hall, 607p
- Li WD (2008) Applied multivariate statistical analysis. Beijing University Press, 255p
- McCammon RB, Botbol JM, Sinding-Larsen R, Bowen RW (1983) Characteristic analysis –1981: final program and a possible discovery. *Math Geol* 15:59–83
- Pan GC (1989) Concepts and methods of multivariate information synthesis for mineral resources estimation, Ph.D. dissertation, University of Arizona, Tucson, 302 p
- Pan GC (1993a) Canonical favorability model for data integration and mineral potential mapping. *Comput Geosci* 19:1077–1100
- Pan GC (1993b) Indicator favorability theory for mineral potential mapping. *Nonrenewable Res* 2:292–311
- Pan GC (1993c) Regionalized favorability theory for information synthesis in mineral exploration. *Math Geol* 25:603–631
- Pan GC (2010) Critical issues in mineral resources appraisal. *Geol Bull China* 29:1413–1429
- Pan GC (2018) General framework of quantitative target selections: Handbook of Mathematical Geosciences, Chapter 21 (Invited Contribution to IAMG’50)
- Pan GC, Harris DP (1992a) Estimating a favorability equation for the integration of geodata and selection of mineral exploration targets. *Math Geol* 24:177–202
- Pan GC, Harris DP (1992b) Decomposed and weighted characteristic analysis for the quantitative estimation of mineral resources. *Math Geol* 24:177–202
- Pan GC, Harris DP (1993) Delineation of intrinsic geological units. *Math Geol* 25:9–39
- Pan GC, Harris DP (2000) Information synthesis for mineral exploration. Oxford University Press, 450p

Fisher, Ronald A.

Frits Agterberg

Central Canada Division, Geomathematics Section,
Geological Survey of Canada, Ottawa, ON, Canada

Biography

Sir Ronald Aylmer Fisher (1890–1962) was born in London, England. Many consider him to be the most important mathematical statistician of the twentieth century. In 1909, he was awarded a scholarship to study at Cambridge University where he obtained a First in mathematics in 1912. From 1919 to 1933, Fisher worked at the Rothamsted Experimental Station developing and applying statistical techniques in agriculture and other fields of science. Next, he became the head of the Department of Eugenics at University College, London. In 1937, he helped P.C Mahalanobis establish the Indian Statistical Institute. From 1940 to 1956, Fisher was a professor at Cambridge University. After his retirement in 1957, he went to Adelaide, Australia, where he was senior research fellow of the Commonwealth Scientific and Industrial Research Organisation (CSIRO) until 1962.

His life is described in detail by his daughter Joan Fisher Box (1978). Until Fisher began his career in Rothamsted, the emerging discipline of mathematical statistics was dominated by Karl Pearson who introduced *inter alia* the Pearson correlation coefficient and the chi-square test for goodness of fit in contingency tables. With respect to these two inventions, Fisher (1915) derived the formula of the correlation coefficient in a two-dimensional normal distribution, and he obtained the correct formula for the number of degrees of freedom to be used in goodness of fit tests (Fisher 1922). Later, he developed many other methods often related to analysis of variance. His first book “*Statistical Methods for Research Workers*” (Fisher 1925) went through 14 editions. Later books included “*The Design of Experiments*” (Fisher

1954). He dominated mathematical statistics worldwide during the first half of the twentieth century.

Fisher had a profound interest in geology. In Fisher (1954), he suggested that geology started out on a quantitative path with Lyell (1833) who wrote *Principles of Geology*. This book contained a 60-page long statistical appendix documenting the subdivision of the Tertiary into stages such as Oligocene and Pleistocene on the basis of their contents of fossil species still living today. According to Fisher, this statistical appendix was omitted from later editions of the book, because during the next 100 years geology evolved as a descriptive, nonmathematical science. Lyell's original idea later became a corner stone of quantitative biostratigraphy. It also fits in with Darwin's later theory of evolution. In genetics, Fisher extensively used mathematical modeling to further develop Mendelian genetics.

In "Dispersion on a Sphere," Fisher (1953) describes a method to treat directions of remanent magnetization in basaltic lava samples originally collected by a graduate student at Cambridge University. The resulting "Fisher distribution" has formed the basis of statistical analysis of measurements in paleomagnetism. In invited talks near the end of his career, Fisher advocated Wegener's original theory of continental drift postulating that paleomagnetic data indicate "large movements of crustal blocks." This was about 10 years before geoscientists generally accepted this idea.

Bibliography

- Box F (1978) R.A. Fisher – the life of a scientist. Wiley, New York
 Fisher RA (1915) Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 10:507–521
 Fisher RA (1922) On the interpretation of χ^2 from contingency tables, and the calculation of P . *J Roy Stat Soc* 85:87–94
 Fisher RA (1925) Statistical methods for research workers. Oliver and Boyd, Edinburgh
 Fisher RA (1953) Dispersion on a sphere. *Proc R Soc A* 217:295–305
 Fisher RA (1954) The design of experiments. Oliver and Boyd, Edinburgh
 Lyell C (1833) Principles of geology. Murray, London

Flow in Porous Media

Daniele Pedretti

Dipartimento di Scienze della Terra "A. Desio", Università degli Studi di Milano, Milan, Italy

Synonyms

Fluid dynamics in porous media

Definition

Flow in porous media is the movement of fluids through the connected pores of a porous material. In the field of geosciences, flow in porous media is typically associated to the movement of groundwater, oil, gas, and other fluids naturally or artificially occurring across geological formations. For instance, aquifers are porous reservoirs in which groundwater flows. Several laws have been derived to quantify flow in porous media, starting from Darcy's law in 1856. Multiple analytical and numerical solutions have been developed to resolve the partial differential equations describing flow in porous media. Other processes, such as mass (solute) and energy (heat) transport, as well as biogeochemical reactions occurring in the subsurface, are directly connected to flow in porous media. Flow in fractured media can be treated as a special case of flow in porous media. Quantifying flow in porous media is notoriously complicated by the limited accessibility to the subsurface, by the lack of data and by the scale dependency of properties and parameters. Uncertainty must be accounted for when dealing with flow in porous formations.

Principles

The concept of flow in porous media is widely utilized in many fields of geosciences. Some well-known applications include: groundwater applications, such as aquifer balance calculations, groundwater well extractions, or managed aquifer recharge; oil and gas extraction from geological reservoirs; aquifer contamination analysis, including transport of reactive biogeochemical species; consolidation and subsidence; saltwater intrusion and salination; storage of energy (heat/cold) in aquifers; CO₂ sequestration in deep geological formations; and nuclear waste storage. The following text addresses mainly groundwater flow in porous media, recognizing, however, that most of the concepts hereby presented can be still applied for other fluids.

Similar to heat and electrical current flows, fluid flow through porous media is a mechanical process, which involves a potential for flow through porous media, i.e., a mechanical energy per unit mass of fluid. A porous material is composed of two parts: the solid matter (or matrix) and the void (or pore) space. The pore space can be occupied by a single fluid phase (e.g., water, air) or by two or more coexisting phases. Total porosity (ϕ) can be defined as:

$$\phi = \frac{V_v}{V_T} \quad (1)$$

where V_v is the void volume and V_T is the total volume of a unit of soil or rock. Total porosity can range from near 0% to

more than 60% depending on the type of geological material, its genetic features, and the physical and chemical processes (e.g., weathering, faulting processes, consolidation, or chemical dissolution) that may have modified the void space with time. Domenico and Schwartz (1990) adopted the terms “the primary porosity” for interstitial porosity and the “secondary porosity” for fracture or solution porosity.

When a pressure difference occurs between two points of a fluid-invaded porous medium, such fluid can flow through the connected pores. The principles of flow mechanics are essentially the same for all fluids, although the scales at which this flow occurs depend on the property of the medium (e.g., permeability), of the fluid (e.g., viscosity and density), and the relative saturation of each fluid in the pore space for multiphase flow.

The first fundamental law describing fluid flow in porous media is Darcy’s law (Darcy 1856). It was first presented by the French engineer Henry Darcy to quantify the results of sand column experiments developed for the water filtration systems of the city of the Dijon in France. Darcy’s law is an empirical law that states that the specific discharge of a flow tube, q (i.e., the volumetric discharge Q divided by a cross section A of the flow tube) is directly proportional to the difference in pressure (P) at the two edges of a flow tube, i.e.:

$$q = \frac{Q}{A} = -\frac{k}{\mu} \nabla P \quad (2)$$

where k is the permeability of the medium and μ is the viscosity of the fluid. The minus sign indicates that flow occurs from higher to lower potential. Darcy’s law is also often expressed in terms of difference in hydraulic heads (h), i.e.:

$$q = -K \nabla h \quad (3)$$

The term K is the hydraulic conductivity of the medium, representing a macroscopic parameter that lumps together the microscopic behavior of the fluid moving in the porous media. As the gradient is dimensionless, the units of q and K are the same (a length over a time, i.e., a velocity). From (3), a pore velocity v can be calculated as:

$$v = q / \phi_e \quad (4)$$

where ϕ_e is the effective porosity, defined as the part of the porosity that is available for the fluid flow.

Darcy’s law is valid for slow, viscous flow approaching the laminar flow conditions (Reynold’s number close to one). While Darcy’s law was originally conceived as an empirical law, it can be rigorously derived from a microscopic description of fluid flow. For instance, it can be derived from the Navier-Stokes equations under the assumption of stationary, creeping, and incompressible flow, with viscous resisting

forces being linear with the velocity. Modifications to the input parameters render Darcy’s law still valid for multiphase flow analysis, such as in the case of water flow in the vadose zone or oil-water systems. A relative permeability factor (k_r) is introduced to scale the saturated hydraulic conductivity K as a function of the phase saturation.

While Darcy’s law has been demonstrated to hold true for a vast number of applications, there are cases when it does not apply. Mitchell and Soga (2005) described at least three scenarios by which nonlinearity between flow velocity and gradient occurs: (1) non-Newtonian water flow properties, (2) particle migrations that cause blocking and unblocking of flow passages, and (3) local consolidation and swelling that is inevitable when hydraulic gradients are applied across a compressible soil.

The interested reader may refer to several text books and monographs (e.g., Bear 1972; De Marsily 1986; Freeze and Cherry 1979) for more information on the topic.

Quantifying Flow in Porous Media

Quantifying flow in porous media can be attained through multiple direct and indirect methods (see, e.g., Delleur 2010).

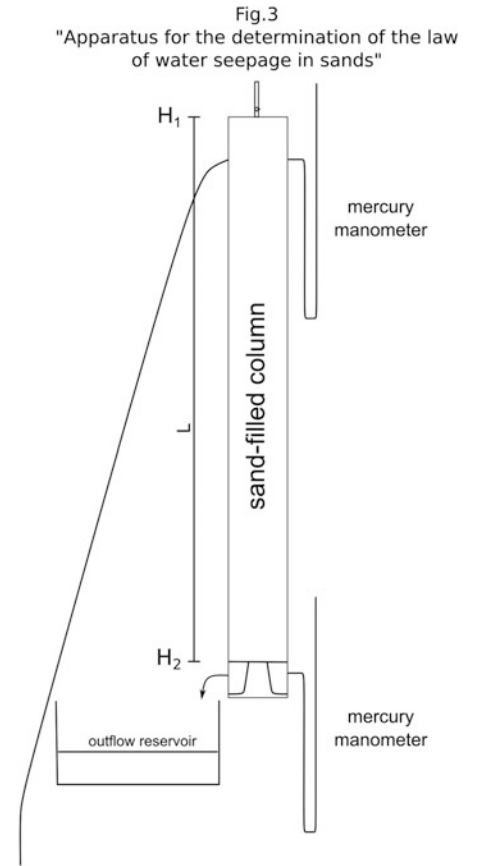
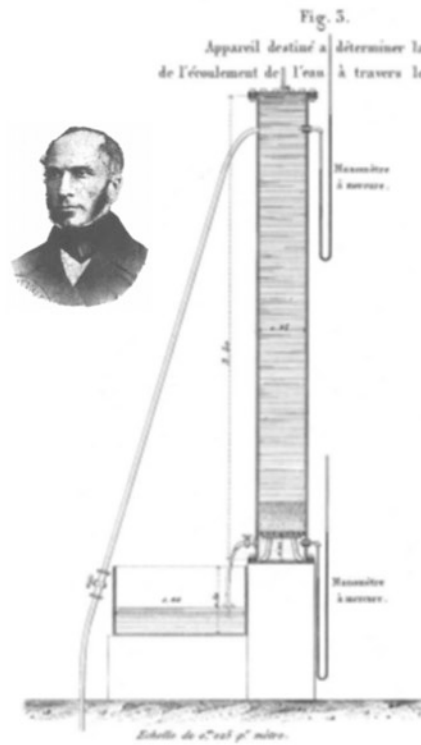
Direct approaches include: borehole flowmeters, which measure horizontal flow velocity and direction of groundwater flow in wells or piezometers; seepage meters and infiltrometers for measuring groundwater–surface water exchange; solute-based methods such as the borehole dilution test. In karst systems, flow rates can be measured in sinking streams, in cave streams, and in springs. Other methods have been revised, for instance, by Essouayed et al. (2019).

Indirect approaches include the estimation of soil properties to be used in flow models. Such properties can be obtained from laboratory scale experiments (constant or falling head permeameters, which are essentially similar to Darcy’s apparatus – Fig. 1) or through field-scale tests, such as slug and pumping tests. Scale dependency of estimated parameters (e.g., K) is connected to the heterogeneous nature of the subsurface. The importance of a specific scale depends on the type of problem requiring flow estimation. Large-scale tests such as pumping tests can integrate multiple heterogeneities, resulting in an averaged response of the system, and are often preferred by hydrogeologists in practical applications. The principle of pumping well tests is simple. Water is pumped from a well, which is untapping the aquifer. The discharge of the well and the changes in water levels are measured with time in the well and in piezometers close to the pumping well at known distances from the well.

Mathematical models have been developed to support the quantification and prediction of flow in porous media as well as connected quantities (e.g., transport). Flow in porous media is described by partial differential equations (pdes)

Flow in Porous Media,

Fig. 1 On the left, Henry Darcy's portrait and the sand column experimental apparatus, as originally reported in Darcy's 1856 monograph "Les fontaines publiques de la ville de Dijon." On the right, the apparatus scheme, illustrating the difference in hydraulic head ($H_1 - H_2$) at the two edges of the column of length L .



requiring initial and boundary conditions as well as a set of input parameters to be resolved. Analytical or numerical methods can be applied to resolve these pdes. An example of a flow equation for porous media, obtained combining Darcy's law with the continuity equation, is:

$$S_s \frac{\partial h}{\partial t} = \frac{\partial}{\partial x_i} \left(K \frac{\partial h}{\partial x_j} \right) + W \quad (5)$$

where t is time, S_s is the specific storage coefficient ($S_s = S/b$, with S being the storativity or the volume of water released from storage per unit decline in H per unit area of the aquifer, and b the aquifer thickness), K is the second-order tensor of the hydraulic conductivity ($K = K_{ij}$), and W is the sink/source term describing the inflow or outflow volumetric flux per unit volume. In Eq. 5, the system is oriented according to Cartesian coordinates. Alternative coordinate systems can be also adopted. Using radial coordinates, for instance, the flow equation can be written as:

$$S_s \frac{\partial h}{\partial t} = \frac{1}{r} \frac{\partial}{\partial r} \left(rK \frac{\partial h}{\partial r} \right) + W \quad (6)$$

where r is the radial distance. This form is particularly useful when flow gradients are forced by a pumping well.

Analytical solutions provide exact solutions to the governing equation and are often relatively simple and efficient to use. This makes analytical solutions particularly useful for certain applications (e.g., risk assessment) requiring quick analysis without the need of sophisticated codes. For instance, certain solutions of the flow equation are programmable in simple spreadsheets. The problem with analytical solutions is that sometimes the properties and boundaries of the flow system are unrealistically idealized to enable obtaining a closed-form solution of the pde.

The Theis (1935) solution provides an example of a widely adopted analytical model for the assessment of flow in porous media. Recognizing the analogy between heat and flow governing equations, Theis presented the first analytical solution of the transient behavior of groundwater head levels in the proximity of a pumping well. The Theis solution can be written as:

$$h = h_0 + \frac{Q_w}{4\pi T} W(u) \quad (7)$$

where h_0 is the initial head, Q_w is the pumping well discharge rate, T is the aquifer transmissivity ($T = Kb$) and $W(u)$ is the exponential integral

$$W(u) = -0.577216 - \ln(u) + u - \frac{u^2}{2 \times 2!} + \frac{u^3}{3 \times 3!} + \frac{u^4}{4 \times 4!} + \dots \quad (8)$$

with

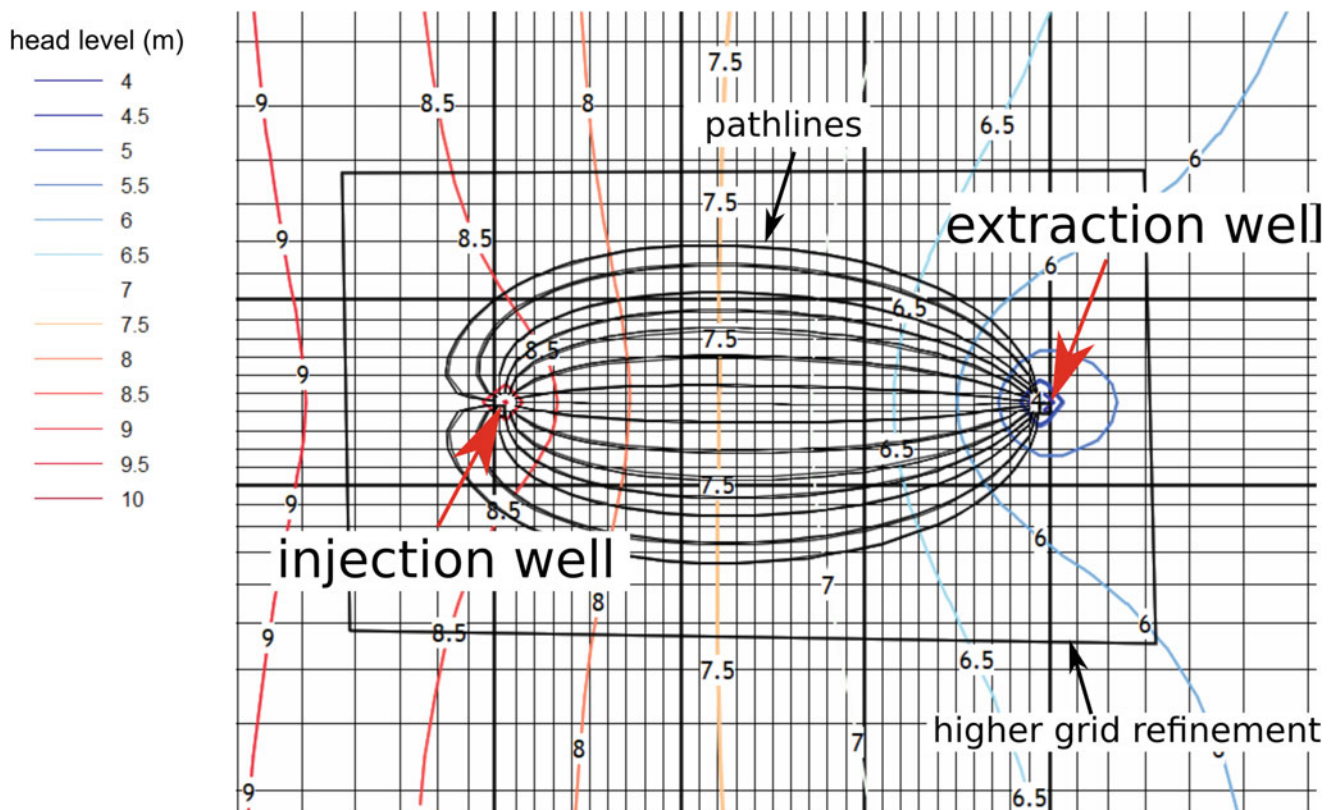
$$u = \frac{r^2 S}{4Tt} \quad (9)$$

where r is the radial distance from the well where the head change is observed.

Numerical flow models have the advantage that complex boundary conditions and properties can be simulated, such that the flow system is more realistically reproduced. With numerical models, continuous variables are replaced with discrete variables that are defined throughout a spatial grid or mesh. The continuous pde defining the space-time change of h (e.g., Eq. 5) is replaced by a finite number of algebraic equations that defines h at specific points. This

system of algebraic equations generally is solved using matrix techniques. Popular numerical flow models includes MODFLOW (<https://www.usgs.gov/mision-areas/water-resources/science/modflow-and-related-programs>), the open-source groundwater flow model distributed by the U.S. Geological Survey and in use since the end of the 1980s (McDonald and Harbaugh 2003). A variety of graphical user interfaces support modelers with the model implementation. Anderson et al. (2015) provides an excellent reading on advanced groundwater numerical modeling.

An example of a numerical flow code applied to a typical groundwater problem is shown in Fig. 2. A dipole (or doublet) well test configuration is simulated under idealized homogeneous aquifer conditions. This is a widely adopted configuration in many modern practical problems involving flow in porous media, including aquifer thermal energy storage or managed aquifer recharge. A regional groundwater flow occurs from left to right in the figure. The upstream well injects water with a recharge rate Q ; the downstream well extracts water with a discharge rate $-Q$. The problem was resolved using the code MODFLOW-2005 (Harbaugh 2005). To assess the resulting geometry of the streamlines, and particularly to check if the injected water is collected by the



Flow in Porous Media, Fig. 2 An application of a numerical model to resolve a typical problem in hydrogeology involving a dipole (or doublet) well configuration. The code MODFLOW-2005 and

MODPATH were adopted to respectively solve for the groundwater flow system and the particle advection from the injection to the extraction well. The rectangle indicates the zone with higher grid refinement

extracting wells, particle tracking is adopted using the code MODPATH (Pollock 1994). The result of this simple modeling exercise suggests that the downstream well is able to collect the injected fluids.

Open Questions and Unresolved Issues

Research on flow in porous media has advanced dramatically in the last decades. Dealing with quantitative flow analysis in porous media remains however challenging. While some questions have remained open and unresolved for decades, new issues have arisen, particularly those related to new societal challenges (e.g., climate change or environmental pollution) and to technological innovation (e.g., the emergence of supercomputers). A short list of currently debated issues regarding flow in porous media includes:

- Uncertainty analysis. The limited amount of (hydro)geological data characterizing the subsurface, which can be highly heterogeneous, renders flow model outcomes ubiquitously uncertain. Stochastic analysis provides a suitable framework to introduce uncertainty analysis in flow modeling of porous media, casting the model outcomes in probabilistic terms. Statistics (e.g., mean and variance) of desired properties (e.g., h) are used for uncertainty quantification and related applications, including risk assessment (Tartakovsky 2013).
- Scale dependency of flow model properties and parameters. Scaling is the process of relating a property's value on one scale to its value on another scale. Flow measurements in porous media provide properties and parameters that depend on the supporting scale of the measuring approach (Hunt and Sahimi 2017; Pedretti et al. 2016). Such scaling effects are also connected to geological heterogeneity and lack of information regarding the soil structures, in which flow occurs. Finding solutions that resolve the change of model parameters across scales, and particularly the concept of “upscaling” (i.e. the use of coarser-scale homogenized models able to effectively reproduce processes occurring at smaller scales) remains an open question for flow in porous media.
- Combining flow and biogeochemical reactions. Physical and biogeochemical processes are strongly related in porous media. Reactive transport modeling constitutes a major research area in many fields of geosciences, particularly due to the complexity of combining multiple nonlinearly coupled processes with computationally efficient solving tools (Steefel et al. 2014). Reactive transport modeling is needed for a variety of applications, spanning from ecohydrological modeling, in which water, carbon, and nutrients are coupled in the so-called soil–plant atmosphere continuum, to the prediction of chemically unstable radionuclides transport or to acid mine drainage under isothermal or non-isothermal conditions.
- Climate change. Little is still known about the effects of climate change on flow dynamics in porous media. Forecasting of changes in the major climatic variables and accurate estimation of groundwater recharge are two important aspects to consider. Estimating recharge is considered one of the most complicated tasks when dealing with flow in geological media.

Conclusion

Flow in porous media is studied in many areas, from biological sciences to the textile industry. In geosciences, it is typically referred to as the fluid dynamics of geological systems. The occurrence and the magnitude of flow in geological systems is directly controlled by the type of soils, reservoirs, and deposits, and the type of moving fluids. Since 1856, it is possible to quantify flow in porous media through a set of equations first obtained by a French engineer called Henry Darcy, after whom a fundamental law has been named. Research is currently very active to find increasingly efficient methods to measure and predict flow in porous media. Uncertainty analysis must be emphasized as an essential tool to make predictions regarding flow in porous media and related issues (e.g., contaminant transport).

Bibliography

- Anderson MP, Woessner WW, Hunt RJ (2015) Applied groundwater modeling: simulation of flow and advective transport, 2nd edn. Elsevier Science B. V., Amsterdam
- Bear J (1972) Dynamics of fluids in porous media. Elsevier, New York
- Darcy H (1856) Les fontaines publiques de la ville de Dijon: exposition et application. ... Victor Dalmont
- De Marsily G (1986) Quantitative hydrogeology. Academic, Orlando
- Delleur JW (2010) The handbook of groundwater engineering. CRC Press
- Domenico P, Schwartz FW (1990) Physical and chemical hydrogeology. Wiley
- Essouayed E, Annable MD, Momtbrun M, Atteia O (2019) An innovative tool for groundwater velocity measurement compared with other tools in laboratory and field tests. J Hydrol X 2:100008. <https://doi.org/10.1016/j.hydroa.2018.100008>
- Freeze RA, Cherry JA (1979) Groundwater. Prentice-Hall, Englewood Cliffs
- Harbaugh AW (2005) MODFLOW-2005: the U.S. Geological Survey modular ground-water model—the ground-water flow process. Tech Methods
- Hunt A, Sahimi M (2017) Flow, transport, and reaction in porous media: percolation scaling, critical-path analysis, and effective medium approximation. Rev Geophys 55:993–1078. <https://doi.org/10.1002/2017RG000558>
- McDonald MG, Harbaugh AW (2003) The history of MODFLOW. Groundwater 41:280–283. <https://doi.org/10.1111/j.1745-6584.2003.tb02591.x>

- Mitchell JK, Soga K (2005) Fundamentals of soil behavior. Wiley, New York
- Pedretti D, Russian A, Sanchez-Vila X, Dentz M (2016) Scale dependence of the hydraulic properties of a fractured aquifer estimated using transfer functions. *Water Resour Res* 52:5008–5024. <https://doi.org/10.1002/2016WR018660>
- Pollock DW (1994) User's guide for: MODPATH/MODPATH-PLOT, version 3: a particle tracking post-processing package for MODFLOW, the U.S. Geological Survey finite-difference ground-water flow model, Documentation of MODPATH. U.S. Geological Survey Open-File Report 94-464, 6 ch
- Steefel CI, Appelo CAJ, Arora B, Jacques D, Kalbacher T, Kolditz O, Lagneau V, Lichtner PC, Mayer KU, Meeussen JCL, Molins S, Moulton D, Shao H, Šimůnek J, Spycher N, Yabusaki SB, Yeh GT (2014) Reactive transport codes for subsurface environmental simulation. *Comput Geosci* 19:445–478. <https://doi.org/10.1007/s10596-014-9443-x>
- Tartakovsky DM (2013) Assessment and management of risk in subsurface hydrology: a review and perspective. *Adv Water Resour* 51:247–260. <https://doi.org/10.1016/j.advwatres.2012.04.007>
- Theis CV (1935) The relation between the lowering of the piezometric surface and the rate and duration of discharge of a well using ground-water storage. *EOS Trans Am Geophys Union* 16: 519–524

Forward and Inverse Stratigraphic Models

Cedric M. Griffiths
 Predictive Geoscience, StrataMod Pty. Ltd., Greenmount,
 WA, Australia
 Department of Exploration Geophysics, Curtin University,
 Perth, Australia

Synonyms

Optimization; Simulation

Definition

A stratigraphic model is a conceptual, physical or mathematical representation of the succession of sedimentary layers beneath the surface of the Earth. Historically such models were static representations of the shape, location, and content of existing layers. Such models were initially built through observation of outcrops, drill holes, and excavations, and have been supplemented since the 1920s by remote petro- and geophysical measurements.

Stratigraphic models have formed the basis not only of academic studies, but a significant proportion of the global economy since pre-Roman times, with rock quarrying, gold,

copper, and lead mining, and more recently hydrocarbon extraction and carbon sequestration.

“Forward” and “inverse” models can be both intuitive and formal, mathematical, ways of examining and testing the inductive and deductive reasoning behind stratigraphic model building.

The traditional “inductive” process of geological interpretation of an observed stratigraphic succession can be seen as a process of “inverse” modeling. A geologist intuitively uses understanding based on training and experience to relate the process that formed the stratal succession to the stratigraphic observations. To a large extent this depends on experience and background, rarely quantified and thus highly subjective.

Forward modeling of stratigraphic processes can be seen as a “deductive” means to quantitatively examine and test this intuitive process understanding by modeling the depositional, diagenetic and tectonic processes forward in geological time and comparing the results with observations. This comparison can be a subjective matching of the model with well or outcrop data, or through a numerical inverse model with a defined error or objective function to be minimized. A forward model may be physical, using a flume-tank, or numerical, using computer implementations of the equations and algorithms representing the physical processes of erosion, transport, sedimentation, burial, diagenesis, and tectonic modification of the stratal succession.

An “Inverse model” can mean somewhat different things mathematically, geophysically, and geologically.

Mathematically the “inverse problem” is the process of solving for “y” given:

$$x = \mathbf{M}y$$

where $\mathbf{M}$ is a matrix of functions relating known values “x” to the unknown values “y”.

Geophysically “x” could be a series of surface measurements of gravity, then if “r” is the distance from the surface measurement point to the subsurface body. “y” the matrix of subsurface rock densities and “G” the universal gravitational constant, “ $\mathbf{M}$ ” is the matrix of G/r^2 for each possible combination of measurement and distance to the subsurface body.

If “ $\mathbf{M}$ ” is a square matrix, with no zero eigenvalues, and is not overdetermined, then this linear inverse problem could be solved exactly by inverting the matrix “ $\mathbf{M}$ ”.

Generally, however, this is not possible in practice and the problem must be solved iteratively using additional prior information and/or by minimizing the error between calculated values of “x” and the actual observations. This process involves a “Forward Model” as discussed below.

The mathematical “Forward Model” consists of an expression of a relationship. This could be a simple equation:

$$x = Fy$$

solving for x given a known function F and known values “ y ”.

Geophysically a forward model example could be the production of a simple synthetic seismic trace by convolving an acoustic source wave with a stratigraphic model of layered acoustic impedances.

Introduction

By modeling a system we can test our understanding of its fundamental properties. Such modeling is often called simulation. Simulation is a form of model building, where sufficient features of the system or object being modeled are included to enable predictions to be made (either qualitatively or quantitatively) concerning the behavior of its real-life analogue. Simulation can be physical, for example flume-tank experiments, mathematical (where analytical solutions exist for predictive equations), algorithmic/computational (where numerical solutions exist), or heuristic/computational (heuristic = “rule-of-thumb”), where knowledge exists in the form of general “understanding” and logical rules rather than equations.

Modeling may be either “forward” or “inverse”. In the forward model we start with what we hope are the major processes acting on a system, wave direction and amplitude, for example, and some boundary conditions, such as beach grain-size distribution, water depth, and beach slope. We let the model run for an appropriate length of geological time and look at the change in the system. If both the process model and the boundary conditions were specified sufficiently, then the end result will be a prediction that can be tested at an appropriate scale. In reality, there will always be higher resolution features that were not predicted by the simulation as a result of “aliasing” at the fixed time intervals and fixed grid dimensions used (see Shannon/Nyquist sampling theory), and we have to hope that these are not significant for our proposed use. Note that although recent simulations use unstructured grids and variable time intervals there will still be spatiotemporal features that are below the simulation resolution. An important point to remember about any modeling is that a model is always a simplification in some aspect. The skill of the modeler is to ensure that the model is neither too simple nor too complex to answer the question being asked. Another way of looking at this is to say that if you do not know what question you are asking, then no model will ever satisfy your requirements. There is no such thing as a generally applicable model.

Forward Stratigraphic Modeling

Numerical forward modeling of any type tests our understanding of a system. To paraphrase Lord Kelvin, “if you can’t put numbers on it – you don’t understand it”, In the case of weather forecasting for example we take the physics of water vapor, air mass movement, radiation, convection, on a variety of scales and test our understanding of the system by making predictions (forecasts) into the future – by hours, days, or months. These predictions are tested each time we go for a walk or plan a holiday. Thus forward stratigraphic modeling can also have several goals:

- Quantitatively investigate our current understanding of a particular depositional/stratigraphic process
- Predict useful rock properties away from observations (between wells, below seismic resolution, etc.)
- Test multiple sedimentological working hypotheses
- Form part of an inverse model iteration

Interest in the forward modeling of sediment depositional processes was boosted by the need to model and simulate beach conditions for landing military vehicles during and after the Second World War. Krumbein provides a glimpse of this work in his 1964 paper on process-response models of “beach phenomena” including the modeling of soft and hard sand areas. His “Technical Memorandum No. 3 for the Beach Erosion Board of the US Corps of Engineers” provided equations for beach energy and grain-size relationships (Krumbein 1947).

Krumbein’s work provided an overview of the types of variables and interrelationships which could be included in a model of beach processes. Bates (1953) (quoted in Harbaugh and Bonham-Carter 1970, p. 458) proposed a hydraulic jet model for delta formation which formed the pre-cursor to Bonham-Carter and Harbaugh’s models of the late 1960s and early 1970s. Culling (1960) looked at erosion and transport in the subaerial environment using a potential-gradient approach (the higher the potential energy slope at a given point, the higher the erosion rate).

Sloss (1962) discussed the use of a simple “hole-filling” type of model in which the shape of a stratigraphic body “ s ” was a function of sediment input “ Q ”, rate of subsidence “ R ”, erosion of deposited sediment “ D ”, and the nature of the material supplied “ M ”. The model could simulate the effects of simple rigid-slab subsidence which increased basinwards from a null-point or hinge-line.

The coal extraction industry in Kansas funded work at the University of Kansas and the Kansas Geological Survey during the 1960s (under Dan Merriam) where many stratigraphic modelers cut their teeth on the Carboniferous cyclothems. Schwarzscher (1966) used differential equations

relating sedimentation rate to water depth and subsidence (quoted in Harbaugh and Bonham-Carter 1970, p. 463). Harbaugh (1966) looked at simulating carbonate ecology and sedimentation using an Algol program. In Britain, King (1968) developed a heuristic and stochastic forward model of shingle-spit formation (quoted in Harbaugh and Bonham-Carter 1970, p. 461). Oertel and Walton (1967) developed a conceptual model that included feedback to give a cyclic component to deltaic sedimentation.

Bonham-Carter and Sutherland (1967, 1968) developed Bates' ideas further with computer programs for jet-flow, assuming that as the river enters a larger body of water the flow velocity distribution is as a "plane jet" (Tetzlaff 1987, p. 17). Empirical functions are used to estimate the distribution of different sediment grain sizes. Each grain trajectory is considered independently. The model "realistically" represents deltaic lobes, levees, and mouth bars (Tetzlaff 1987, p. 20).

Harbaugh and Bonham-Carter (1970) further developed Sloss' geometric space-filling model with some refinements. They included sediment-load dependent subsidence and a time-lag that succeeded in modeling sigmoidal chronosomes (bodies of rock bounded by isochronous surfaces) with interfingering grain-size. The model worked in an algorithmic way where the basin floor was divided into a two-dimensional series of columns and sediment was passed into the basin from one end and allowed to accumulate (or not) on any particular column depending on the available accommodation (distance between wave base and previous column top).

Harbaugh and Bonham-Carter (1970) summarized the developments to date and presented examples of different types of flow models (Chaps. 6 through 9). This is one of the most useful introductions to the subject.

Since the 1980s there has been an increase in the number and variety of forward stratigraphic modeling computer programs as discussed in Paola (2000) and Burgess (2012), and the decade to 2021 has seen a rapid increase in the number and variety of practical applications.

Inverse Stratigraphic Modeling

Inverse modeling takes the given, observed, state of a system and tries to derive both its initial state and the processes acting to produce the observed state. In simple physical cases, where a single linear process has acted on a few variables, this may be possible. As an example, a rough estimate of the initial height above a given point of a steel ball rolling down an inclined plane may be derived by measuring the velocity of the ball at that point. If we wanted to predict the initial height to several significant digits however, then we would have to include information about surface roughness, friction, material properties, etc. Note also that in this case we cannot use an

inversion approach to predict horizontal distance of the measurement point from the release point. In the inverse modeling case we have to understand the processes involved and their relation to the question that we wish to ask.

Both forward and inverse models have their weak points. A given present-day state can be derived from an initial set of conditions via several different routes, or a different route combined with different initial conditions can result in an identical present-day state. In other words it is relatively easy to prove sufficiency but almost impossible to prove necessity. A sedimentological example of this is the "upward-coarsening unit," where it is relatively easy to show that a prograding marine unit **can** have produced that grain-size distribution, but impossible to show that **only** a prograding marine unit could have produced it. Both forward and inverse modeling of natural systems suffer from the phenomena of "non-uniqueness." This is the property of being able to approach the same result by an infinite number of different paths. Burton et al. (1987) emphasized the need for prior information in any inverse scheme to constrain the solutions. Today the more-or-less global acceptance of a defensible eustatic curve from the Devonian has provided a valuable (and testable) prior.

One of the most challenging issues in inverse stratigraphic modeling is the choice of an "objective/cost/loss/error function" to be minimized by the inverse scheme. The objective/error function to be minimized should represent the aim of the simulation. For example if the simulation is being run to predict the presence of porous rocks in an unexplored part of a basin, then an appropriate objective function could be the difference between modeled and predicted porosity distribution at a known well location. If there is a large error between the predicted and observed porosity distribution, then the model is wrong in some way and the forward model parameters will have to be changed and the model re-run. A typical (and poorly understood by users) problem here is the difference in **volumetric** resolution of the model and the constraint or test data. Both the difference in volume sampled and the relative measurement/prediction errors need to be understood. All physical measurements sample a volume, and all stratigraphic models predict volumes, but if there is an order of magnitude difference between the two volumes, it is necessary to adjust the objective function accordingly.

In its simplest form, inverse stratigraphic modeling can be achieved by running a group of forward models with systematic changes to the input parameters, using the human eye to select the best model from a visual inspection of the results, then systematically changing the input parameters to and repeating the process. This was described in Wijns et al. (2004):

Underlying most forward modelling exercises in geosciences is an implicit inverse question. A formal inverse approach to

geological modelling has only recently started to develop due mostly to two reasons: first, standard geological modelling is extremely computationally intensive, and second, it is hard to develop numerical cost functions, to drive the inverse search, that can capture enough geological knowledge to make the search meaningful. Creating a cost function is difficult because it involves distilling into a single value an assessment of the geometries, volumes, and positions, as well as determining how geologically reasonable a solution may be.

We employ interactive inversion to tackle the second problem. The concept is of such simplicity as to make it appear as little more than a novelty: replace the numerical evaluation of solution quality with a subjective value provided by an expert user. Underneath the simplicity there is a very powerful concept. We replace what in most cases is an artificial numerical function with the computational power of an expert brain. This provides not only the best 'geological processing power' we currently have, but also intuition and expertise, in the form of theoretical knowledge, experience, and a priori information.

Intuitive as the above approach is, it is also slow, and with large, complex real-world stratigraphic models a computer-based convergence method is needed.

Numerical stratigraphic inverse modeling has a shorter history than forward modeling, which in part reflects its difficulty and reliance on very fast processing power to achieve solutions. Some of the first approaches have been discussed by Lessenger, Cross, and Lerche in various papers in the 1980s; Guérillo, Bornholdt, Nordlund, and Duan in the 1990s; and Duan, Charvin, and Falivene since 2000.

The (fractal-like) incompleteness of the stratigraphic record (Plotnick 1986) at any given location also complicates stratigraphic inversion. Erosion and non-deposition are part of the sedimentation process, but are represented by surfaces rather than sediment bodies. Typically, at any given location, the amount of "missing" geological time exceeds the time represented by sediment.

Forward Stratigraphic Models

Mathematical and computational forward stratigraphic models have evolved in many diverse schools since the 1950s. A good introduction to the basic principles behind these various approaches can be found in Granjeon (1996).

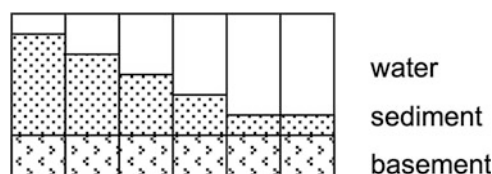
The approaches can be divided into broad subcategories, from the simplest to the more complex as follows.

Geometrical Models

Models that do not describe the process itself, but the geometric results of the process, usually filling or partially filling accommodation space. Many of these models are based on pioneering work by Sloss (1962). He developed a simplistic model for the relationship between sediment shape (**S**), volume (**Q**) of sediment supplied, rate of subsidence (**R**), rate of dispersal of sediment (**D**), and the "nature of the material supplied" (**M**). He developed models for situations where

the relationship between input sediment volume and subsidence rate varied, showing that both regressive and transgressive phases and cyclical patterns of lithosomes (sensu Wheeler and Mallory (1956)) could be generated by sediment supply variation. Sloss's subsidence term implicitly rather than explicitly included relative sea level change (as a uniform subsidence). Sloss discussed the use of the model in passive margin and foreland settings, and also discussed turbidite development. His treatment was on a purely logical, conceptual basis. These ideas were taken up by Harbaugh and Bonham-Carter in Chap. 9 of their 1970 publication, which provides a mathematical treatment of Sloss's conceptual model. The following is taken from this chapter.

A section through the basin topography is divided into a series of two-dimensional columns Δx wide which provide the basis for sediment accumulation.



The basic premise is very straightforward. Sediment enters the section from a predetermined point. Part of the sediment load is deposited on the first column encountered and the remainder is passed to the next column. The amount deposited on a particular column depends on (a) the amount of sediment available for deposition, and (b) the accommodation space in that column. Mass is conserved by accounting for all sediment as it moves throughout the 2D section. The accounting system is very simple. Let the sediment load (mass) entering the basin be L and let the proportion of this load that is deposited on the first column be k . Thus the amount deposited is kL and the remaining load that is transferred to the next column is $L - kL$, or $L(1 - k)$. This can be generalized in that the amount deposited n columns from the start point ($n \cdot \Delta x$ distance units away) is $kL(1 - k)^{n-1}$. The sediment mass entering this column is $L(1 - k)^{n-1}$. The value k relates to the transmissibility of sediment in the direction of mass transport and must be specified beforehand. Different values of k can be used for different grain-size fractions. K is not a physical parameter but is heuristic in nature (Harbaugh and Bonham-Carter 1970, pp. 377–379). The value k can also be interpreted as slope of a particular sediment type. Different sediments may have different k values.

Changes in accommodation space due to relative sea level fluctuation, subsidence, and compaction are handled heuristically by constraining this simple algorithm. Harbaugh and Bonham-Carter (1970, p. 379) discuss three constraints, although more are possible and will be discussed later. The three situations discussed are:

- Where the sediment-water interface is above or equal to base level (as specified for a given particle-size class), then deposition is not allowed and all remaining load is passed to the next cell.
- Where the sediment-water interface is only slightly below base level only part of the sediment load is deposited as a proportion of $kL(1 - k)^{n-1}$. Sediment is base-leveled and the remainder passed on to the next column.
- The sediment-water interface is well below base-level for all particle-size fractions and $kL(1 - k)^{n-1}$ is deposited.

Over the years additional constraints have been added by many packages, such as angle of repose for various grain-size mixtures, slope failure, etc. Subsidence can be either an instantaneous or lagged (occurring after a defined time interval) function of sediment loading or external tectonics. There are in practice few limits to the algorithmic complexity that can be built in to this type of heuristic model. A FORTRAN IV program to carry out this type of modeling is provided by Harbaugh and Bonham-Carter (1970, pp. 382–384), and this program formed the precursor to many of today's geometric models. Quite complex geometries and sedimentological situations can be built in. The relation to physical principles is however tenuous to say the least, and such models function best as tools where full access to the source code is necessary for the programmer-geologist to test ideas expressed as “if-statements” in the code. As a general “off-the-shelf” modeling package they are only of use if the heuristics used are explicitly stated and agree to one's prior assumptions about the way sediments behave in any particular circumstance.

Topography-Control (Diffusion, or Potential-Gradient) Models

Models in which the transport and deposition of sediment is a function of the potential gradient on the preexisting (and continually changing) sediment mass/surface rather than the nature of the preceding cell as is the case in a geometric model. In the following discussion mass-concentration and topographic height (energy potential) are treated as equivalent and interchangeable. In other words a heap of sediment that will slide down-hill (diffuse) is the same as a concentration of sodium ions that will diffuse throughout a block of porous sandstone from a high potential to a low potential. Sediment can be considered to follow the same rules, that is, it is difficult to get sediment to climb uphill.

These models are based on solutions to Darcy flow or Ohms Law type equations: Fick's first law of diffusion states (after Harbaugh and Bonham-Carter 1970, p. 239):

$$J_x = -k\delta_c/\delta_x$$

where:

J_x is the mass flux.

k is the diffusion coefficient.

c is the concentration of material.

x is the coordinate direction parallel to flow and perpendicular to the reference surface.

In its three-dimensional, time-dependent form, derived by combining the first law with an equation of continuity (conservation of mass) such as

“sediment volume in minus sediment volume out equals accumulation of sediment”

For zero accumulation the continuity equation can be simplified to (after Harbaugh and Bonham-Carter 1970, p. 216)

$$\frac{\delta V_x}{\delta x} + \frac{\delta V_y}{\delta y} + \frac{\delta V_z}{\delta z} = 0$$

where V is the volume of an incompressible fluid over time interval t .

In the form of a mass-flux equation this becomes:

$$\frac{\delta J_x}{\delta x} + \frac{\delta J_y}{\delta y} + \frac{\delta J_z}{\delta z} = \frac{\delta c}{\delta t}$$

(Harbaugh and Bonham-Carter 1970, p. 239).

Combining a continuity equation and Fick's first law gives Fick's second law:

$$k \left[\frac{\delta^2 c}{\delta x^2} + \frac{\delta^2 c}{\delta y^2} + \frac{\delta^2 c}{\delta z^2} \right] = \frac{\delta c}{\delta t}$$

where k is an isotropic diffusion coefficient (the same in all directions).

This equation can be modified to give diffusion coefficients that vary with direction, with depth (Rivenaes 1992), with location, and with time.

In its two-dimensional form the equation appears as:

$$k \left[\frac{\delta^2 c}{\delta x^2} \right] = \frac{\delta c}{\delta t}$$

where c is the topographic elevation and x is the distance from an arbitrary coordinate origin.

This is saying that the change in height (elevation) with time at any particular point in a section is a function of a constant times the change in slope at that point. In its finite difference form this equation becomes (after Harbaugh and Bonham-Carter 1970, p. 241, Eq. 6–49):

$$c_{t+1,j} = c_{t,j} + k \left(\frac{\Delta t}{(\Delta x)^2} \right) (c_{t,j+1} + c_{t,j-1} + c_{t,j})$$

which is simply saying that the increase in height at any given location after a time interval Δt is a function of the height (potential) difference between that location and another location Δx away at the previous moment in time. The product of the constant of proportionality “ k ” (the diffusion coefficient) and the time step/column width term:

$$k \left(\frac{\Delta t}{(\Delta x)^2} \right)$$

is constrained to lie within the range of 0.0–0.5. A value zero gives a diffusion rate of zero and there is no change of slope with time (Harbaugh and Bonham-Carter 1970, p. 244). In this form the equation functions in a similar way to the geometric model which adds sediment according to heuristics which replace the single coefficient.

Paola in Angevine et al. (1990, pp. 97–101) discusses the derivation of the linear diffusion equation and the assumptions involved, specifically in the case of fluvial systems.

One practical issue with diffusion-based multi-grain forward models at all scales is the choice of “diffusion coefficients” that have sedimentological meaning beyond the simulation process. A realistic basin-scale simulation requires the definition of many diffusion coefficients for the various geological processes involved, and there are very few studies that relate these values to real measurable sedimentological or environmental processes. The coefficients can be chosen to provide a model result that matches observations to a specified degree, but inverting this result to derive insight into the geological processes is not straightforward.

Hydraulic Process-Response Models

These are models in which fundamental laws concerning hydraulic flow are used in three dimensions to model sediment transport and deposition in three dimensions. Most are based on solutions to simplified versions of the Navier-Stokes equations describing flow in three dimensions for an isotropic Newtonian fluid. The book by Tetzlaff and Harbaugh (1989) provides a useful introduction to this type of modeling. The following passage is taken from Tetzlaff and Harbaugh (1989, p. 14).

“The Navier-Stokes equations combine the continuity equation (Eq. 1) and the momentum equation (Eq. 2) to provide a mathematical description of flow for an isotropic Newtonian fluid. An isotropic Newtonian fluid is a fluid whose properties are uniform in all directions and that follows Newton’s laws of motion.

The continuity equation incorporates the conservation of mass:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \rho \mathbf{q} = 0 \quad (1)$$

where

ρ = fluid density

t = time

$\mathbf{q}$ = flow velocity vector

whereas the momentum equation describes changes in momentum by the fluid:

$$\rho \left(\frac{\partial \mathbf{q}}{\partial t} + (\mathbf{q} \cdot \nabla) \mathbf{q} \right) = -\nabla p + \nabla \cdot \mu \mathbf{U} + \rho(\mathbf{g} + \Omega \mathbf{q}) \quad (2)$$

where

p = pressure

μ = fluid viscosity

$\mathbf{U}$ = Navier-Stokes tensor

$$= \begin{bmatrix} 2\frac{\partial u}{\partial x} - \frac{2}{3}\nabla \cdot \mathbf{q} & \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} & \frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} \\ \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} & 2\frac{\partial v}{\partial y} - \frac{2}{3}\nabla \cdot \mathbf{q} & \frac{\partial v}{\partial z} + \frac{\partial w}{\partial y} \\ \frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} & \frac{\partial v}{\partial z} + \frac{\partial w}{\partial y} & 2\frac{\partial w}{\partial z} - \frac{2}{3}\nabla \cdot \mathbf{q} \end{bmatrix}$$

u, v, w = components of flow velocity $\mathbf{q}$ in the direction of x, y , and z axes, respectively,

Ω = Coriolis tensor

$$= \begin{bmatrix} 0 & 2\omega \sin \phi & -2\omega \sin \phi \\ -2\omega \sin \phi & 0 & 0 \\ -2\omega \cos \phi & 0 & 0 \end{bmatrix}$$

ω = angular rotational velocity of the earth

ϕ = latitude

$\mathbf{g}$ = gravitational acceleration vector

If the fluid is homogeneous, incompressible, and remains at constant temperature, the density ρ and viscosity μ can be considered constant”.

Although the Coriolis acceleration $\Omega \mathbf{q}$ is so small in most deltaic simulations that it can be ignored, marine currents on an oceanic scale are strongly affected by the Coriolis acceleration, which should therefore be included for basin-scale models.

Sediment transport and deposition within this flow regime is another, separate, problem. One solution uses a bed-load transport equation that transports sediment as a function of bed roughness, flow depth, and flow rate.

Comparison Between Various Equations for Bed-Load Transport

Author	Formula
DuBoys (1879)	$Q_s = B_1 n^4 \frac{Q^4}{h^{2/3}}$
Meyer-Peter and Muller (1948)	$Q_s = B_2 n^3 \frac{Q^4}{h}$
Shields (1936)	$Q_s = B_3 \frac{n^4}{d} \frac{Q^4}{h^{2/3}}$
SedSim (1980)	$Q_s = B_4 \frac{n^2}{h^{1/3}} Q^4$
Q_s = sediment discharge rate	B_i = constant
n = Manning's roughness coefficient	d = sediment particle diameter
Q = water flow rate = $ Q $	h = flow depth

Some of the more successful 3D hydraulic process-response models use a combined Eulerian/Lagrangian “Marker-in-Cell” approach (Harlow 1964) that allows intra-cell fluid vectors to be used to simulate 2.5 dimension free-surface flow under many flow conditions.

Recent models include Delft3D (Lesser et al. 2004; Flores 2011) and Badlands (Salles and Hardiman 2016).

Mixed Algorithmic, Heuristic, and Fuzzy-Logic Models

In this case a pragmatic/heuristic mixture of model is used under different circumstances. For example, large-scale clastic features may be filled geometrically/by diffusion, while carbonates use a partial process model, and coal is deposited heuristically. The Shell model “StrataGem” operated in this way (Lawrence et al. 1987). FUZZIM is a model that uses fuzzy-sets and fuzzy-rules to model three-dimensional stratigraphy (Nordlund 1999a, b). Other modeling approaches are described by Ritchie et al. (1999), and Lessenger (1993), Lessenger and Cross (1996), Lessenger and Lerche (1999). Duan has modeled carbonates (Duan et al. 1998) as has Bosence et al. (1998).

Inverse Stratigraphic Models

As stated previously, the process of geological study is one of inference or inversion, the attempt to understand the earth's history through intuitive or numerical process understanding.

The sediments are a sort of epic poem of the earth. When we are wise enough, perhaps we can read in them all of past history. (Carson 1951)

The stratigraphical record is a lot of holes tied together with sediment. (Ager 1973)

Our best record of past conditions on Earth, for most of its history, comes from strata in sedimentary basins. Clues to this history are housed in the composition, spatial organization, chemistry, and fossils of strata. The record allows us to interpret the Earth system response to past episodes of climate change. These interpretations are not only important for characterizing Earth's past but also aid our ability to predict responses to ongoing climate change. To read Earth's epic poem, though, we

must solve one of the most complicated inverse problems known to science (Straub et al. 2020).

Stratigraphic inverse problems are nonlinear and generally not solvable analytically.

In general terms, properly formulating and solving a stratigraphic inverse problem demands that one has:

1. A very good understanding of the physical processes and how they may be simplified, e.g. at its simplest, a sedimentary basin consists of accommodation filled with sediment at a certain rate, from certain locations over a certain time period
2. Good experimental strategy and quality measurements
3. Most importantly, effective computational techniques

Without a good understanding of the physical processes, basically nothing can be done. Quality measurement data are essential because they will decide the quality of the solution of the inverse problem. This includes not only the accuracy of the observational or experimental data but also the precise knowledge of the characteristics of the measurement data in terms of the noise content (noise level, frequency, etc.).

Stratigraphic observational data are problematic in this context. Outcrop data exist because the strata at that location are somehow atypical (more resistant to weathering/faulted/etc.) – the rest have been eroded away. Aerial/satellite imagery are snapshots of a current process, possibly transient and atypical for that transient environment. Well/core data are sparse in most parts of the world and the well locations are often not selected to represent the average strata but some atypical feature of economic interest.

To numerically solve the inverse problem in stratigraphy, a numerical forward model is necessary. As discussed in the section on forward stratigraphic modeling, a wide range of numerical forward model types are available. However, due to the iterative nature of inverse modeling there are specific requirements of a forward model that are necessary as part of an inverse scheme:

For inverse modeling purposes the forward stratigraphic model should:

- Make predictions that are directly comparable to real, measurable, stratigraphic data
- Simulate those aspects of stratigraphy that are of interest in a specific study (e.g., “location of offlap-break over time,” or “porosity of Permian sands in an undrilled part of the basin,” or “timing and character of carbonate deposition in the basin”)
- Simulate necessary and sufficient processes for the depositional system being investigated
- Be capable of implementation on a digital computer and have as short a run time as possible
- Have an appropriate spatial and temporal resolution

The outputs from the forward model should be independent and as sensitive as possible to the system parameters to be inversely identified. For example, marine stratal patterns evolve in space as a function of the relationship between accommodation and sediment supply. Accommodation can change as a function of initial topography, tectonic subsidence, sediment accumulation, sea level change, isostatic response, etc. The inverse model could attempt to identify all or some of these variables, or simply invert for change in accommodation in space and time.

The simplistic basic requirements of several inverse modeling schemes are as outlined in the pseudo-code below (from Alazzam and Lewis (2013)):

```

Define objective function (g)
Initialize the values for all parameters: m, t
Generate (m) feasible solutions from the search space using a forward model
    Evaluate the fitness from the objective function (g)
    Optimal solution = the best solution.

i=1
Do while i < t,
    Generate m/2 solutions from the neighbourhood of the best solution
    Generate m/2 solutions from the search space
    Find the best solution from the (m) generated solutions

    If the best solution is better than the optimal solution
        Optimal solution=best solution
    End If
End DO

```

The various components of this scheme have changed over the past 40 years, which is an indication of the difficult nature of the problem. Taking each component in turn:

Define objective function (**g**)

g – the objective function – is a “fit for purpose” measure of success in matching the model with some aspect of reality. It is often expressed as a numerical distance to approach zero in a successful simulation, but can be a visual comparison between a complex model and reality (see Wijns et al. 2004) or the shape of a sedimentary feature (Zhang et al. 2021). Stratigraphically it could be the thickness and modal grain sizes of the strata at one or more known well locations.

Initialize the values for all parameters: **m**, **t**

Generate (**m**) feasible solutions from the search space using a forward model

This step involves running the forward model using either a selection of values sampled using a variety of approaches

from the permissible ranges for each variable in the forward model, or a subset of the variables with others being pre-determined (e.g., an assumed eustatic curve and a sediment compaction rate).

In a “genetic” inverse model these variable values may be coded as a binary string (Bornholdt et al. 1999) since mutations on a binary string affect each order of magnitude with equal probability.

In a Bayesian Monte Carlo Markov Chain (MCMC) approach (Charvin et al. (2009a, b), Gilks et al. (1996)):

the total problem for inversion is to understand the uncertainty on this optimal, or best, model, or distribution of all possible models. Both aspects require some quantitative measure of how well a given model can reproduce the observed data. This is known as the misfit or likelihood function, the latter of which can be thought of as a measure of the probability of reproducing the observed data, given a model with a particular set of parameter values. This can be written as $p(D_{obs} | X)$, where $a | b$ means a given, or conditional on, b .

The aim of the inversion scheme is to construct a reliable representation of the posterior probability distribution $p(D_{obs} | X)$. Given this, it is then straightforward to extract individual models (e.g. the best or expected model), to construct probability distributions for individual model parameters, and to examine correlations between parameters. One widely used method to achieve these goals is Markov chain Monte Carlo (MCMC), a stochastic sampling-based approach. MCMC algorithms aim to construct a random walk around the model space (i.e. a Markov chain) which has the posterior probability distribution as its limiting stationary distribution. Useful introductions to this methodology are given in Gilks et al. (1996) and Denison et al. (2002). In a few words, MCMC operates by generating a large number of samples from the model space, according to well-defined rules, and the distribution of those samples provides a good approximation of the underlying posterior distribution. Charvin et al. (2009a, b)

Surrogate models may be used to reduce the computational cost associated with uncertainty propagation of the most relevant inputs to a selected range of outputs (Gervais et al. 2018).

Evaluate the fitness from the objective function (**g**)

Where “fitness” is the measure of the best “match” to the objective function or the minimum distance to data. This could be a simple regression r^2 value in the case of a series of thicknesses at a well, or an N-dimensional matrix.

In the case of string matches at well locations (Duan 2017) has shown that the Levenshtein distance (Levenshtein 1966) is useful from the point of view of sedimentology and stratigraphy “where comparison of sedimentary facies successions should account for the following aspects: (1) facies types and their division in sections, including special facies such as erosional surfaces; (2) the thickness of each identified and divided facies (maybe repeated) in sections; and (3) the vertical order or sequence of the coded facies in each

sedimentary facies succession.” The Levenshtein distance is also capable of allowing for facies changes according to Walther’s Law (facies in neighboring depositional environments tend to be adjacent in vertical succession’).

The choice of objective function in inverse modeling is critical. There is a tendency to use mathematical solutions taken from non-stratigraphic studies rather than those that can have direct “stratigraphic fitness” such as conforming to Walther’s Law or chromosome dimensions.

Optimal solution = the best solution.

The best combination of input variables as defined by the fitness is then used as a good measure for the local optimal solution and it is also initially set as the best known solution for the next iteration.

Do while $i < t$,
Generate $m/2$ solutions from the neighbourhood of the best solution

In the next iteration, take the best solution and use variable values close to those from the fittest to form $m/2$ new simulations

Generate $m/2$ solutions from the search space

Use variable **values from the entire variable set** to form $m/2$ new simulations to add to the fittest.
Calculate the fitness for all m simulations.

Find the best solution from the (m) generated solutions

If the best solution is better than the optimal solution
Optimal solution=best solution
End If
End DO

Lessenger and Lerche (1999) discuss the strategy and tactics of inverse model construction in stratigraphy.

Current Investigations, Controversies, and Gaps in Current Knowledge

The above Forward Stratigraphic Modeling categories are to some extent arbitrary, and there is a good deal of overlap in practice in many operational programs. The difference between the models is becoming blurred as commercial pressure forces developers to include features that cannot be handled by one approach alone. There remains however room for all types of model at appropriate scales. For studying the entrainment, transport, distribution, and deposition of clastics in locations such as beaches, harbors, and estuaries over short time intervals and at resolutions of a few meters then hydraulic process models are of interest, such as Delft3D and Sedsim. For pipe-line scour studies, platform and

loading-quay effects on sediment build-up, and pollutant dispersal including heavy minerals from drilling operations, and cases where basin/channel topography is complex and the problem is three-dimensional in nature and short-term, then Delft3D may be appropriate.

A new generation of parallelized, unstructured grid models such as Badlands, and those based on Python code are now available (Salles and Hardiman 2016; Salles et al. 2018a, b).

Limitations to Present Models

Explicit numerical forward stratigraphic modeling of cohesive clays and flocculation is limited to a few high resolution/short time interval studies at present, and given the influence of flocculated clays on channel migration in deltaic environments this could be an area of interest in future.

The explicit inclusion of vegetation in forward stratigraphic models is also a potential future area of interest as dune stabilization by vegetation is a major preservation mechanism (Salles et al. 2011), and the role of mangroves and seagrass as sedimentation and erosion controls in tropical littoral environments needs to be explicitly modeled as climate changes.

The use of unstructured grids and massive parallelization of code will see larger and more comprehensive forward and inverse models evolve over the next decade.

Summary

The ready availability of relatively cheap fast computers with graphic processing units (GPU’s) that lend themselves to massively parallel processing, combined with simulation code that is open-source and readily available, will hopefully lead to most undergraduate stratigraphy courses being based around forward and inverse stratigraphic modeling exercises where three-dimensional stratigraphic hypotheses can be numerically tested in real-time and the consequences of sea level change and the coastal response to changed wave and current environments will be more widely understood. A stratigraphy forecast will become as common as the weather forecast (but hopefully more reliable. . .).

Geology is on the cusp of a revolution in stratigraphic education and prediction that is reaching maturity after a 40-year gestation period.

Bibliography

- Ager DV (1973) The nature of the Stratigraphical record. Wiley, p 114
Alazzam A, Lewis HW (2013) A new optimization algorithm for combinatorial problems. *Int J Adv Res Artif Intell* 2(5):63–68

- Alt H, Godau M (1995) Computing the Fréchet distance between two polygonal curves. *Int J Comput Geom Appl* 5(1–2):75–91
- Angevine CL, Heller PL, Paola C (1990) Quantitative sedimentary basin modeling. Continuing Education Course Note Series 32 AAPG, Tulsa
- Baker MR (1988) Quantitative interpretation of geological and geophysical well data, Ph.D. Dissertation, The University of Texas at El Paso, El Paso, Texas, p 145
- Bates CC (1953) Rational theory of delta formation. *AAPG* 37: 2119–2162
- Blum J, Dobranszky G, Eymard R, Masson R (2006) Identification of a stratigraphic model with seismic constraints. *Inverse Prob* 22(4): 1207–1226
- Bonham-Carter GF, Harbaugh JW (1968) Simulation of Geologic Systems, an overview. Computer contribution 22, Kansas Geological Survey
- Bonham-Carter GF, Sutherland AJ (1967) Diffusion and settling of sediments at river mouths: a computer simulation model. *Gulf Coast Assoc Geol Soc Trans* 17:326–338
- Bonham-Carter GF, Sutherland AJ (1968) Mathematical model and Fortran IV program for computer simulation of deltaic sedimentation, Computer Contribution 24, Kansas Geological Survey
- Bornholdt S, Westphal H (1998) Automation of stratigraphic simulations: quasi-backward modelling using genetic algorithms. In: Mascle A, Puigdefabregas C, Luterbacher HP, Fernandez M (eds) *Cenozoic Foreland Basins of Western Europe*, vol 134. Geological Society Special Publications, pp 371–379
- Bornholdt S, Nordlund U, Westphal H (1999) Inverse stratigraphic modelling using genetic algorithms. In: Harbaugh JW, Watney WL, Rankey EC, Slingerland RL, Goldstein RH, Franseen EK (eds) *Numerical experiments in stratigraphy: recent advances in stratigraphic and sedimentologic computer simulations*, vol 62. SEPM Society for Sedimentary Geology, pp 85–90. <https://doi.org/10.2110/pec.99.62.0069>
- Boschetti F, Moresi L (2001) Interactive inversion in geosciences. *Geophysics* 66(4):1226–1234
- Boschetti F, Wijns C, Moresi L (2002) Effective exploration and visualisation of geological parameter space. *Geochem Geophys Geosyst* 4:10
- Bovy B, Braun J, Cordonnier G, Lange R, Yuan X (2020) The FastScape software stack: reusable tools for landscape evolution modelling. In EGU General Assembly Conference Abstracts
- Burgess PM (2012) A brief review of developments in stratigraphic forward modelling, 2000–2009. In: *Regional geology and tectonics: principles of geologic analysis*. Elsevier, Amsterdam/Boston
- Burton R, Kendall CGStC, Lerche I (1987) Out of our depth: on the impossibility of fathoming eustasy from the stratigraphic record. *Earth Sci Rev* 4:237–277
- Cancès C, Granjeon D, Peton N, Tran QH, Wolf S (2017) Numerical scheme for a stratigraphic model with erosion constraint and non-linear gravity flux. *FVCA 8 International Conference on Finite Volumes for Complex Applications VIII*, Jun 2017, Lille, France. pp 327–335
- Carson R (1951) *The sea around us*. Oxford University Press, New York
- Charvin K, Hampson GL, Gallagher K, Labourdette R (2009a) A Bayesian approach to inverse modelling of stratigraphy, part 1. *Basin Res* 21:5–25
- Charvin K, Hampson GL, Gallagher K, Labourdette R (2009b) A Bayesian approach to inverse modelling of stratigraphy, part 2: validation and sensitivity tests. *Basin Res* 21:27–45
- Craig RG (1979) A simulation model of landform erosion. Phd The Pennsylvania State University
- Cross TA (1988) Controls on coal distribution in transgressive-regressive cycles, Upper Cretaceous, Western Interior, U.S.A. In: Wilgus CK et al (eds) *Sea-level changes – an integrated approach*, vol 42. Society of Economic Paleontologists and Mineralogists Special Publication, pp 371–380
- Cross TA, Lessenger MA (1997) Sediment volume partitioning: rationale for stratigraphic model evaluation and high-resolution stratigraphic correlation. In: Sandvik KO et al (eds) *Predictive high-resolution stratigraphic sequence stratigraphy*
- Cross TA, Lessenger MA (1999) Construction and application of a stratigraphic inverse model. In: Harbaugh JW, Watney WL, Rankey EC, Slingerland RL, Goldstein RH, Franseen EK (eds) *Numerical experiments in stratigraphy: recent advances in stratigraphic and Sedimentologic computer simulations*, vol 62. SEPM Society for Sedimentary Geology, pp 69–84
- Culling WEH (1960) Analytical theory of erosion. *J Geol* 68:3
- Denison DGT, Holmes CC, Mallick BK, Smith AFM (2002) *Bayesian methods for nonlinear classification and regression*. Wiley, Chichester
- Duan T (2017) Similarity measure of sedimentary successions and its application in inverse stratigraphic modeling. *Pet Sci* 14:484–492
- Duan T, Griffiths CM, Cross TA (1998) Adaptive stratigraphic forward modeling: making forward modeling adapt to conditional data. *AAPG Annual convention and exhibition*, Salt Lake City
- Duan T, Griffiths CM, Johnsen SO (1999) Conditional simulation of 2-D parasequences in shallow marine depositional systems by using attributed controlled grammar. *Comput Geosci* 25(6):667–681
- Duan T, Cross TA, Lessenger MA (2001) Reservoir- and exploration-scale stratigraphic prediction using a 3-D inverse carbonate model. *AAPG annual convention*, Denver
- Flores JS (2011) Process-based modelling of the Brent delta: influence of paleobathymetry from the Oseberg Fm. pinch out on the wave dominated Brent delta progradation. North Sea Norwegian sector – Huldra field. Delft University of Technology, MSc thesis
- Gervais V, Ducros M, Granjeon D (2018) Probability maps of reservoir presence and sensitivity analysis in stratigraphic forward modeling. *AAPG Bull* 102(4):613–628
- Gilks WR, Roberts GO, George EI (1992) Adaptive direction sampling. *Journal of the Royal Statistical Society. Series D (the statistician)* 43, 1, special issue: conference on practical Bayesian. *Statistics* 1992(3):179–189
- Gilks WR, Richardson S, Spiegelhalter DJ (1996) *Markov chain Monte Carlo in practice*. Chapman & Hall, London
- Granjeon D (1996) *Modélisation stratigraphique déterministe: Conception et applications d'un modèle diffusif 3D multilithologique.. Stratigraphie*. Université Rennes 1, PhD 1996. Français. Access through HAL 2011 Id: tel-00648827 <https://tel.archives-ouvertes.fr/tel-00648827>
- Griffiths CM, Duan T (2000) Quantitative comparison of observed stratigraphy and that predicted from forward modelling. In: *Proceedings of the 31st IGC, Rio de Janeiro, August 2000*
- Griffiths CM, Duan T, Mitchell A (1996) How to know when you get it right: a solution to the section comparison problem in forward modelling. In: *Proceedings numerical experiments in sedimentology conference*, 15–17 May 1996. University of Kansas
- Harbaugh JW (1966) Mathematical simulation of marine sedimentation with IBM 7090/7094 computers. Computer Contribution 1, Kansas Geological Survey
- Harbaugh JW, Bonham-Carter GF (1970) Computer simulation in geology. Wiley Interscience, p 575
- Harlow FH (1964) The particle-in-cell computer method for fluid dynamics. In B Alder (ed). *Comput Phys* 3:319–343
- Hern C, Nordlund U, van der Zwan K, Ladipo K (2001) Forward prediction of Aeolian systems using fuzzy logic, constrained by

- data from recent and ancient analogues. *Geologie en Mijnbouw/Netherlands J Geosci* 80(1):53–70
- King CAM (1968) SPITSIM. In: The use of computers in geomorphological research, British geomorphological research group symposium, Occ Paper 6. Geo Abstracts, Norwich, pp 63–72
- Krumbein WC (1947) Shore processes and beach characteristics. United States Beach Erosion Board, Technical memorandum #3
- Krumbein WC (1968) Fortran IV computer program for simulation of transgression and regression with continuous-time Markov models. Computer Contribution 26, Kansas Geological Survey
- Krumbein WC, Hall JV (1964) A geological process-response model for analysis of beach phenomena, Northwestern University Technology Report 8, ONR Contract 1228(26). Northwestern University, Evanston
- Lawrence DT, Doyle M, Snelson S, Horsfield WT (1987) Stratigraphic modeling of sedimentary basins. SEG Technical Program Expanded Abstracts: 407–408
- Lessenger MA (1989) A stratigraphic inverse simulation model. In: Proceedings from the American Geophysical Union Chapman conference on sea level estimation (abstr)
- Lessenger MA (1993) Forward and inverse simulation models of Stratal architecture and facies distributions in marine shelf to coastal plain environments, Ph.D. Dissertation, Colorado School of Mines, Golden, Colorado, p 182
- Lessenger MA, Cross TA (1991) A stratigraphic inverse simulation model. *AAPG Bull* 75(3):7–10
- Lessenger MA, Cross TA (1996) An inverse stratigraphic simulation model – is stratigraphic inversion possible? *Energy Explor Exploit* 14(6):627–637
- Lessenger MA, Lerche I (1999) Inverse modeling. In: Harbaugh JW, Watney WL, Rankey EC, Slingerland RL, Goldstein RH, Franseen EK (eds) Numerical experiments in stratigraphy: recent advances in stratigraphic and sedimentologic computer simulations, vol 62. SEPM Society for Sedimentary Geology, pp 29–31
- Lesser GR, Roelvink JA, van Kester JATM, Stelling GS (2004) Development and validation of a three-dimensional morphological model. *Coast Eng* 51:883–915
- Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions and reversals. *Cybern Control Theory* 10(8):707–710
- Martinez PA, Harbaugh JW (1993) Simulating nearshore environments, Computer methods in the geosciences 12. Pergamon Press, p 265
- Nordlund U (1996) Formalizing geological knowledge; with an example of modeling stratigraphy using fuzzy logic. *J Sediment Res* 66(4): 689–698. <https://doi.org/10.1306/D42683E2-2B26-11D7-8648000102C1865D>
- Nordlund U (1999a) FUZZIM: forward stratigraphic modeling made simple. *Comput Geosci* 25(4):449–456
- Nordlund U (1999b) Stratigraphic modeling using common sense rules. In: Harbaugh JW, Watney WL, Rankey EC, Slingerland RL, Goldstein RH, Franseen EK (eds) Numerical experiments in stratigraphy: recent advances in stratigraphic and sedimentologic computer simulations, vol 62. SEPM Society for Sedimentary Geology, pp 85–90
- Nordlund U, Silfversparre M (1994) Fuzzy logic – a means for incorporating qualitative data in dynamic stratigraphic modeling. In: International association for mathematical geology, ... conferences and proceedings, pp 265–266
- Oertel G, Walton EK (1967) Lessons for a feasibility study for computer models of coal-bearing deltas. *Sedimentology* 9:157–168
- Oldenburg DW (1974) The inversion and interpretation of gravity anomalies. *Geophysics* 39(4):389–581
- Paola C (2000) Quantitative models of sedimentary basin filling. *Sedimentology* 47(1):121–178
- Plotnick RE (1986) A fractal model for the distribution of stratigraphic hiatuses. *J Geol* 94(6):885–890
- Revil A, Jardani A (2013) Forward and inverse modeling. In: The self-potential method: theory and applications in environmental geosciences. Cambridge University Press, Cambridge, pp 110–153
- Ritchie BD, Hardy S, Gawthorpe RL (1999) Three-dimensional numerical modeling of coarse-grained clastic deposition in sedimentary basins. *J Geophys Res* 104:17759–17780
- Rivenaes JC (1992) Application of a dual-lithology, depth-dependent diffusion equation in stratigraphic simulation. *Basin Res* 4:133–146
- Salles T, Duclaux G (2014) Combined hillslope diffusion and sediment transport simulation applied to landscape dynamics modelling. *Earth Surf Process Landf* 40:823–839
- Salles T, Hardiman L (2016) Badlands: an open-source, flexible and parallel framework to study landscape dynamics. *Comput Geosci* 91:77–89
- Salles T, Griffiths CM, Dyt C (2011) Aeolian sediment transport integration in general stratigraphic forward modeling. *J Geol Res* 2011:1–12
- Salles T, Ding X, Brocard G (2018a) pyBadlands: a framework to simulate sediment transport, landscape dynamics and basin stratigraphic evolution through space and time. *PLoS One* 13(4): e0195557
- Salles T, Pall J, Webster JM, Dechnik B (2018b) Exploring coral reef responses to millennial-scale climatic forcings: insights from the 1-D numerical tool pyReef-Core v1.0. *Geosci Model Dev* 11:2093–2110
- Sambridge M (1999) Geophysical inversion with a neighbourhood algorithm—I. Searching a parameter space. *Geophys J Int* 138(2): 479–494
- Schwarzacher W (1966) Sedimentation in subsiding basins. *Nature* 210:1349–1350
- Sloss LL (1962) Stratigraphic models in exploration. *J Sediment Petrol* 32(3):415–422
- Straub KM, Duller RA, Foreman BZ, Hajek EA (2020) Buffered, incomplete, and shredded: the challenges of reading an imperfect stratigraphic record. *J Geophys Res Earth* 125:1–44
- Tetzlaff DM (1987) A simulation model of clastic sedimentary processes [unpublished PhD dissertation] Dept. Applied Earth Sciences, Stanford University, Stanford, California. p 345
- Tetzlaff DM, Harbaugh JW (1989) Simulating clastic sedimentation, Computer methods in the geosciences 8. Van Nostrand, New York, p 202
- Tucker GE, Hutton E, Piper MD, Campforts B, Gan T, Barnhart KR, Kettner A, Overeem I, Peckham SD, McCreedy L, Syvitski J (2021) Numerical modeling of Earth's dynamic surface: a community approach. Community Surface Dynamics Modeling System (CSDMS) Technical Report & Preprint submitted to EarthArXiv
- Wan L, Bianchi V, Hurter S, Salles T (2020) Evolution of a delta-canyon-fan system on a typical passive margin using stratigraphic forward modelling. *Mar Geol* 429(106310):1–18
- White N, Bellingham P (2002) A two-dimensional inverse model for extensional sedimentary basins 1. Theory *J Geophys Res* 107(B10):2259
- Whitten EHT (1964) Process-response models in geology. *GSA Bull* 75(5):455–464
- Wijns C, Poulet T, Boschetti F, Dyt C, Griffiths CM (2004) Interactive inverse methodology applied to stratigraphic forward modelling. Geological Society of London Special Publication, Geological Prior Information, pp 147–156
- Zhang J, Flaig P, Wartes M, Aschoff J, Shuster M. (2021) Integrating stratigraphic modelling, inversion analysis, and shelf-margin records to guide provenance analysis: An example from the Cretaceous Colville Basin, Arctic Alaska. *Basin Research* 24;33(3):1954–66

Fractal Geometry and the Downscaling of Sun-Induced Chlorophyll Fluorescence Imagery

Juan Quiros-Vargas¹, Bastian Siegmann¹, Alexander Damm^{2,3}, Ran Wang⁴, John Gamon^{4,5}, Vera Krieger¹, B. S. Daya Sagar⁶, Onno Muller¹ and Uwe Rascher¹

¹Institute of Bio- and Geosciences, IBG-2: Plant Sciences, Forschungszentrum Jülich GmbH, Jülich, Germany

²Department of Geography, University of Zurich, Zürich, Switzerland

³Eawag, Swiss Federal Institute of Aquatic Science & Technology, Surface Waters – Research and Management, Dübendorf, Switzerland

⁴Center for Advanced Land Management Information Technologies, School of Natural Resources, University of Nebraska–Lincoln, Lincoln, NE, USA

⁵Department of Earth and Atmospheric Sciences and Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada

⁶Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, Karnataka, India

Synonyms

Geometry of nature

Definition

The fractal geometry is a young branch of mathematics that studies the configuration of complex shapes and phenomena formed by the repetition of patterns.

Introduction

“Clouds are not spheres, mountains are not cones, coastlines are not circles, and bark is not smooth, nor does lightning travel in a straight line” (Mandelbrot 1982). The fractal geometry was developed by Benoît Mandelbrot in the 1970s to describe those and all other chaotic geometries observed in nature. The theory states that natural phenomena can be characterized as a repetition of self-similar (isotropic) or self-affine (anisotropic) patterns with dimensions, not limited to the Euclidean 1-, 2-, and 3-D planes. Thus, the complexity of fractal geometries is described by the so-called Fractal Dimensions (FDs), real numbers within the [1, 4] interval. This means that the more complex a line, a polygon, or a prism is, the higher to 1 (but lower than 2), 2 (but lower than 3), and 3 (but lower than 4) its dimension is, respectively.

Applications of the fractal geometry in science are nearly as many as natural phenomena exist. One example is the Remote Sensing (RS)-based investigation of scale dependencies of dynamic processes in vegetation, especially between individual leaves and the ecosystem. Since such dynamic processes are composed by nonlinear phenomena, the implementation of linear methods for its analysis might lead to local approaches with limited applicability beyond specific circumstances. The use of the fractal geometry, instead, can greatly contribute to RS of vegetation and help understanding image features as interconnected and scale-independent spatiotemporal patterns.

Fractal Geometry and the Power Law (PL) Distribution of Sun-Induced Chlorophyll Fluorescence (SIF)-Emitting Objects

Several approaches exist to mathematically prove the existence of fractal geometry in the distribution of RS-measured objects; the computation of universal Power Laws (PLs), also known as scaling laws, is one of them (Nagajothi et al. 2021). A PL distribution indicates the presence of fractal geometry composed by patterns where few occurrences of large magnitudes and numerous occurrences of small magnitudes are observed regardless the scale. The function of a PL distribution is described with the following equation:

$$y = \alpha x^\beta$$

where α is a constant representing the scale, and β is the exponent which represents the dimension of the function (closely related to the fractal dimension). Examples of PL distribution can be found almost in any area of research. For instance, in economics, a PL can describe how the richness in a country is distributed in few individuals holding huge amounts of money and many individuals with scarce monetary resources. Another example can be found in linguistics, where the frequency of words in a text follows a PL as well, with few words repeated thousands of times, and many words used only in few occasions.

Hypothetically, PLs can be used to describe the distribution of RS-measured properties of vegetation like the emitted SIF signal, since: (i) Its heterogeneity is related to spatial patterns defined by factors like canopy structure, shape, and size of agricultural fields or the organization of tree species in a forest; and (ii) SIF, as a red light emitted from vegetation, is more clearly aggregated across spatial scales (pixel sizes) than other more uniform signals like reflectance. Therefore, this chapter focusses on the analysis of the distribution of SIF emitting objects in an agricultural field to confirm whether they present a PL distribution (fractal geometry) or not, and

how the scaling and dimension factors behave across spatial scales.

Case Study: “Fractal Geometry for SIF-Downscaling”

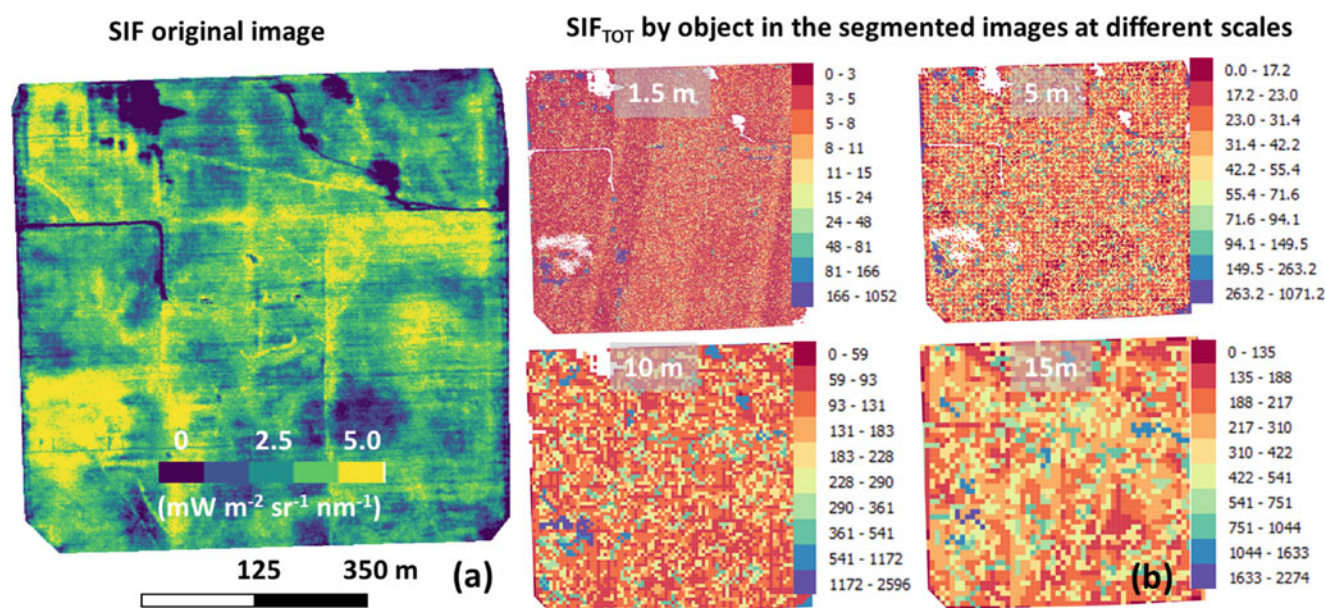
Context: The world is confronted with substantial socioeconomic challenges. Agricultural production must ensure food security for a growing population, while it is imperative to protect and sustainably manage natural resources. Cropland expansion at the cost of natural ecosystems is not an option; therefore, large-scale farming must employ technology and advanced management to improve yields. RS offers possibilities to contribute making this possible. Recently available RS platforms equipped with SIF sensors constitute a powerful tool, since SIF shows large sensitivity to actual photosynthetic efficiency of crops (Mohammed et al. 2019). Such information is complementary to commonly used Vegetation Indices (VIs) and offers new pathways to assess plant health and, thus, optimize crop management. The forthcoming Fluorescence Explorer (FLEX) satellite mission of the European Space Agency (ESA) was planned within this context. Its relatively coarse resolution of 300 m, however, limits applicability for small field sizes and to investigate inter-field heterogeneity of croplands. Thus, downscaling satellite SIF products (understood as the increase of the spatial resolution) is currently a research subject of utmost importance.

Previous studies have addressed the SIF-downscaling considering its physical relation with explanatory variables like

the land surface temperature (Duveiller et al. 2020), and using machine learning-based statistical relations with VIs (Zhang et al. 2020). Another approach with applications on plant phenotyping was presented by Krämer et al. (2021), who analyzed the potential use of aggregated SIF pixel to extract representative values for crop plots representing few m^2 . Despite the great relevance of these contributions, a more versatile downscaling approach capable to run in a diversity of ecosystems is still not available but needed. We hypothesize that the flexibility of the fractal theory as a complexity measure together with the discontinuity of self-similar geometries captured in RS data (Sun et al. 2006) enables new strategies to advance downscaling approaches of SIF.

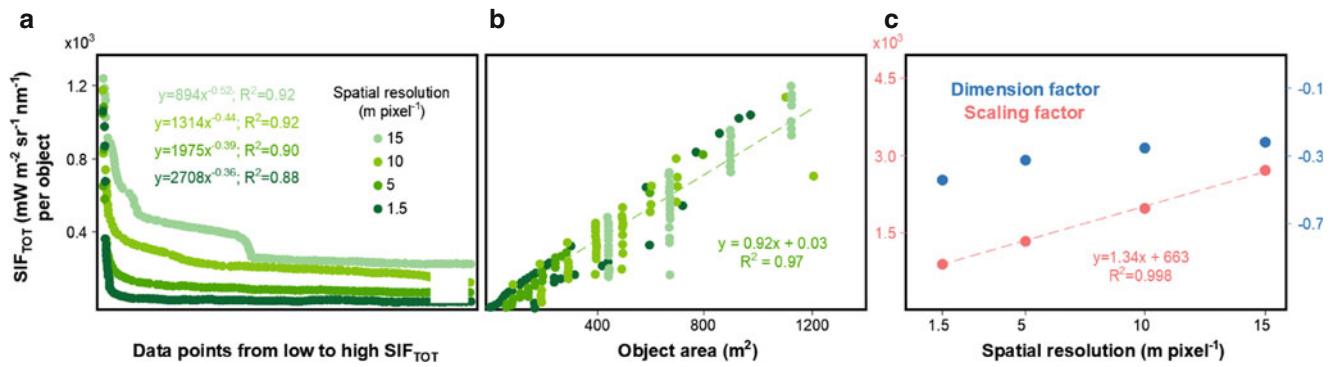
Aims: This study aims (i) to analyze if the aggregation of SIF over vegetation objects follows a universal PL distribution (i.e., representing fractal geometry) across spatial scales, and (ii) to evaluate how the scaling and dimension components of the PLs behave across spatial scales.

Analysis: We analyze the fractal geometry vegetation (SIF-emitting) objects through the computation of PLs at 1.5, 5, 10, and 15 m pixel^{-1} resolutions. Far red SIF at 740 nm (hereafter named “SIF”) in a 65 ha soybean field was retrieved from an imaging fluorometer (IBIS, Specim, Oulu Finland) with the Spectral Fitting Method (SFM, Cogliati et al. 2015; Fig. 1a). Only vegetation pixels were considered and afterward segmented into individual homogeneous objects through unsupervised classification (Fig. 1b). For the first aim, the total SIF (SIF_{TOT}) of each object was calculated and its distribution plotted in order to identify the presence of a PL distribution. For the second aim, the scaling



Fractal Geometry and the Downscaling of Sun-Induced Chlorophyll Fluorescence Imagery, Fig. 1 Sun-Induced chlorophyll Fluorescence (SIF) image (a) and the respective segmented images

considering individual aggregation scales, i.e., 1.5, 5, 10, and 15 m pixel^{-1} (b)



Fractal Geometry and the Downscaling of Sun-Induced Chlorophyll Fluorescence Imagery, Fig. 2 Power Law (PL) distributions of the total Sun-Induced chlorophyll Fluorescence (SIF_{TOT}) data observed

and dimension factors of the PLs from the different spatial resolutions were compared in order to understand if they follow scale-independent patterns.

Results: Besides the high intrafield variability observed in Fig. 1, the distribution of the SIF_{TOT} information aggregated by objects follows the PL distribution at all analyzed scales (Fig. 2a). Such order, characterized in the function equations, describes how the data is arranged in patterns of few occurrences of objects with high SIF_{TOT}, and numerous occurrences of objects with low SIF_{TOT} emission. For downscaling purposes, the information about known variables related to the analyzed feature, like the object area (Fig. 2b), could potentially be used to infer the SIF_{TOT} of the polygons. Furthermore, the linear increase of the scaling factor and the nearly constant dimension factor for 1.5 to 15 m pixel⁻¹ sizes (Fig. 2c) might indicate the scale-invariant property of the SIF signal aggregation patterns within the range of analyzed resolutions.

Summary

Our results indicate that the distribution of the SIF_{TOT} information aggregated by object follows the PL distribution for 1.5 to 15 m pixel⁻¹ resolutions, i.e., the fractal geometry is present in the distribution of individual SIF-emitting objects at all analyzed scales. Moreover, the linear ($R^2 = 0.998$) increase of the PL-scaling component, and the nearly constant dimension across the scales, might be an indicator of the scale-invariant property of the SIF signal aggregation patterns within the range of analyzed resolutions.

In theory, similar PL characterizations of high resolution SIF-explanatory variables could help to understand how to spatially disentangle SIF_{TOT} represented in a coarse pixel (e.g., measured by FLEX) in the SIF_{TOT} contributions of individual vegetation objects within the measurement footprint. Further studies might address that subject across diverse

ecosystems to provide more evidence on the suitability of this proposed approach to downscale and link coarse-scale-retrieved SIF_{TOT} with field-level SIF_{TOT} information.

Cross-References

- [Allometric Power Laws](#)
- [Fractal Geometry in Geosciences](#)
- [Fractal Landscape](#)
- [Multifractals](#)
- [Remote Sensing](#)
- [Scaling and Scale Invariance](#)

Bibliography

- Cogliati S, Verhoef W, Kraft S, Sabater N, Alonso L, Vicent J, Moreno J, Drusch M, Colombo R (2015) Retrieval of sun-induced fluorescence using advanced spectral fitting methods. *Remote Sens Environ* 169: 344–357
- Duveiller G, Filippini F, Walther S, Köhler P, Frankenberg C, Guanter L, Cescatti A (2020) A spatially downscaled sun-induced fluorescence global product for enhanced monitoring of vegetation productivity. *Earth Syst Sci Data* 12:1101–1116
- Krämer J, Siegmund B, Thorsten K, Onno M, Rascher U (2021) The potential of spatial aggregation to extract remotely sensed sun-induced fluorescence (SIF) of small-sized experimental plots for applications in crop phenotyping. *Int J Appl Earth Obs Geoinf* (in press)
- Mandelbrot B (1982) *The fractal geometry of nature*. Freeman, San Francisco, p 1982
- Mohammed GH, Colombo R, Middleton EM, Rascher U, van der Tol C, Nedbal L, Goulas Y, Pérez-Priego O, Damm A, Meroni M, Joiner J, Cogliati S, Verhoef W, Malenovsky Z, Gastellu-Etchegorry JP, Miller JR, Guanter L, Moreno J, Moya I, Berry JA, Frankenberg C, Zarco-Tejada PJ (2019) Remote sensing of solar-induced chlorophyll fluorescence (SIF) in vegetation: 50 years of progress. *Remote Sens Environ* 231:111177
- Nagajothi K, Rajashekara HM, Sagar BSD (2021) Universal fractal scaling laws for surface water bodies and their zones of influence. *IEEE Geosci Remote Sens Lett* 18(5):781–785

- Sun W, Xu G, Gong P, Lliang S (2006) Fractal analysis of remotely sensed images: a review of methods and applications. *Int J Remote Sens* 20(22):4963–4990
- Zhang Z, Xu W, Qin Q, Long Z (2020) Downscaling solar-induced chlorophyll fluorescence based on convolutional neural network method to monitor agricultural drought. *IEEE Trans Geosci Remote Sens* 59(2):1012–1028

Fractal Geometry in Geosciences

Qiuming Cheng^{1,2} and Frits Agterberg³

¹School of Earth Science and Engineering, Sun Yat-Sen University, Zhuhai, China

²State Key Lab of Geological Processes and Mineral Resources, China University of Geosciences, Beijing and Wuhan, China

³Central Canada Division, Geomathematics Section, Geological Survey of Canada, Ottawa, ON, Canada

Definition

A fractal is an object or feature characterized by its fractal dimension that differs from the integer Euclidian dimension of the space in which the fractal is imbedded. On the one hand, fractals are often closely associated with the random variables studied in mathematical statistics; on the other hand, they are connected with the concept of “chaos” that is an outcome of some types of nonlinear processes. Fractals are phenomena measured in terms of their presence or absence in boxes belonging to arrays superimposed on the domain of study in 1-D, 2-D, or 3-D space. Several phenomena that originally were thought to be fractals turned out to multifractals, which are “measures” representing of how much of a feature is present within the boxes used for measurement. Multifractals are spatially intertwined fractals.

Introduction

The word “fractal” was coined by Mandelbrot (1975). Evertsz and Mandelbrot (1992) explain that fractals are phenomena measured in terms of their presence or absence in boxes belonging to arrays superimposed on the domain of study in 1-D, 2-D, or 3-D space, whereas multifractals are either spatially intertwined fractals (*cf.* Stanley and Meakin 1988) or mixtures of fractals in spatially distinct line segments, areas, or volumes that are combined with one another. During the past 40 years, the fractal geometry of many natural features in Nature has become widely recognized (see, e.g., Mandelbrot 1983; Cheng 1994; Stoyan and Stoyan 1994, Turcotte 1997; Sagar et al. 2004, Carranza 2008; Ford and

Blenkinsop 2009; Agterberg 2012; Cheng 2006, 2008, 2012; Korvin 1992; Zuo and Wang 2016; Agterberg 2020). Fractals in geoscience either represent the end products of numerous, more or less independent, processes (e.g., coastlines and topography), or they result from nonlinear processes, many of which took place millions of years ago within the Earth’s crust. Although a great variety of fractals can be generated by relatively simple algorithms, theory needed to explain fractals of the second kind generally is not so simple, because previously neglected nonlinear terms have to be inserted into existing linear, deterministic equations (Lovejoy and Schertzer 2018).

Many geological features display random characteristics that can be modeled by adopting methods of mathematical statistics. A question to which new answers are being sought is: Where does the randomness in Nature come from? Nonlinear process modeling is providing new clues to answers. Spatial distribution of chemical elements in the Earth’s crust and features such as the Earth’s topography that traditionally were explained by using deterministic process models now also are modeled as fractals or as multifractals. Which processes have produced phenomena that are random and often characterized by non-Euclidian dimensions? The physico-chemical processes that have resulted in the Earth’s present configuration were essentially deterministic and “linear” but globally as well as locally they may display random features that can only be modeled by adopting a nonlinear approach that is increasingly successful in localized prediction. The situation is analogous to the relation between climate and weather. Longer-term climate change can be modeled deterministically, but short-term weather shows random characteristics that are best modeled by adopting a nonlinear approach in addition to the use of conventional deterministic equations.

This article reviews fractal modeling of solid Earth observations and processes with emphasis on topography, thickness measurements, geochemistry, and hydrothermal processes. There is significant overlap with multifractals that also will be discussed. Special attention is paid to improvements in goodness of fit and prediction obtained by nonlinear modeling. Some examples are as follows: The Pareto distribution is closely associated with fractal modeling of metal distribution within rocks or surficial cover in large regions. The concentration-area (C-A) method (Cheng et al. 1994) is a useful new tool for geochemical prospecting to help delineate subareas with anomalously high element concentration values that can be targets for further exploration involving drilling. Cheng (2016) has introduced the concept of fractal density that replaced ordinary density of substances with multifractal characteristics. The spatial distribution of ore deposits within large regions or within worldwide permissive tracts often is fractal or multifractal.

Two frequency distribution models frequently used in the mathematical geosciences are the lognormal and the Pareto.

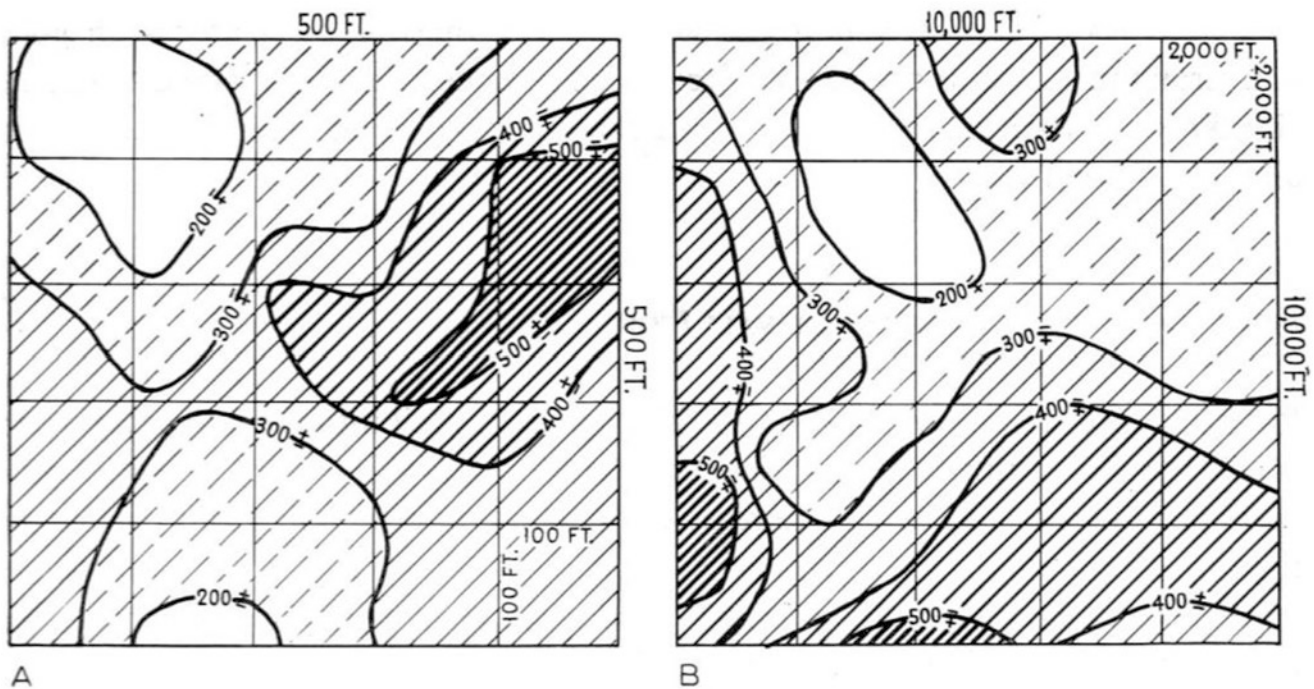
The normal distribution that underlies much, especially early, theory of mathematical statistics can be regarded as a lognormal distribution with negligibly small coefficient of variation (ratio of mean and standard deviation). Both the lognormal and the Pareto are the so-called stable frequency distributions with generating processes that involve summations of random variables with different (usually unknown) frequency distributions. The Pareto applies to tails of frequency distributions only. Both types of frequency distribution often can be explained as the result of multifractal cascade-type modeling.

Fractal Dimension Estimation

Fractals can arise in chaotic situations. A famous example of chaos is Poincaré's (1899) study of the three-body problem in mechanics. Small differences in the initial conditions produce very great differences in the outcomes. Likewise, Lorentz (1963) discovered how miniscule changes in the initial state of a time series computer simulation experiment grew exponentially with time to yield results that did not resemble one another at all, although the equations controlling the process were the same in all experiments. The different outcomes orbited along a bounded region in 3-D space known as a chaotic attractor with fractal properties.

Fractals also may reflect self-similarity. Mandelbrot (1975) discovered that the increased detail of many different types of natural phenomena when they are studied at larger-scale maps or quantified with longer yardsticks often display self-similarity and can be quantified by estimating their fractal dimension. An example of self-similarity persistent across large areas is shown in Fig. 1 from Krige (1966). The two maps represent 2-D moving averages of gold inch-dwt values in the Klerksdorp goldfield in South Africa. In Fig. 1a, these averages were computed for overlapping square areas measuring 500 feet on a side; in Fig. 1b, the squares measured 10,000 feet on a side. Thus, each average value contoured in Fig. 1b was based on about 400 as many original values as the average contoured in Fig. 1a. Nevertheless, the resulting patterns are similar in appearance indicating self-similarity.

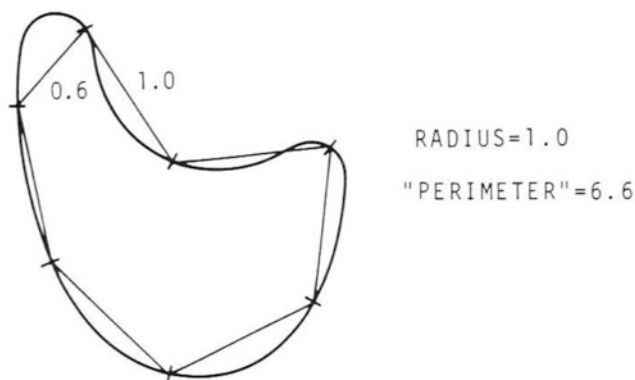
A second example of fractal dimension estimation is given in Figs. 2, 3, and 4. The perimeter of an object in two-dimensional space can be estimated by approximating the true perimeter by a sequence of straight-line segments (yardsticks or unit radii) as illustrated in Fig. 2. These segments have the same length except the last one in the sequence. In general, the estimated perimeter underestimates the true perimeter except when the line segments become infinitely short. Mandelbrot (1975) applied this method to measure the length of the coastline of Britain. A log-log plot



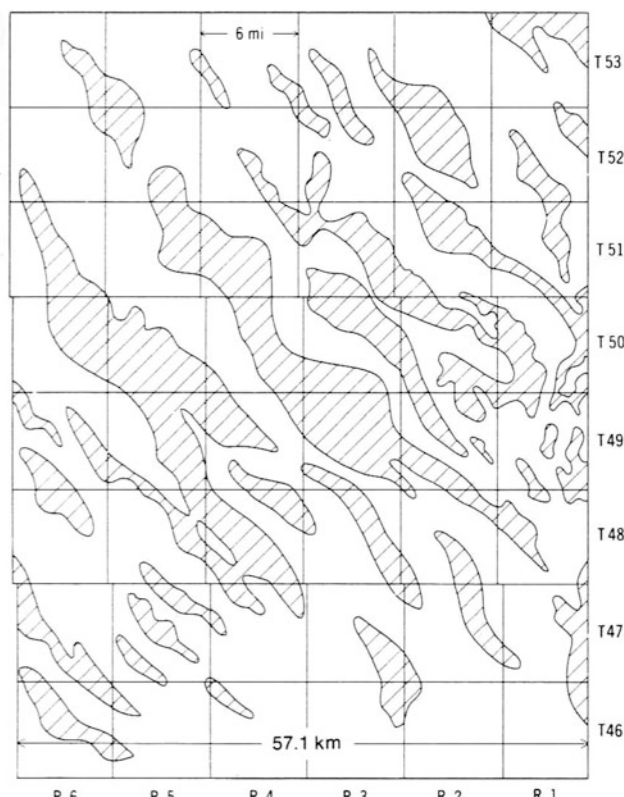
Fractal Geometry in Geosciences, Fig. 1 Gold inch-dwt trend surfaces in the Klerksdorp goldfield based on 2-dimensional moving averages for two areas with similar average gold grades. (a) Moving averages of 100 × 100 ft. areas within a mined-out section of 500 × 500 ft. (b) Moving averages of 2,000 × 2,000 ft. areas within mined-out section of

10,000 × 10,000 ft. Although the area covered in the case of (b) is 400 times as large as that in the case of (a), the resulting patterns are similar indicating scale-independence across large areas in this goldfield (Source: Krige 1966, Fig. 3)

of the measured length of coastline (y) against segment length (x) for segments greater than a critical limit value could be approximated by a straight line, $y = a + b \cdot x$. He defined the fractal dimension D as being equal to $2 - b$. For very short yardsticks, $b = 0$ and $D = 2$. A simple example of application of this approach to a geological map is to be shown in Figs. 3 and 4.



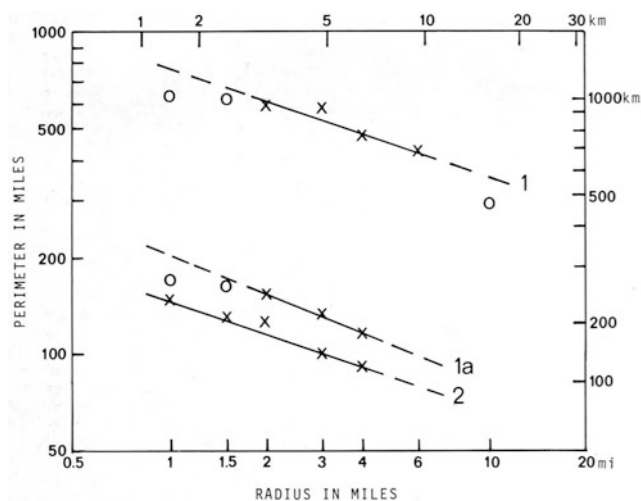
Fractal Geometry in Geosciences, Fig. 2 Measurement of the perimeter of a 2-dimensional object depicted on a geological map (Source: Agterberg 1980, Fig. 2)



Fractal Geometry in Geosciences, Fig. 3 Thirty-foot contour map of Sparky sandstone thickness in Lloydminster area, Alberta, mapped by Burnett and Adams (1977) (Source: Agterberg 1980, Fig. 3)

Figure 3 is a contour map of 30 feet (9.14 m) thickness of the oil-rich Sparky sandstone unit near Lloydminster, Alberta, originally constructed by Burnett and Adams (1977). In Agterberg (1980) the perimeters of these objects were measured using yardsticks of different lengths for the method explained in Fig. 1, with the results shown in Fig. 4 (Case 1). When the unit radius is set at 1.5 miles (2.4 km) or less, the measured perimeter provides a reasonably good approximation of the combined length of the contours in Fig. 3. However, for radii greater than 1.5, a shorter “perimeter” is measured. For yardsticks between 2 and 6 miles in length (Case 1 in Fig. 4), the measurements fall approximately on a straight line with slope equal to 0.66 from which it follows that $D \approx 1.34$. Yardsticks of 10 miles or longer were not used because then only few relatively large objects can be measured.

Burnett and Adams (1977) also published a more detailed Sparky sandstone thickness map for a much smaller subarea within the area of Fig. 3. For the objects shown in Fig. 2, this subarea gave $D \approx 1.38$. The more detailed thickness map (not shown here) gave $D \approx 1.34$ (Case 2 in Fig. 4). The difference between results for the subarea and the area of Fig. 2 is that shorter yardsticks could be used in Case 2, because the 30 ft. Sparky thickness contours are shown in much greater detail on this larger-scale map. Tentatively, it may be assumed that the degree of smoothing used for the smaller-scale map was the same as that for the larger-scale map on which more detail could be shown. The fractal dimension is independent of the degree of smoothing. This example also has been considered



Fractal Geometry in Geosciences, Fig. 4 “Perimeter” measured for 2-dimensional objects depicted in Fig. 3 by method illustrated in Fig. 2 (Case 1). Results for smaller subarea within the area of Fig. 2 are shown as Case 1a. Perimeters measured on a more detailed (larger scale) map are shown as Case 2. The three lines of best fit show similar slopes but results for smaller units of distance to measure perimeters on the larger scale map fit in with results for larger units of distance indicating scale independence (Source: Agterberg 1980, Fig. 7)

by Mandelbrot (1983). It illustrates that the principle of self-similarity can be applicable to geological maps produced at different scales. Self-similarity or independence of scale is a feature of multifractals as well.

An early success of fractal geometry in the geosciences was the modeling of Brownian landscapes by Mandelbrot (1977). Realistic representations of mountainous terrains with dimensions less than 3 and for elevation contours less than 2 were obtained in a variety of computer simulation experiments. Later, multifractal models also were used for the modeling of topography (Lovejoy and Schertzer 2007). These authors have reviewed the history of fractal geometry pointing out that, early on, several mathematicians and geophysicists had identified problems of differentiability and integrability in connection with some types of natural phenomena. For example, Perrin (1913) wrote: "Consider the difficulty in finding the tangent to a point of the coast of Brittany ... depending on the resolution of the map the tangent would change. The point is that a map is simply a conventional drawing in which each line has a tangent. On the contrary, an essential feature of the coast is that ... at each scale we guess the details which prohibit us from drawing a tangent." With respect to problems of integrability, Steinhaus (1954) stated: "The left bank of the Vistula when measured with increased precision would furnish lengths ten, hundred, or even a thousand times as great as the length read off a school map. A statement nearly adequate to reality would be to call most arcs encountered in nature as not rectifiable."

Among the early pioneers, Lovejoy and Schertzer (2007) list Vening Meinesz (1951) who argued that the power spectrum $P(k)$ of the Earth's topography has the scaling form $k^{-\beta}$ where k is a wave number ($= 2\pi \times$ frequency) and $\beta \approx 2$. Vening Meinesz (1951) originally based his scaling model on a development of the Earth's topography in terms of spherical harmonics up to the 16th order making a separation between continents and oceans. He showed that mean square elevation (y) is approximately related to order x of harmonic according to the power law relation $y = C \cdot x^{-\beta}$ where C and β are constants. Orders of spherical harmonics are analogous to wave numbers in a periodogram which is a commonly used tool in time series analysis. A multifractal series of values is characterized by a straight line on the log-log plot of its power spectrum (Agterberg 2014). Vening Meinesz (1964) refined his earlier analysis by development in spherical harmonics of the Earth's topography to the 31st order. His later curves suggest that β is slightly less than 2 for $n > 5$, although there may exist an increase in y for higher orders, especially for continental topography. Although the precise value of β is not known exactly, it is clear that the Earth's topography on the continents as well as on the ocean floor approximately displays the same type of fractal/multifractal behavior.

Separation of Geochemical Anomalies from Background by Fractal Methods

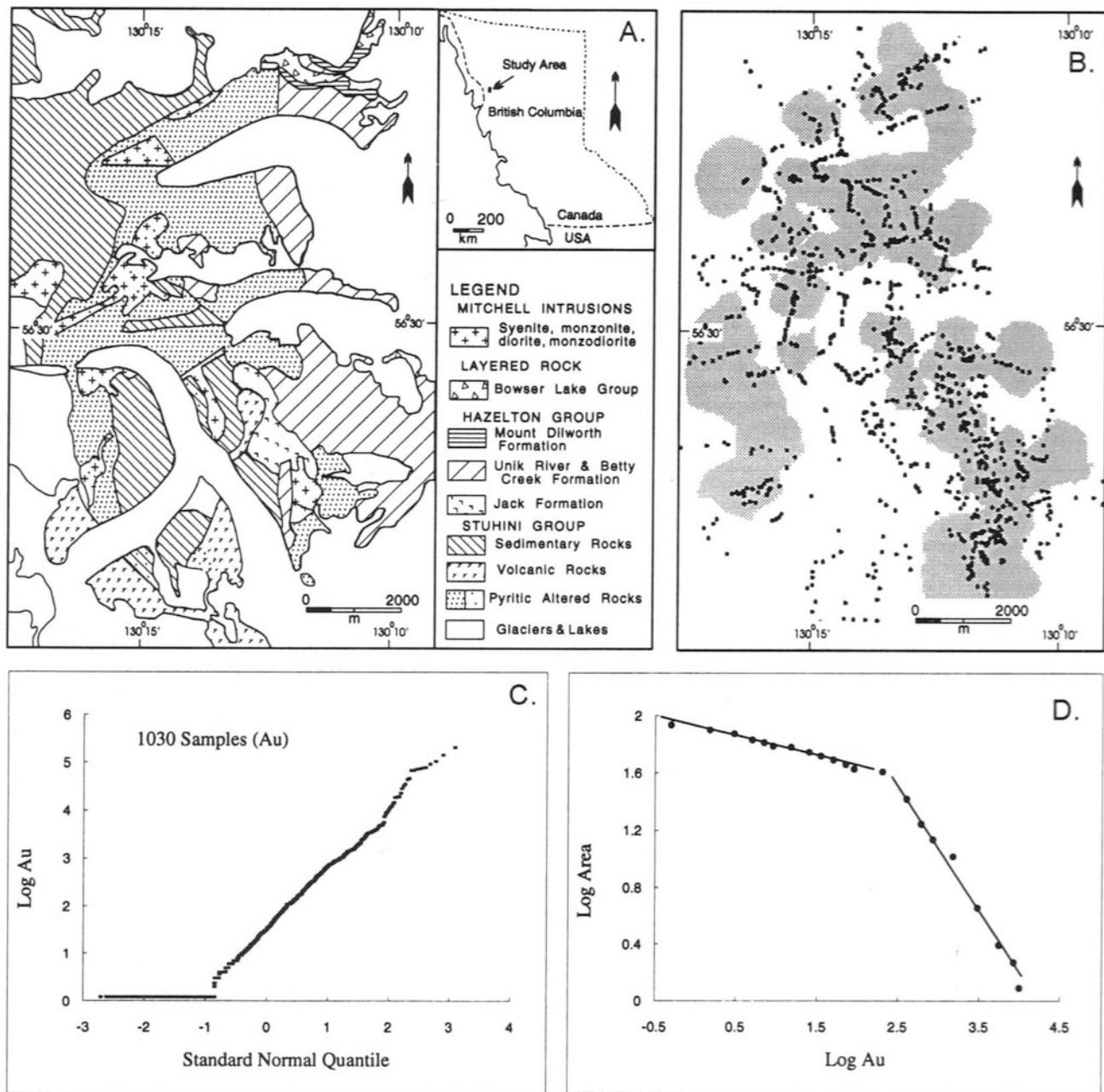
Cheng (1994) has studied the spatial distribution of 1030 concentration values of gold and associated elements in partially altered volcanic rocks in the Mitchell-Sulphurets area, northwestern British Columbia (Fig. 5a and b). A plot of Au content in surface samples from altered and unaltered rocks using logarithmic probability paper ($=$ lognormal $Q-Q$ plot, Fig. 5c) suggests that the gold concentration values satisfy a single, positively skewed frequency distribution that is approximately lognormal. The alteration pattern of Fig. 5b closely resembles the >200 ppb Au pattern of Fig. 6 showing a sequence of geochemical isopleths (isolines or contours for concentration values in ppb Au). In Fig. 5d, log-log paper is used to plot the cumulative area of surface rocks with larger concentration values against the contour value. The altered and unaltered rocks show two different power-law relations. This indicates that a power-law-based approach can be useful for the delineation of geochemical anomalies in addition to commonly used approaches directly based on element frequency distributions and contour maps (see, e.g., Sinclair 1991).

From the isopleths of Fig. 6, an optimum threshold for separating anomalies from background areas can be selected by means of the log-log plot for the element concentration-area relation shown in Fig. 5d. This threshold coincides with a sudden change in the rate of decrease of the area enclosed by higher value contours on the log-log plot. Suppose $A(\rho)$ denotes the area with concentration values greater than the contour value ρ . This implies that $A(\rho)$ is a decreasing function of ρ . If v represents the threshold, the following empirical model often provides a good fit to the data for different elements in the study area:

$$A(\rho \leq v) \propto \rho^{-\alpha_1}; A(\rho > v) \propto \rho^{-\alpha_2}$$

where $\propto$ denotes proportionality; α_1 and α_2 are different exponents.

Figure 7 is a log-log plot satisfying the preceding two power-law relations for Au. All areas were computed from separate contour maps such as those shown in Fig. 6. Pairs of estimated exponents and corresponding optimum thresholds for 28 elements or oxides were presented by Cheng (1994, Table 1). The thresholds delineate anomalous areas. Comparison of the areas above and below the threshold of 200 ppb Au on the contour map (Fig. 6) with the geological map (Fig. 5a) shows significant spatial correlation between the areas with Au concentration above 200 ppb and Au-associated alteration zones. This concentration-area plot approach (C-A model, cf. Cheng et al. 1994) has become widely used in geochemical exploration (Cheng 1999).



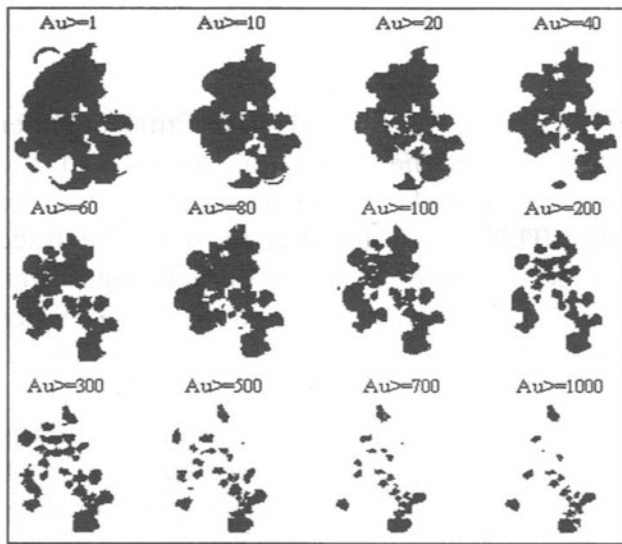
Fractal Geometry in Geosciences, Fig. 5 Mitchell-Sulphurets mineral district, northwestern British Columbia. Four plots originally constructed by Cheng (1994): (a) Simplified geology after R.V. Kirkham (personal information, 1993) including outline of alteration zones; (b) Sampling sites and (smoothed) isopleth for 200 ppb gold

concentration value; (c) Lognormal $Q-Q$ plot (Au terminations below 2 ppb detection limit are shown as horizontal line); (d) Log-log plot for area enclosed by isopleths (including circumference of pattern for 200 ppb in b). Straight lines are least-squares fits (log base: 10 in ppb) (Source: Cheng (1994), Fig. 3)

Fractal Perimeter-Area Relation

The following perimeter-area ($L-A$) relationship was introduced by Mandelbrot (1983) for similarly shaped sets: $L(\delta) = C \cdot \delta^{(1-D)} \sqrt{A(\delta)^D}$ where C is a constant. A modification of this equation was introduced by Cheng (1994) as follows (*cf.* Cheng et al. 1994). For yardstick δ , the estimated length and area

of a pattern in 2-D can be expressed as: $L(\delta) = L_0 \delta^{(1-D)}$; $A(\delta) = A_0 \delta^{(2-D)}$. Theoretically, for a group of similarly shaped sets, there exist power-law relationships between any two measures of area and perimeter. For example, two areas A_i and A_j enclosed by contours i and j are related to their perimeters L_i and L_j as follows:



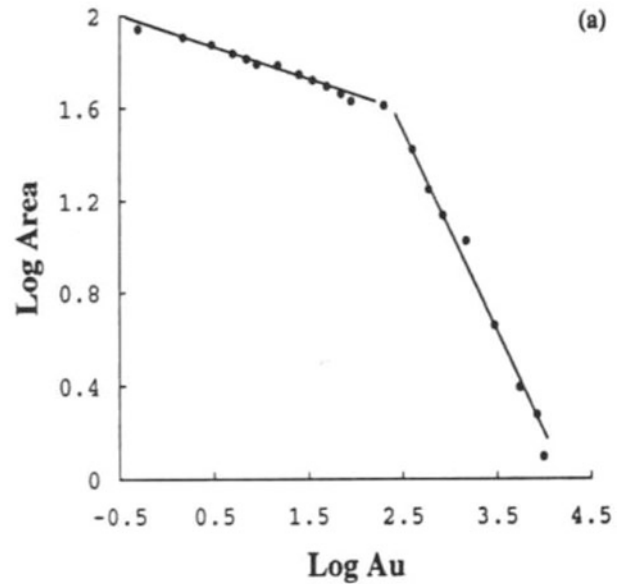
Fractal Geometry in Geosciences, Fig. 6 Successive binary patterns for separate contours of Au, Mitchell-Sulphurets area (Source: Cheng et al. 1994, Fig. 5)

$$\frac{L_j(\delta)}{L_i(\delta)} = \left[\frac{A_j(\delta)}{A_i(\delta)} \right]^{D_{AL}/2}$$

where δ is the common yardstick used for measuring both areas and perimeters. The fractal dimension D_{AL} satisfies $D_{AL} = 2D_L / D_A$, where D_L and D_A denote fractal dimensions of perimeter and area, respectively. The contours for Au in Fig. 6 show similar shapes suggesting that the current perimeter-area modeling approach is valid.

Cheng (1994) subjected element concentration maps as shown for Au in Fig. 6 to perimeter-area analysis as follows. Experimental results for Au, Cu, and As are shown on log-log plots in Fig. 8 using perimeters and areas for different contours with concentration values above the previously defined threshold using the element concentration-area method (Fig. 5d for Au). There appear to be two different fractal dimensions in these diagrams on the right side of Fig. 8. The first of these (D_1) for yardsticks less than 300 m is equivalent to the so-called “textural” fractal dimension (Kaye 1989, p. 27). It is close to 1 for the elements considered and can be explained as the result of smoothing during interpolation. The other estimate of D_L (D_2) ranges from 1.14 to 1.33 and is probably representative of the true geometry of anomalous areas for the elements considered. It is noted that the least square estimates of D_1 and D_2 in Fig. 8 are subject to considerable uncertainty because they are based on relatively few points.

From any two of the three fractal dimensions (D_L) for perimeter, D_{AL} for perimeter-area relation, and D_A for area, the third one can be obtained by means of the relation $D_A = 2D_L / D_{AL}$. Theoretically, D_A cannot be greater than

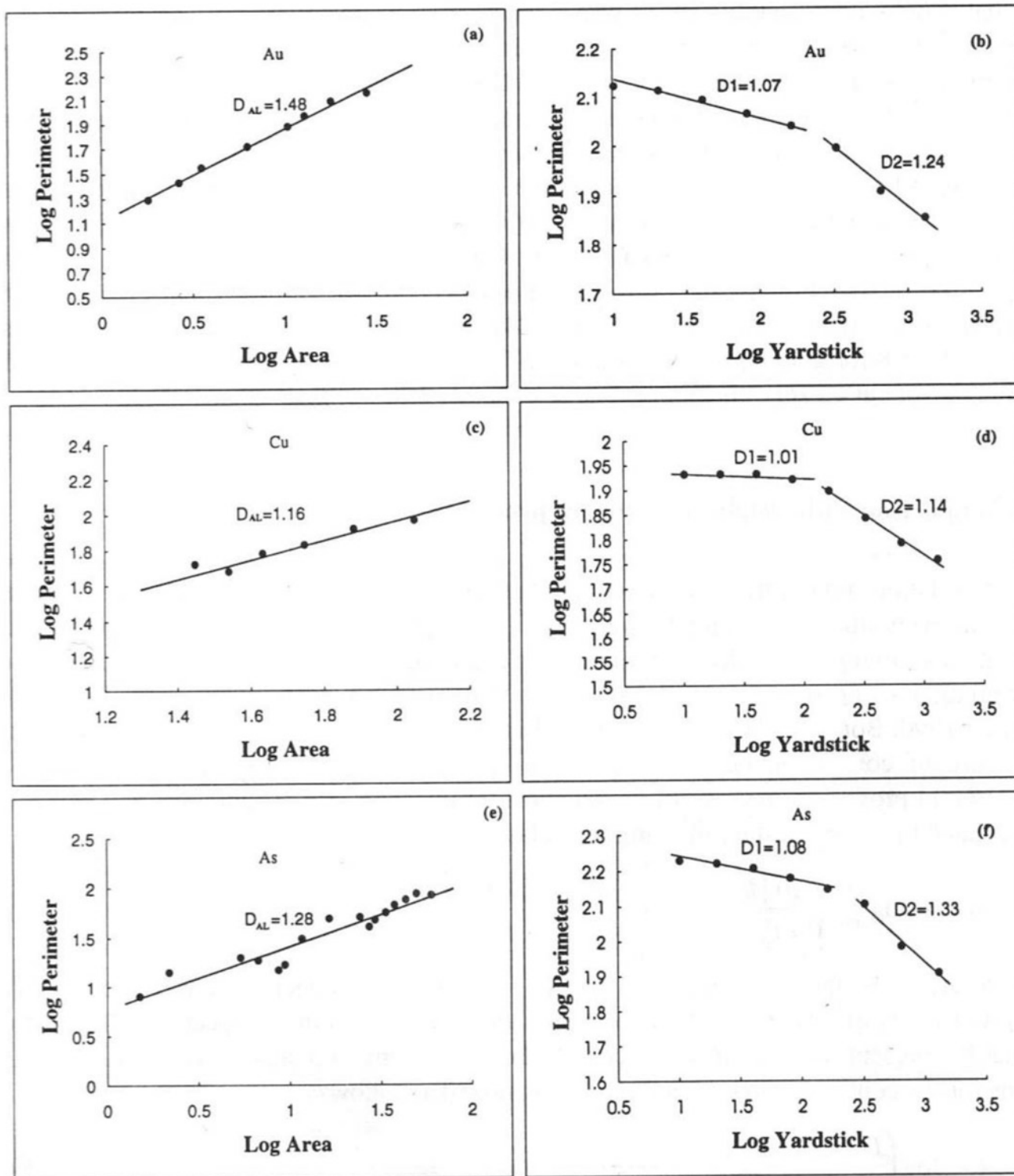


Fractal Geometry in Geosciences, Fig. 7 Log-log plot representing the relationship between areas bounded by contours as shown in Fig. 6. Straight-line segments were fitted by least squares (Source: Cheng et al. 1994, Fig. 6)

2. In most applications of the perimeter-area method, it is set equal to 2 so that $D_L = D_{AL}$. The latter relation holds approximately true for Cu and As ($D_{AL} = 1.16$ and $D_A = 1.96$ for Cu and $D_{AL} = 1.28$ and $D_A = 2.07$ for As). For Au, the estimated value of D_{AL} (1.48) is significantly greater than the estimated structural fractal dimension (1.24). From the relation $D_A = 2D_L / D_{AL}$, it follows that $D_A = 1.68$ for gold which is less than 2. These results indicate that the distribution of Au in the alteration zones is more irregular than that of Cu or As.

The Multifractal Spectrum

Multifractals arose in physics and chemistry as a generalization of (mono)fractals (Meneveau and Sreenivasan 1987; Feder 1988; Lovejoy and Schertzer 2018). Multifractals can be regarded as spatially intertwined monofractals (Stanley and Meakin 1988). Mandelbrot (1989; see also Evertsz and Mandelbrot 1992) has emphasized that multifractals apply to continuous spatial variability patterns, whereas monofractals are for binary Yes-No type patterns. The relation between monofractals and multifractals also was considered by Herzfeld et al. (1995) who showed that the ocean floor could not be modeled as a monofractal. Better results were obtained by a multifractal model after incorporation of a nonstationary component (Herzfeld and Overbeck 1999). The multifractal spectrum is the tool par excellence in the study of multifractals. In this spectrum, a monofractal plots as a single spike because it has only one singularity associated with it. Examples of multifractal spectra for geoscience data



Fractal Geometry in Geosciences, Fig. 8 Relationships between perimeters and anomalous areas using contours with different values greater than thresholds for (a) Au, (c) Cu, and (e) As. Log-log plots on right side show relationships between estimated lengths of the perimeters

of the anomalous areas for (c) Au, (d) Cu and (f) As with variable yardstick. D1 and D2 are for textural and structural fractals dimensions, respectively. All solid lines were obtained by method of least square. (Source: Cheng et al. 1994, Fig. 8)

include Gonçalves et al. (2001), Pahani and Cheng (2004), and Arias et al. (2012).

A 1-D example of a multifractal is as follows. Suppose that $\mu = x \cdot \epsilon$ represents total amount of a metal for a line segment of length ϵ and x is the metal's concentration value. In the multifractal model, it is assumed that (1) $\mu \propto \epsilon^\alpha$ where $\propto$ denotes proportionality and α is the singularity exponent corresponding to x and (2) $N\alpha(\epsilon) \propto \epsilon^{-f(\alpha)}$ represents total number of line segments of length ϵ with concentration value x , and $f(\alpha)$ is the fractal dimension of these line segments.

There are three different ways in which a multifractal spectrum can be computed: histogram method, method of moments, and direct determination. The first two methods are described in Evertsz and Mandelbrot (1992). The histogram method is intuitively appealing because it is easy to understand. However, in practice it is better to use the method of moments with Legendre transform to be discussed here. This method produces better results faster than the histogram method although the latter can be useful in the study of frequency distributions of multifractals. The third method (direct determination) was developed by Chhabra and Jensen (1989).

Evertsz and Mandelbrot (1992) make a clear distinction between (1) Hölder exponent or singularity $\alpha(x) = \lim_{\epsilon \rightarrow 0} \frac{\log_\epsilon \mu\{B_\epsilon(x)\}}{\log_\epsilon \epsilon}$ at a point x and (2) "coarse" Hölder exponent $\alpha = \frac{\log_\epsilon \mu\{B_\epsilon(x)\}}{\log_\epsilon \epsilon}$ measured for a volume $B_\epsilon(x)$ around the point x . The mass partition function is: $\chi_q(\epsilon) = \sum_{N(\epsilon)} \mu_i^q = \int N_\epsilon(\alpha) (\epsilon^\alpha)^q dx$ with $N_\epsilon(\alpha) = \epsilon^{-f(\alpha)}$ where $f(\alpha)$ represents fractal dimension. It follows that $\lim_{\epsilon \rightarrow 0} \chi_q(\epsilon) = \int \epsilon^{q\alpha - f(\alpha)} d\alpha$ keeping in mind that $\alpha (= \alpha(q))$ also is a function of q . At the extremum, $\frac{\partial \{q\alpha - f(\alpha)\}}{\partial \alpha} = 0$ and $\frac{\delta f(\alpha)}{\delta q} = q$ for any moment q . Writing $\tau(q) = q\alpha - f(\alpha)$, it follows that $\frac{\delta \tau(q)}{\delta q} = q$. The multifractal spectrum satisfies $f(\alpha) = q\alpha - \tau(q)$. If the multifractal spectrum $f(\alpha)$ exists, the so-called codimension satisfies $C_1 = f(\alpha) - E$ where E is the Euclidian dimension (cf. Evertsz and Mandelbrot 1992).

In practice, a feature such as element concentration in rock samples is measured in blocks of different sizes. The mass partition function $\chi(\epsilon, q)$ then is the sum of all measurements raised to the power q for blocks with cell edge ϵ . Terms such as "mass partition function" were borrowed from physical chemistry (see, e.g., Evertsz and Mandelbrot 1992). The slopes of the lines on a log-log plot of partition function against size measure are estimates of the mass exponent $\tau(q)$. The singularity $\alpha(q)$ is the first derivative of $\tau(q)$ with respect to q , and the fractal dimension satisfies $f(\alpha) = q \cdot \alpha(q) - \tau(q)$.

A multifractal case history study in 2-D was provided by Cheng (1994) for gold in the Mitchell-Sulphurets area. The example of Fig. 9 shows how the multifractal spectrum was obtained for 100 ppb Au cutoff. Different grids with square

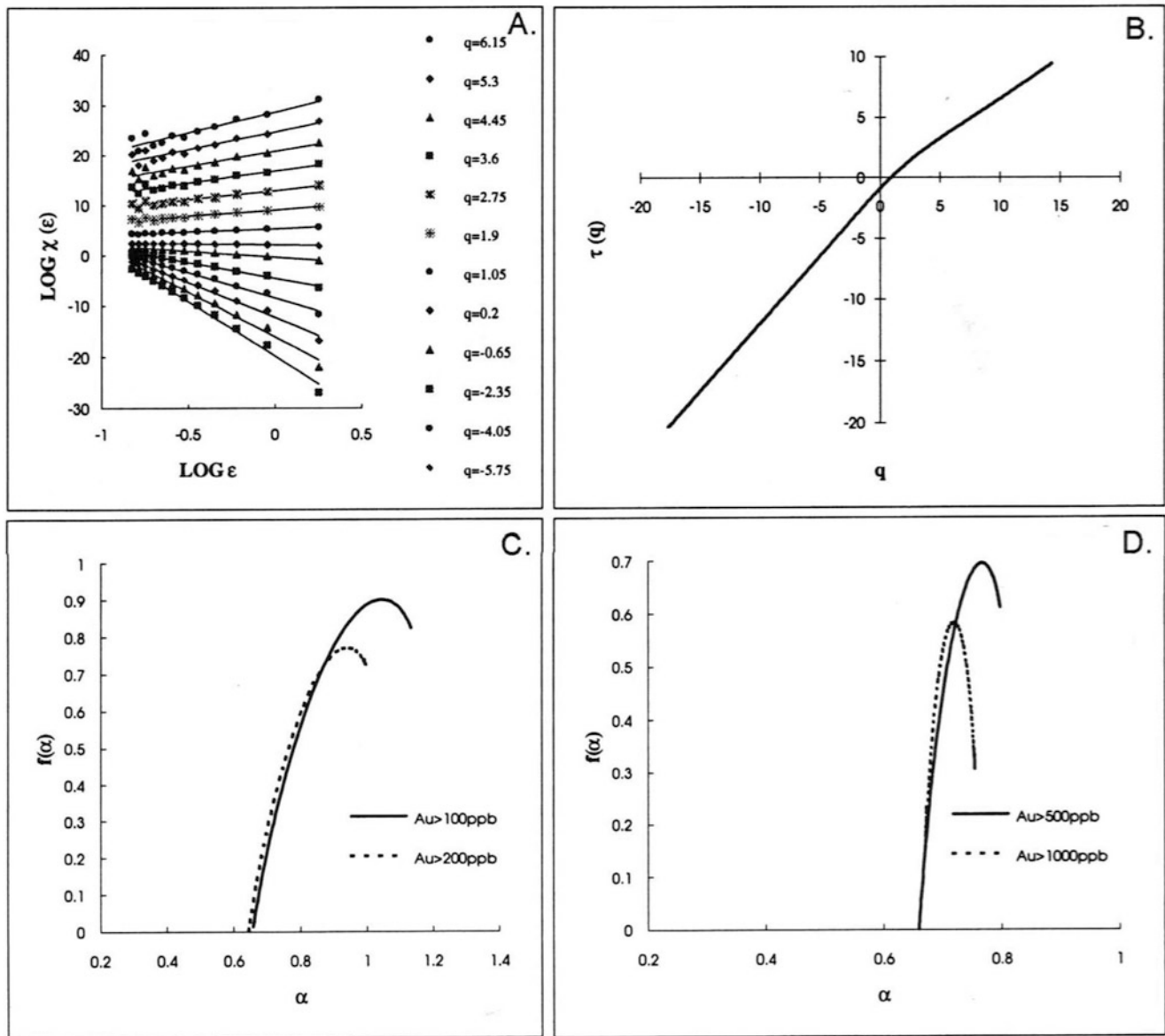
cells measuring ϵ on a side were superimposed on the study area. The average gold value was determined for each grid cell containing one or more samples with Au > 100 ppb. Each average value was raised to the power q . The sum of the powered values satisfies a straight-line relationship for any q in a multifractal model. This aspect is verified in Fig. 9a. The slopes of many best-fitting straight lines (including those in Fig. 9a) are shown as the function $\tau(q)$ in Fig. 9b. The first derivative of $\tau(q)$ with respect to q gives $\alpha(q)$, and the multifractal spectrum $f(\alpha)$ follows from the relation $f(\alpha) = q \cdot \alpha(q) - \tau(q)$ (solid lines in Fig. 9c and d).

The multifractal nature of the gold deposits in the Mitchell-Sulphurets area is shown in the spectra for different cutoff values in Fig. 9c and d. A fractal model would have resulted in a spectrum consisting of a single spike characterized by two constants: the fractal dimension $f(\alpha)$ and the singularity α . The four spectra of Fig. 9c and d are approximately equal on the left side, which is representative of the largest concentration values. The point where the multifractal spectrum reaches the α -axis and the slope of the curve at this point together determines the approximate area-concentration power-law relation on the right side in Fig. 9d (Agterberg et al. 1993; Cheng et al. 1994). The maximum value of $f(\alpha)$ in Fig. 9c and d decreases with increasing cutoff value. In general, if $f(\alpha) < 2$ in 2-D space, it can be assumed that the multifractal measure is defined on fractal support (cf. Feder 1988).

Singularity Analysis

Singularity analysis is a relatively simple method based on fractal geometry (cf. Cheng 2007; Cheng and Agterberg 2009). The following examples illustrate the advantages of this approach. Figure 10 is the Gejiu Mineral District example (Cheng and Agterberg 2009). This area of about 4000 km² contains 11 large tin deposits (shown as triangles in Fig. 10a). Several of these, including the Laochang and Kafang deposits, are tin-producing mines with copper extracted as a by-product. Eight of these occur in the eastern part of the area where there is considerable pollution related to mining. Cheng and Agterberg (2009) applied singularity analysis to data from 1,056 stream sediment samples collected as part of the Chinese regional geochemistry reconnaissance project (Xie et al. 1997).

Original sample locations were 2 km apart both in the north-south and east-west directions. Twelve of these sampling points (see Fig. 10b) were used, for example, to illustrate the calculations involved in singularity analysis. Only two of these examples are reproduced here in Fig. 11. Tin concentration values from within the same 26 × 26 km² cells each contain 169 sampling points. The singularities can be estimated by fitting straight lines on log-log plots of either the tin concentration value (x) versus ϵ using $x = c \cdot \epsilon^{-\alpha/2}$, or



Fractal Geometry in Geosciences, Fig. 9 Construction of multifractal spectrum of gold in Mitchell-Sulphurets mineral district (see Fig. 5). (a) $\chi(\epsilon)$ represents sum of gold values greater than 100 ppb averaged for cells and raised to the power q ; cell edge ϵ in km; log base 10; lowest sets of solid triangles and squares are for q equal to -7

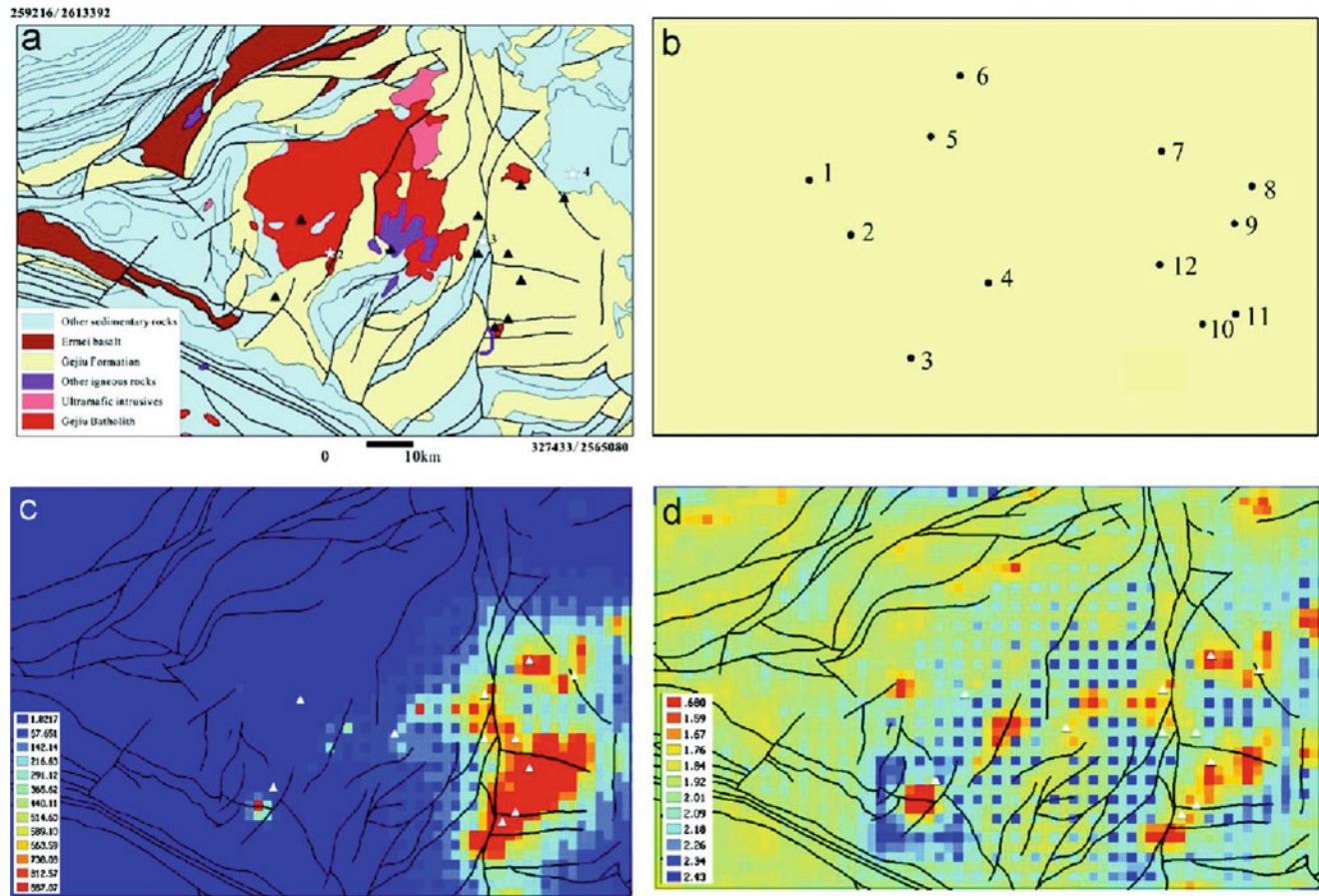
0.45 and -9.15 , respectively. (b) Slopes $\tau(q)$ of best-fitting straight lines including those shown in a as function of q . (c and d) Multifractal spectra for four different cut-off values. It is noted that, contrary to a multifractal, a fractal is characterized by a single dimension that would have a spectrum consisting of a spike (Source: Agterberg 1995, Fig. 4)

amount of metal $\mu = x \cdot \epsilon^{-2}$ versus ϵ using $\mu = c \cdot \epsilon^{-\alpha}$. In Fig. 10d, both methods are yielded similar estimates of the singularity α . The resulting pattern shows a number of relatively small anomalies, some of them approximately coinciding with large tin deposits.

A moving average map (Fig. 11c) was constructed for the same stream sediment samples using weighted averages of the 169 tin concentration values in surrounding squares. The inverse squared distance was used for each weight. Like ordinary kriging or using another moving average method (cf. Figure 1), this method yields a generalized picture in

which local variability is smoothed out. The pattern of Fig. 11c shows pollution of water in streams related to mining in the eastern part of the Gejiu area instead of the localized anomalies in Fig. 11d that may provide clues on occurrences of large tin deposits that have not yet been discovered.

Another example of local singularity analysis is shown in Figs. 12 and 13 after Xiao et al. (2012) for Ag and Pb-Zn deposits in northwestern Zhejiang Province, China. The southeastern part of this study area is relatively poorly explored but stream sediment survey data were available for the entire study area. Local singularity analysis as used for tin



Fractal Geometry in Geosciences, Fig. 10 Geological setting of tin deposits in Gejiu area and map patterns derived from tin concentration values in stream sediment samples. (a) Simplified geology after Yu (2002). Solid lines indicate faults; triangles represent tin deposits; UTM coordinates apply to map corner points. (b) Locations examples of estimation of local tin singularity in Fig. 11. (c) Distribution of tin

concentration values in 1000 stream sediment samples using inverse distance weighted moving average. Eight of 11 large tin deposits fall on large anomaly in eastern part of area that is due to environmental pollution related to mining. (d) Distribution of local tin singularities. Tin deposits tend to occur in places where local singularity is less than 2 (Source: Cheng and Agterberg 2009, Fig. 1)

in the Gejiu example was applied to lead in stream sediments in this other example clearly outlining potential new target areas for further exploration.

Model of de Wijs

Mandelbrot (1983) has pointed out that a model introduced by de Wijs (1951) was the first example of application of fractal geometry in geoscience. This model for metal concentration values in blocks of rock was originally developed as follows: Suppose any block with metal concentration model ζ is repeatedly divided into halves with concentration values $(1 + d)\zeta$ and $(1 - d)\zeta$ where d is the so-called coefficient of dispersion. Then, the frequency distribution for metal concentration values in increasingly smaller blocks satisfies the so-called negative binomial distribution that rapidly approaches lognormal form. If there are p subdivisions, the

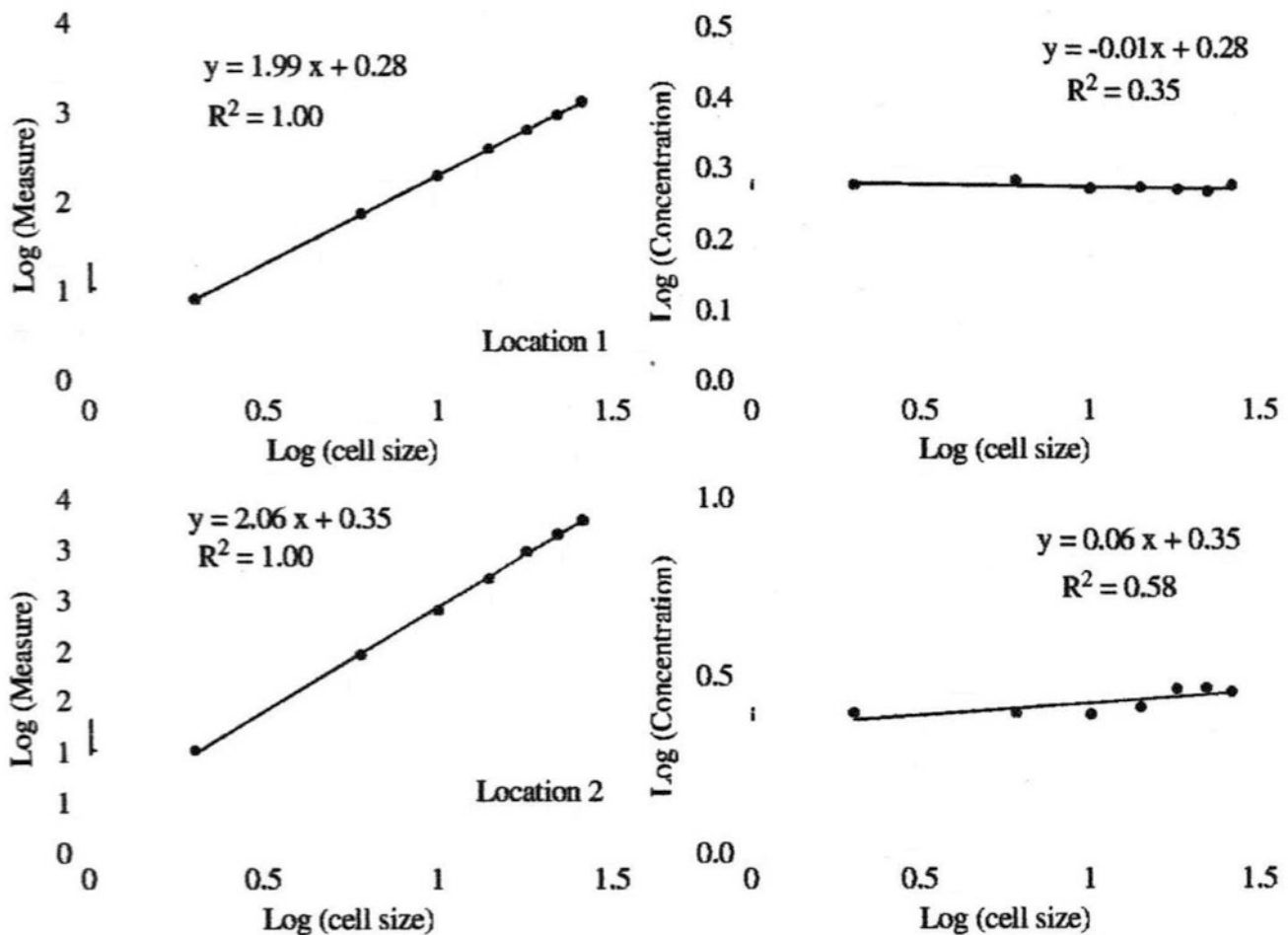
log-binomial distribution of the $\binom{p}{K}$ concentration values of the resulting $n = 2^p$ blocks is

$$X(p, K) = \zeta \cdot (1 + d)^{p-K} (1 - d)^K$$

where K satisfies the binomial distribution with $\mu(K) = p/2$ and variance $\sigma^2(K) = p/4$ (cf. Agterberg 1974, p. 322). This log-binomial has mean $\mu(X) = \zeta$ and variance $\sigma^2(X)$ approaching to:

$$\sigma^2(X) = \frac{p}{4} \cdot \left[\ln \frac{1-d}{1+d} \right]^2$$

A two-dimensional example of application of the model of de Wijs is shown in Fig. 14 for $p = 14$ and $d = 0.4$ (from Agterberg 2007). Here, the overall mean value was set equal to 1 and the blocks were taken to be squares, and at each



Fractal Geometry in Geosciences, Fig. 11 Examples of estimation of local singularity at locations shown in Fig. 10b. Cell size is length of side of square cells with half-widths set equal to 1, 3, 5, ..., 13 km. Plots on left and right show relations between values of μ and ξ versus cell size, respectively. Value of μ is sum of tin concentration values in all $2 \text{ km} \times 2 \text{ km}$ cells contained within the larger cells centered on a location with value of $\xi = \mu/\epsilon^2$; base of logarithms is 10. Straight lines were fitted

by ordinary least squares method (log of value regressed on log of cell size); R^2 = multiple correlation coefficient squared. Typically, all seven points fall on the best-fitting straight line although small deviations do occur (e.g., second point at Location 2). Slopes of straight-lines on left provide estimates of local singularity α and those on the right are for $\alpha - 2$ (Source: Cheng and Agterberg 2009, Fig. 2)

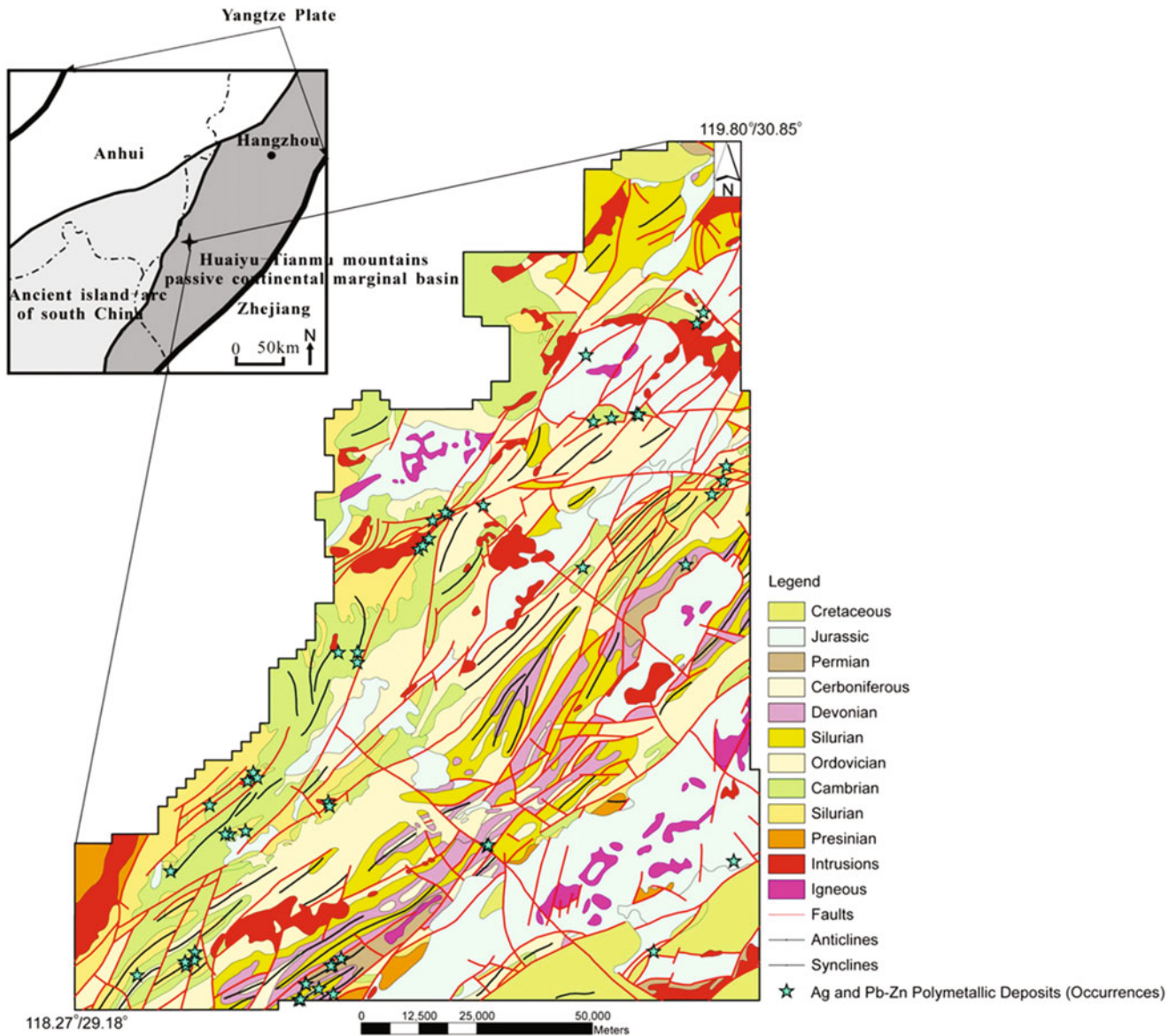
iteration a square block was subdivided into 4 blocks with concentration values $(1-d)^2$, $(1-d) \cdot (1+d)$, $(1+d) \cdot (1-d)$, and $(1+d)^2$, respectively. Values greater than 4 were truncated to more clearly display most of the spatial variability. Later, Mandelbrot (1989) clarified that the model of de Wijs produces a multifractal instead of a monofractal. It offers a simple explanation for the widespread occurrence of lognormality.

Multifractal Spatial Correlation

De Wijs (1951) developed his original model on the basis of 118 zinc concentration values for 2-m channel samples from a drift in the Pulacayo Mine in Bolivia. This data set has been subjected to many different kinds of statistical analysis including Matheron (1962), Agterberg (1974), Cheng and

Agterberg (1996), Chen et al. (2007), Lovejoy and Schertzer (2007), and Cheng (2014). Several of these different applications are described in Agterberg (2014) which also includes good results obtained by means of Whittle (1962)'s 3-D correlogram method. Figure 15 shows the original data together with the results of a signal-plus-noise model, which is close to what would be obtained by kriging or cubic spline-curve fitting. These other techniques are basically different from the fractal-multifractal methods because they assume that the true values to be estimated are masked by white noise that is to be eliminated.

Cheng (1994; also see Cheng and Agterberg 1996) derived general equations for the spatial covariance, correlogram, and semi-variogram of any scale-independent multifractal including the model of de Wijs. This model is for sequences of adjoining blocks along a line. Experimentally, their



Fractal Geometry in Geosciences, Fig. 12 Geological map of northwestern Zhejiang Province, China (Source: Xiao et al. 2012, Fig. 1)

semi-variogram resembles Matheron's semi-variogram for infinitesimally small blocks. Extrapolation of the corresponding spatial covariance function to infinitesimally small blocks would yield infinitely large variance when h approaches zero. The autocorrelation function of a multi-fractal satisfies:

$$\rho_k(\epsilon) = \frac{C\epsilon^{\tau(2)-2}}{2\sigma^2(\epsilon)}[(k+1)^{\tau(2)+1} - 2k^{\tau(2)+1} + (k-1)^{\tau(2)+1}] - \frac{\xi^2}{\sigma^2(\epsilon)}$$

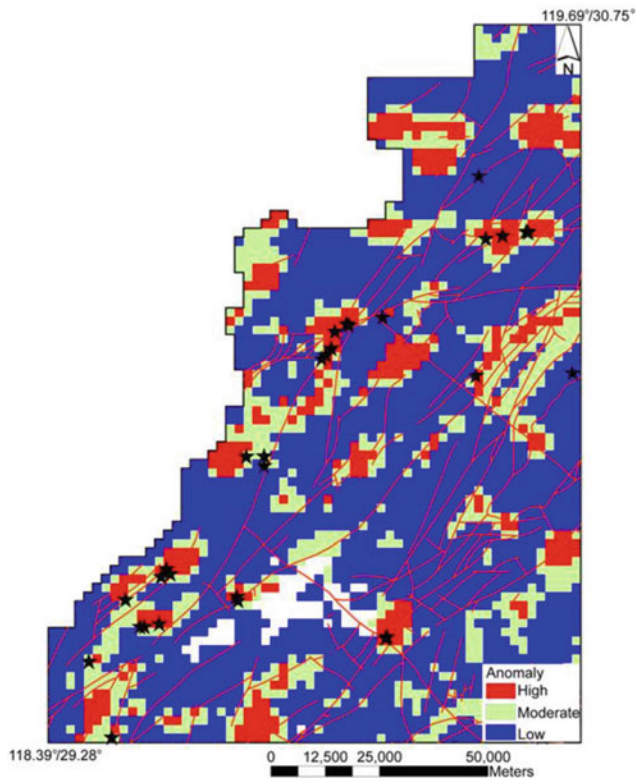
where C is a constant, ϵ represents length of line segment for which an average zinc concentration value is assumed to be representative, $\tau(2)$ is the second-order mass exponent, ξ represents overall mean concentration value, and $\sigma^2(\epsilon)$ is the variance of the zinc concentration values. The unit interval ϵ is

measured in the same direction as the lag distance h . The index $k = \frac{1}{2}h$ is an integer value.

The Pulacayo orebody provides an example. The most important parameter on which this approach is based is the second-order mass exponent $\tau(2)$. In this application, it is estimated as the slope of the straight line fitted by least squares in Fig. 16 with $\tau(2) = 0.979 \pm 0.019$. Estimation for the 118 Pulacayo zinc values using an ordinary least squares model with $\tau(2) = 0.979$ gave:

$$\hat{\rho}_k = 4.37[(k+1)^{1.979} - 2k^{1.979} + (k-1)^{1.979}] - 8.00$$

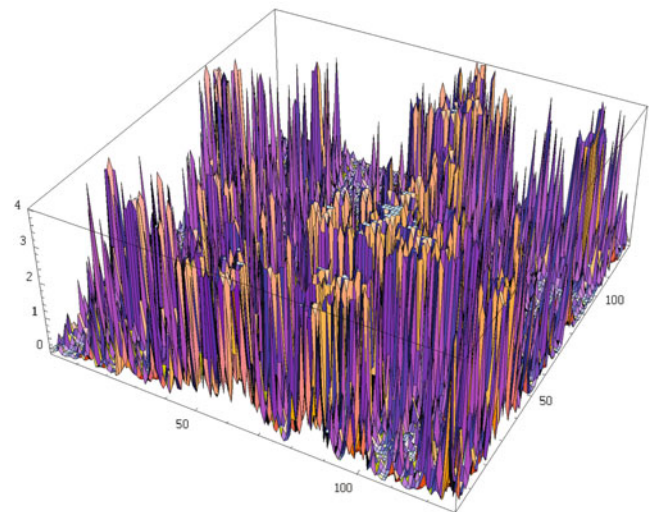
The semi-exponential autocorrelation previously used for smoothing the zinc values of Fig. 15 is shown as a broken line in Fig. 17 together with the autocorrelation coefficients (for



Fractal Geometry in Geosciences, Fig. 13 Raster map for Fig. 12 showing prospecting target areas for Ag and Pb-Zn deposits delineated by comprehensive singularity anomaly method (cell size is 2 km × 2 km). Known Ag and Pb-Zn deposits are indicated by stars (Source: Xiao et al. 2012, Fig. 13)

lag distance $h > 0$) to which it was fitted. Its “nugget effect” at the origin would explain about half of total variability of the zinc values. On the other hand, use of the multifractal correlogram (solid line in Fig. 17) shows a continuous increase of spatial correlation toward the origin. The method used for fitting the multifractal correlogram does not apply when the lag distance becomes very small ($h < 0.07$) in Fig. 17 so that the true white noise at the origin cannot be estimated by this method. By means of local singularity mapping, white noise can be estimated to represent only about 2% of total variability of the zinc values. It represents measurement error and strong decorrelation at microscopic scale.

Figure 18a and b show the multivariate semi-variogram that satisfies the equation: $\gamma_k(\epsilon) = \xi_2(\epsilon) \left[1 - \frac{1}{2} \left\{ (k+1)^{\tau(2)+1} - 2k^{\tau(2)+1} + (k-1)^{\tau(2)+1} \right\} \right]$ with $\tau(2) = 0.979$ and $\xi_2(\epsilon) = 391.49$ in comparison with experimental semi-variogram values estimated from the 118 Pulacayo zinc values using arithmetic and log-log scales, respectively. As shown by Cheng and Agterberg (1996, Eq. (23)), if $\tau(2)$ is only slightly less than 1, the preceding theoretical equation can be approximated by $\gamma_k(\epsilon) = -\xi_2(\epsilon) \cdot \epsilon^{\tau(2)-1} \log_e \left[\frac{1}{2} \{ \tau(2) + 1 \} \tau(2) k^{\tau(2)-1} \right]$. Figure 18c



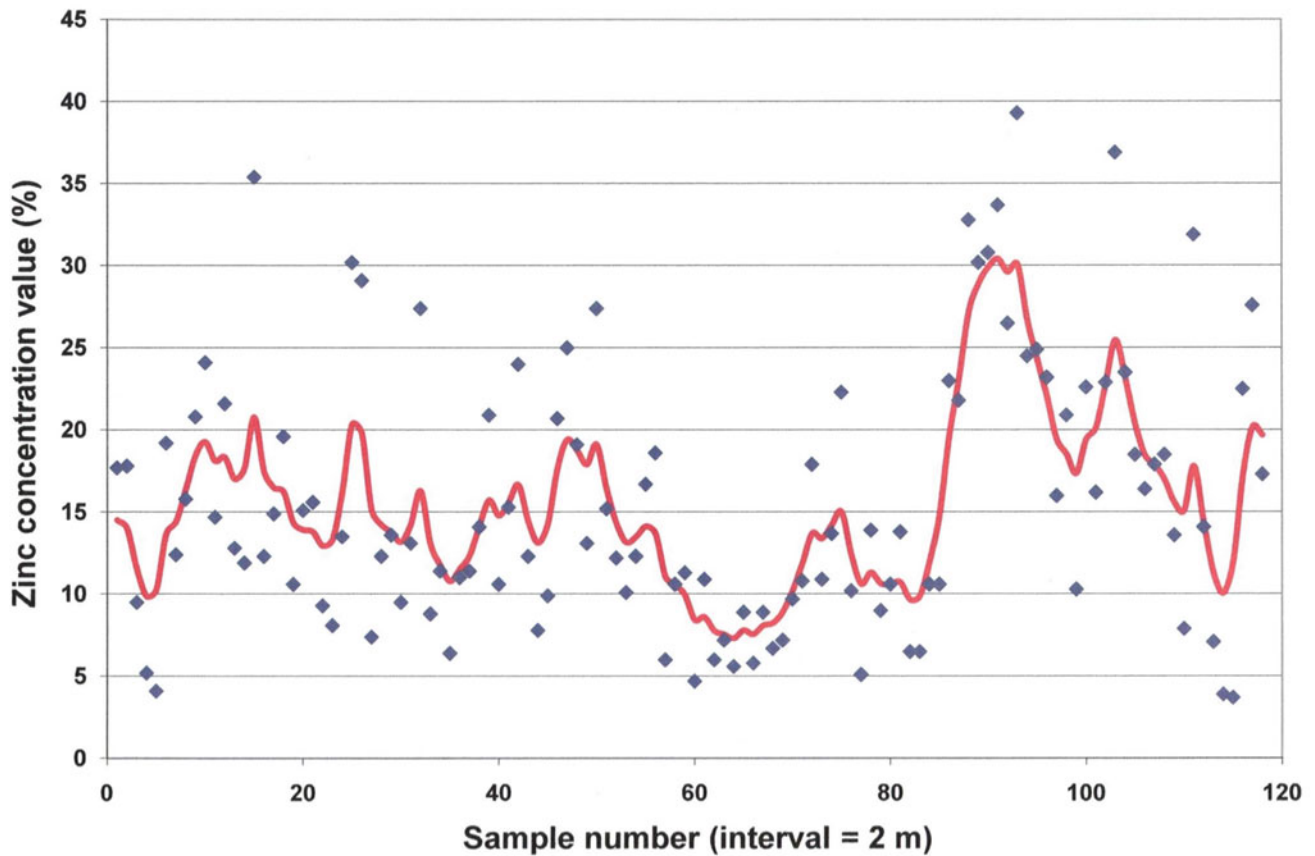
Fractal Geometry in Geosciences, Fig. 14 Realization of model of de Wijs (in 2-D for $d = 0.4$ and $N = 14$). Overall average value is equal to 1. Values greater than 4 were truncated (Source: Agterberg 2007, Fig. 3)

shows the experimental semi-variogram values on a graph with logarithmic scale in the horizontal direction only. The straight line in Fig. 18c was fitted by least squares. This approximate semi-variogram model provides a good fit. It is almost exactly equal to the logarithmic semi-variogram originally obtained by Matheron (1962, p. 180; also see his Table 6.1) who expanded his version of the model of de Wijs incorporating the geometry of channel samples taken at regular distance along the Pulacayo mining drift.

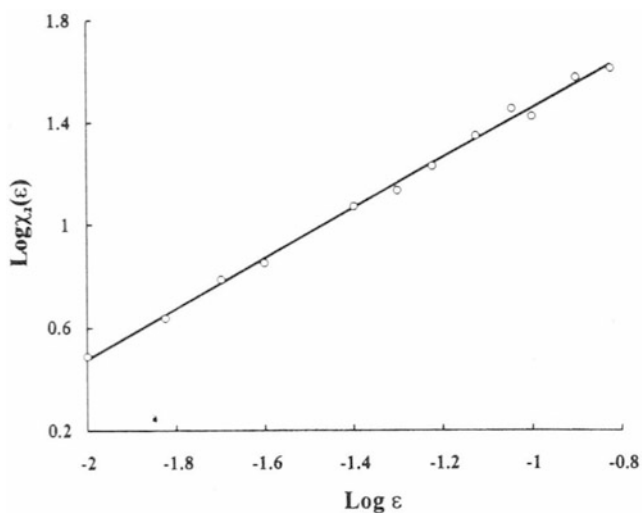
The example of Figs. 16, 17, and 18 shows analogy with the approach taken in singularity analysis. In both approaches, the assumption that there is a nugget effect is replaced with the assumption that autocorrelation becomes stronger over very short distances instead of being reduced to zero. This fundamental difference in approach is discussed in more detail in Agterberg (2014).

Fractal/Multifractal Modeling of Point Patterns

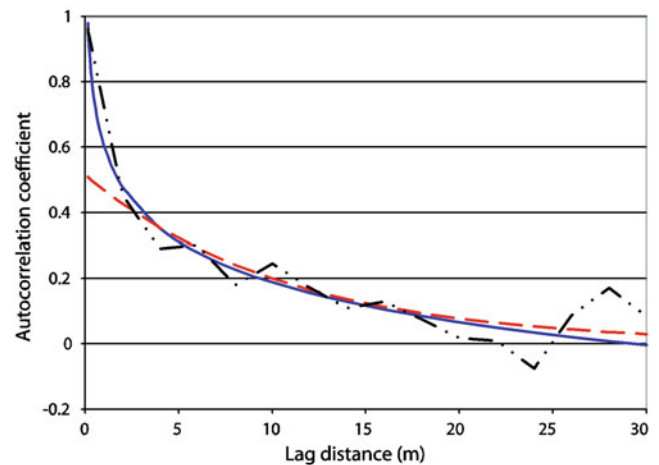
Fractal and multifractal models can be used for pattern analysis of locations of mineral deposits on maps. Traditionally, two-dimensional point location distributions for mineral deposits were considered to satisfy classical statistical (primarily Poisson or negative binomial) distribution models with constant means and variances. However, evidence has been increasing that the means and variances depend on size of measurement area used for counting number of deposit points contained within it. Fractal and multifractal models both result in power-law relations between point density (number of points per unit of area) and size of study area. For fractal point patterns, the power-law exponent provides an estimate of the cluster dimension (*cf.* Feder 1988).



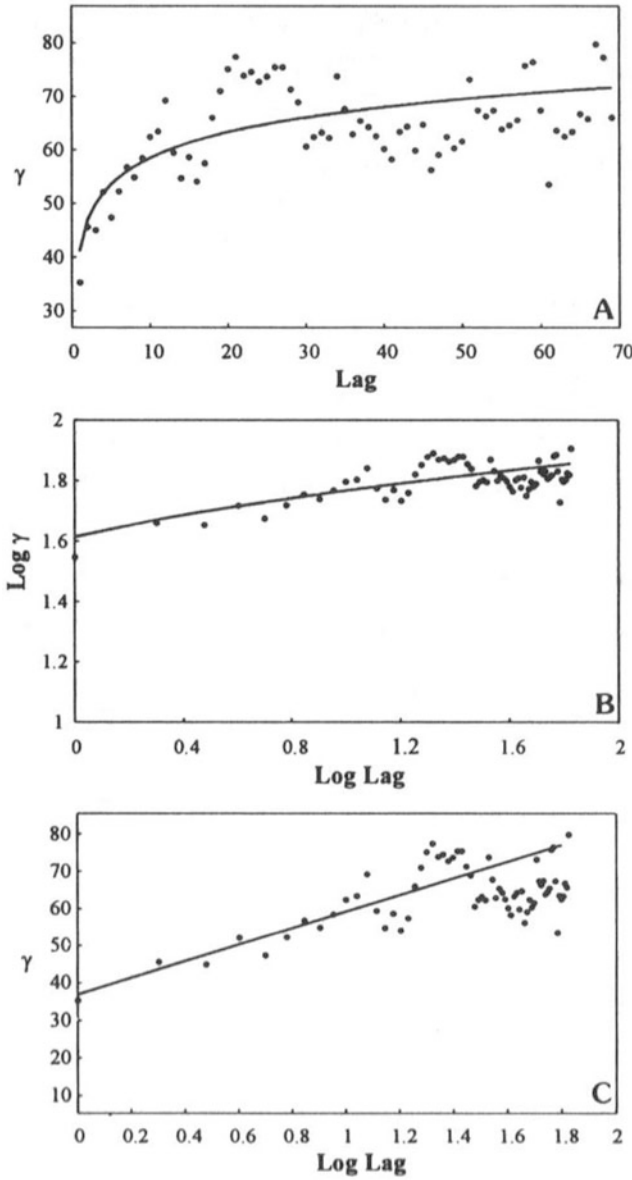
Fractal Geometry in Geosciences, Fig. 15 Pulacayo Mine zinc concentration values for 118 channel samples along horizontal drift (original data from De Wijs 1951). Solid curve is “Signal” retained after filtering out “Noise” (Source: Agterberg 2012, Fig. 1)



Fractal Geometry in Geosciences, Fig. 16 Pulacayo zinc example. Log-log plot for relationship between $\chi_2(\epsilon)$ and ϵ (Source: Cheng and Agterberg 1996, Fig. 3)



Fractal Geometry in Geosciences, Fig. 17 Estimated autocorrelation coefficients (partly broken line) for 118 zinc concentration values of Fig. 15. Broken line represents best-fitting semi-exponential function used to extract “signal” in Fig. 15. Solid line is based on multifractal model that assumes continuance of self-similarity over distance less than the sampling interval (Source: Agterberg 2012, Fig. 3)



Fractal Geometry in Geosciences, Fig. 18 Estimates of semi-variogram γ_k . A, Solid line obtained by linear regression after setting $\tau(2) = 0.979$. B, Log-log plot of A. C, Logarithmic approximation (Source: Cheng and Agterberg 1996, Fig. 5)

A different approach is needed to determine if a point pattern is multifractal instead of fractal. The purpose of this section is to present a simplified account of multifractal pattern analysis as it was first proposed by Cheng (1994). Later this novel approach was also published in Cheng and Agterberg (1995).

Suppose that A is a rectangular area in two-dimensional space with size $n \in_1 \times n \in_2$ where $\in_1 = k \in$ and $\in_2 = l \in$, with $\in$ being a unit distance. Consequently, A can be divided in $n \cdot m$ subareas $A_{ij} (\in_1, \in_2)$. The partition function of the number of points $N\{A_{ij} (\in_1, \in_2)\}$ within $A_{ij} (\in_1, \in_2)$ is

$$\chi_q(\epsilon_1, \epsilon_2) = \sum_{i=1}^n \sum_{j=1}^m [N\{A_{ij}(\epsilon_1, \epsilon_2)\}]^q$$

for $-\infty \leq q \leq \infty$. For multifractal $N\{A_{ij} (\in_1, \in_2)\}$, the partition function has a power-law relation with the cell sizes for any q , or

$$\chi_q(\epsilon_1, \epsilon_2) \propto \epsilon_1^{\tau_1(q)} \cdot \epsilon_2^{\tau_2(q)}$$

For a square cell with side $\in = \in_1 = \in_2$ ($k = l$), it follows that

$$\chi_q(\epsilon) \propto \epsilon^{\tau(q)}; \quad \tau(q) = \tau_1(q) + \tau_2(q)$$

The singularity exponent $\alpha(q)$ and multifractal spectrum $f(\alpha) = f\{\alpha(q)\}$ can be obtained successively by

$$\alpha(q) = \frac{d\tau(q)}{dq}; f(\alpha) = q\alpha(q) - \tau(q)$$

Setting $q = 2$, it follows that

$$E[N^2\{A_{ij}(\epsilon_1, \epsilon_2)\}] = c(k\epsilon)^{\tau_1(2)+1} (l\epsilon)^{\tau_2(2)+1}$$

where c is a constant and E denotes mathematical expectation. For an isotropic process

$$\tau_1(2) = \tau_2(2) = \tau(2)/2.$$

Point processes have been studied in spatial statistics (Ripley 1976; Diggle 1983; Cressie 1991; Stoyan and Stoyan 1994). Suppose x and y denote two points in the plane and $r = \|x-y\|$ represents the distance between them. The first-order intensity function (λ) and the second-order intensity function (λ_2) are used to describe isotropic, stationary point patterns. Although originally developed for patterns with an average point density that is independent of size of study area, the two intensity functions (λ and λ_2) can also be used for nonstationary multifractal point patterns. Diggle (1983) has defined the function $K(r) = \lambda^{-1} \cdot E$ (number of additional points within distance r from an arbitrary point). The expected number of additional points within distance r from an arbitrary point at 0 then satisfies

$$\lambda K(r) = 2\pi\lambda^{-1} \int_0^r \lambda_2(r) r dr$$

Conversely

$$\lambda_2(r) = \lambda^2 (2\pi r)^{-1} \frac{dK(r)}{dr}$$

The practical advantage of using $K(r)$ is that Ripley's

estimator (Ripley 1976; Diggle 1983) can be used to correct for edge effects that generally are strong in two-dimensional space. If r_{ij} represents the distance between two points labeled i and j within A , the edge effect corrected estimate of $K(r)$ satisfies

$$\hat{K}(r) = \frac{|A|}{N(A)^2} \sum_{i \neq j} w_{ij}^{-1} I_i(r_{ij})$$

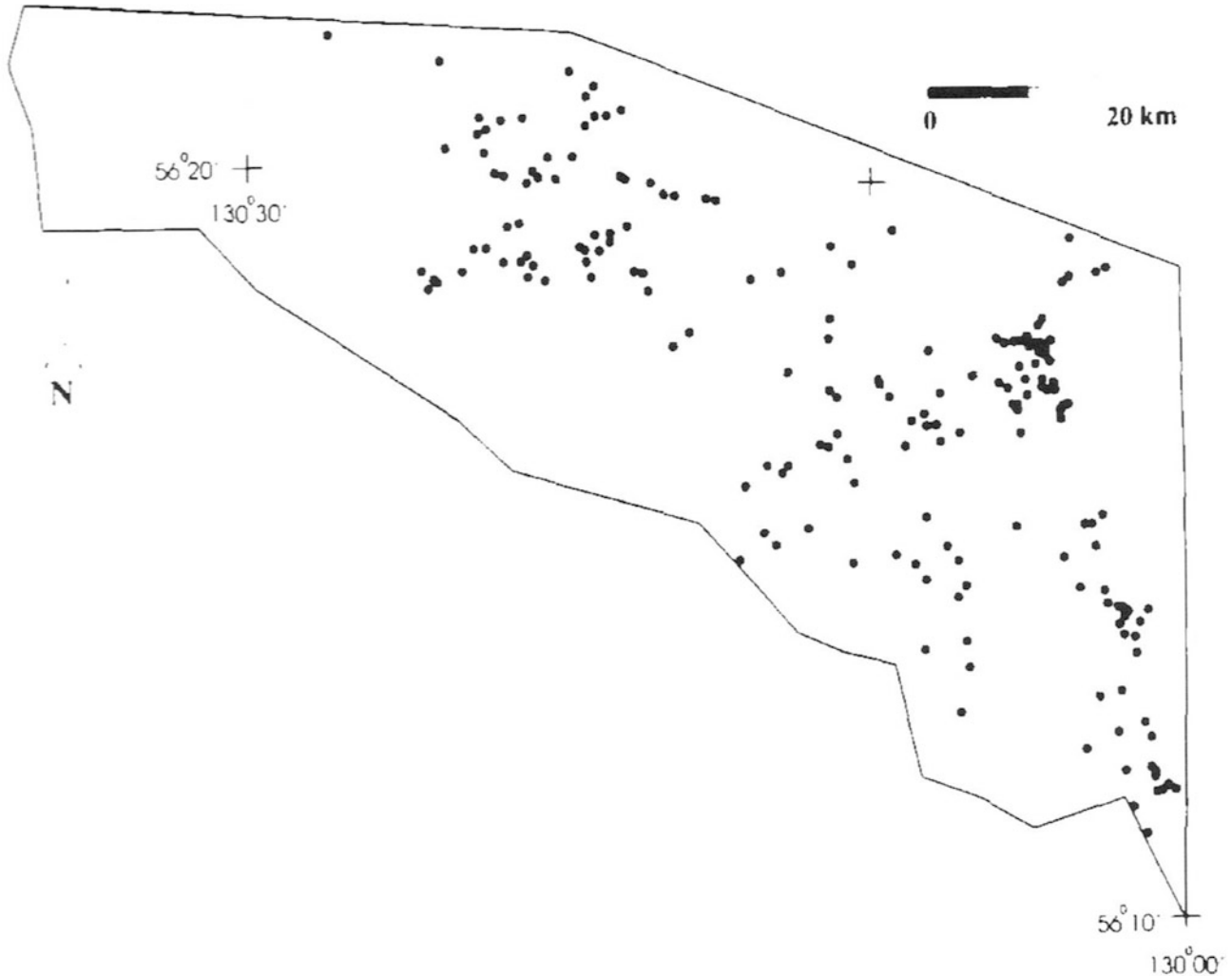
where $I_r(r_{ij})$ is an indicator function that assumes the value 1 if $r_{ij} < r$; 0 otherwise. The weight w_{ij} represents the proportion of the circumference of the circle around point i with radius r_{ij} that lies within A . In Cheng and Agterberg (1995), it is shown that setting $r = k \in \sqrt{2}$ yields the approximate power-law relation

$$\lambda_2(r) \cong Cr^{\tau(2)-2}$$

where C is a constant. For sufficiently small ϵ

$$K(r) \cong C_2 r^{\tau(2)}; \quad C_2 = \left[\frac{|A|}{\mu(A)} \right]^2 \cdot \frac{2\pi C}{\tau(2)}.$$

Figure 19 shows the first geochemical exploration example to which the preceding technique was applied. It consisted of a set of 175 Au occurrences in an irregularly shaped area on the Iskut River map sheet (Minfile, Map 104B, 1989), north-western British Columbia (Cheng and Agterberg 1995). In order to use the multifractal model, cell sizes (ϵ) ranging from 3 to 20 km were used. Au occurrences were counted on eight gridded maps. Care was taken to account for edge effects. Figure 20a shows five plots of $\log \chi_q(\epsilon)$ versus $\log \epsilon$ with integer values of q ranging from 0 to 4, from which the experimental relationship between $\tau(q)$ and q , as well as $f(\alpha)$ against α were obtained (see Fig. 20b and c).



Fractal Geometry in Geosciences, Fig. 19 Gold mineral occurrences in Iskut River map sheet, northwestern British Columbia (from B.C. Minfile Map 104B, 1989) (Source: Cheng and Agterberg 1995, Fig. 4)

For example, three estimates of $\tau(q)$ accompanied by their 95% confidence intervals are $\hat{\tau}(0) = -1.34 \pm 0.15$, $\hat{\tau}(2) = 1.22 \pm 0.07$, and $\hat{\tau}(4) = 3.07 \pm 0.19$, respectively. As explained in Cheng and Agterberg (1995), the fractal model would have resulted in the following linear relationships:

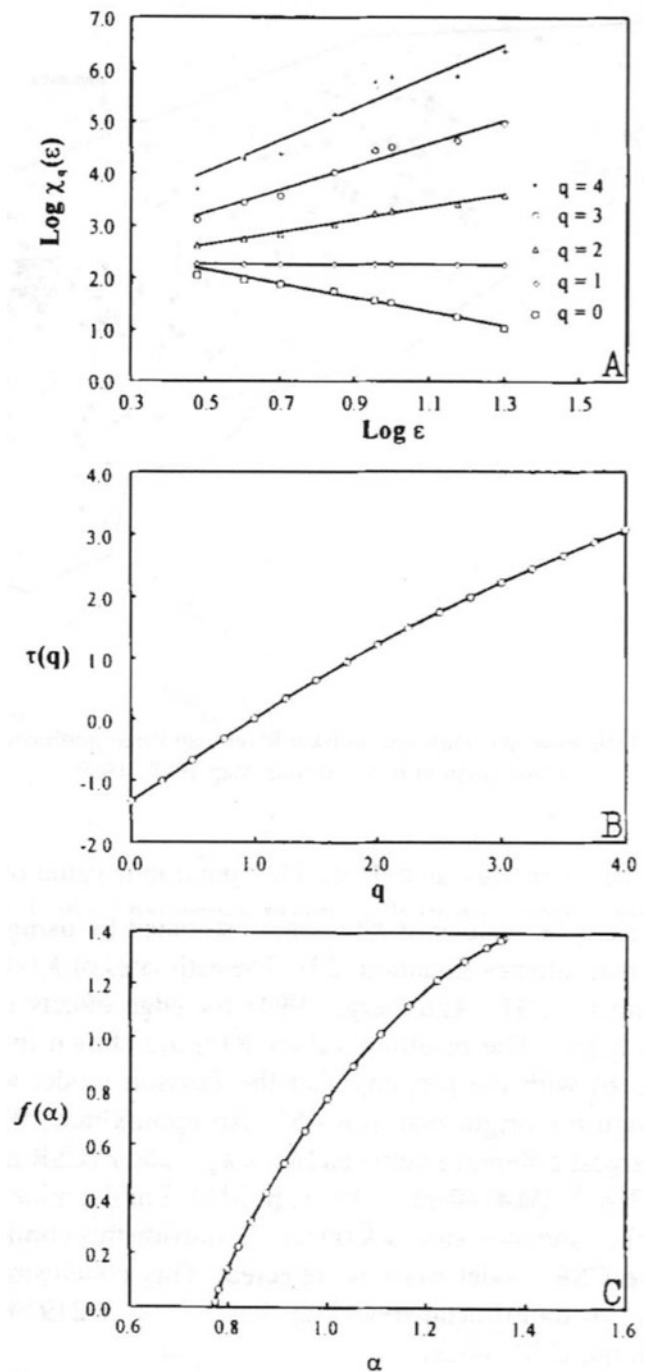
$$\frac{\tau(q)}{q-1} = \frac{\tau(p)}{p-1}.$$

Consequently, for $q = 4$ and $p = 0$, the fractal model would have resulted in $\tau(4)/3 + \tau(0) = 0$. However, from the preceding considerations, it follows that $\hat{\tau}(4)/3 + \hat{\tau}(0) = 0.31 \pm 0.10 \neq 0$ proving that the Au occurrence point pattern is multifractal instead of monofractal. Further results for the example of Fig. 20 are shown in Fig. 21.

Another example of this approach described in Agterberg (2013, 2014) was based on occurrences of (1) volcanogenic massive sulfide (“Cyprus + Kuroko”) deposits, (2) porphyry copper deposits, and (3) podiform chromium deposits in many different “permissive tract” areas in the world (definitions and data according to Singer and Menzie 2010, Table 4.1). Deposit density (= number of deposits divided by permissive tract area) plotted against permissive tract area produces a linear pattern on the log-log plot for each of these three deposit types (*cf.* Singer and Menzie 2010, Fig. 4.5). Normally, one would expect deposit density to be independent of size of permissive tract area, and for stationary point processes the slopes of best-fitting lines on plots of this type would be zero. However, slopes of the three lines fitted by least squares are -0.62 (38 permissive tracts for massive sulfides), -0.63 (33 for porphyry coppers), and -0.53 (28 for chromium deposits), respectively. Singer and Menzie demonstrated that these estimated slopes are significantly different from zero pointing out that size of tract area can be used to predict deposit density. The correlation coefficients for these three linear associations (-0.801 , -0.788 , and -0.709 , respectively) are significantly less than zero. Agterberg (2013, 2014) assumed that the observed negative slope estimates could best be interpreted as estimates of $2 - D_c$ with D_c representing fractal cluster dimension. According to the statistical model described in this section, however, a multifractal explanation with $\hat{\tau}(2)$ values of 1.38 (volcanogenic massive sulfides), 1.39 (porphyry copper), and 1.47 (chromium deposits) provides a possible explanation as well.

Lognormal and Pareto Size-Frequency Distributions for Mineral Resources

The International Union of Geological Sciences has outlined actions required by the geoscience community pertaining to the comprehensive evaluation and future supply of mineral



Fractal Geometry in Geosciences, Fig. 20 Multifractal analysis results for Au mineral occurrences of Fig. 19. A, Log-log plot of mass-partition function for selected values of q , dots represent estimated values, solid lines obtained by LS fitting. B, $\tau(q)$ vs. q . C, Multifractal spectrum showing $f(\alpha)$ vs. α (Source: Cheng and Agterberg 1995, Fig. 5)

resources (Nickless et al. 2014). Such actions are required in the following three fields: (1) inventory, (2) statistical modeling of known resources, and (3) estimation of future resources. Patiño Douce (2016a, b, c, d) has published four important papers that are helpful in achieving these aims,

showing, for example, that for copper there would be a deficit of about 2.39×10^9 t (tons) by the end of this century if recent discovery rates are maintained. For comparison, according to the USGS Mineral Commodity Summaries (2015), proven copper reserves currently are 0.68×10^9 t. According to Patiño Douce (2016d), estimated copper resources are 2.32×10^9 t, whereas new demand by 2100 will probably be 4.70×10^9 t. Consequently, estimated future copper deficit is approximately equal to forward projected known copper resources. Using a different statistical method, this forecast was confirmed by Agterberg (2017b) who estimated copper resources to be discovered by the end of this century at 2.77×10^9 t with a 95% confidence interval of $\pm 0.994 \times 10^9$ that contains Patiño Douce's estimate.

Patiño Douce (2016b)'s paper is accompanied by a supplementary database with sizes and grades for 20 metals in thousands of known worldwide mineral deposits. For example, data for 2541 copper deposits were compiled from as many as 49 different sources. Patiño Douce (2016b) fitted lognormal distributions to the metal deposit size-frequency distributions in this data base pointing out that the logarithmic (base e) standard deviation ranges from about 2 to 3 for different metals, although average metal deposit sizes are greatly different. Later, both Patiño Douce (2016c) and Agterberg (2017a) showed that the largest deposits for different metals can be described by means of Pareto distributions. In the Pareto-lognormal metal size-frequency distribution model of Agterberg (2017a, b, 2018a, b, c, 2020), the lognormal has a Pareto upper tail separated from the central lognormal by a bridge zone. This model recognizes both (1) lognormality of metal content of ore deposits from within smaller regions and those belonging to different mineral deposit types (see, e.g., Singer and Menzie 2010) and (2) Pareto size-frequency distribution of the largest deposits but also for the smallest metal deposits that exhibit Pareto size-frequency distributions as well. For most metals, the size-frequency distributions are exceedingly skew. For example, the 50% of worldwide copper deposits with less than median size contain only 0.4% of all copper contained in the 2541 deposits while 60% of all copper is contained in the 2% largest deposits. Consequently, the tails of these size-frequency distributions can only be studied for very large samples.

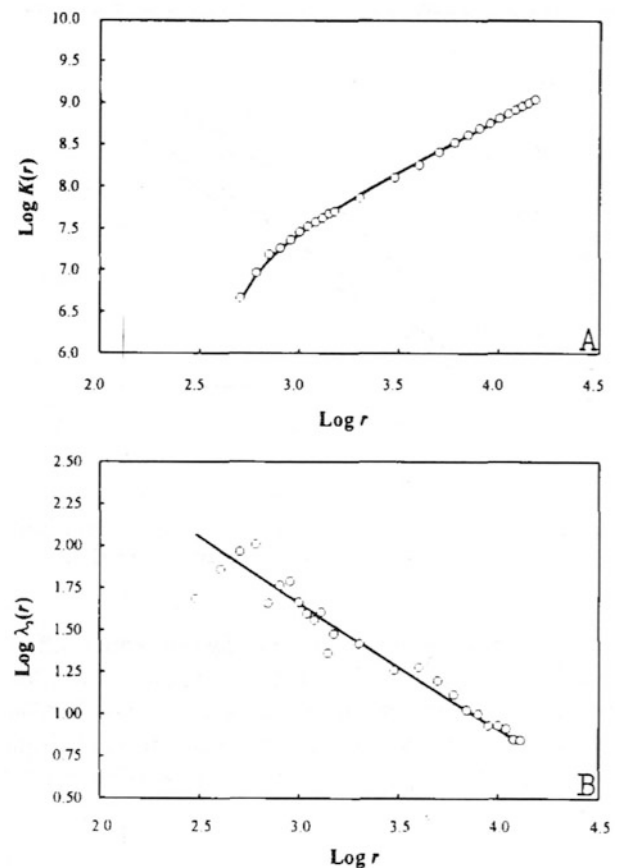
The Pareto-lognormal model for metal deposits provides an alternative to other size-frequency distribution models, which until about 30 years ago almost exclusively were based on the lognormal model. However, as pointed out by Mandelbrot (1983, p. 263), oil and other natural resources have Pareto distributions and "this finding disagrees with the dominant opinion, that the quantities in question are lognormally distributed. This difference is extremely significant, the reserves being much higher under the hyperbolic than under the lognormal law." Pareto size-frequency distribution

modeling of the largest deposits has been used by many authors including Drew et al. (1982) and Crovelli (1995) for oil and gas fields and Turcotte (1997) for metal deposits. Like the lognormal, the Pareto-lognormal distribution is not universally applicable to all elements, several of which show bimodal or multimodal size-frequency distributions (Fig. 21).

The cumulative frequency distribution for the Pareto-lognormal distribution $F(x) = F(\log x)$ can be written as

$$F(\log x) \approx \Phi\left(\frac{\log x - \mu}{\sigma}\right) + H(\log x - \mu) \cdot B_1(\log x) \cdot (\log x - \mu)^{-\alpha} + H(\mu - \log x) \cdot B_2(\log x) \cdot (\mu - \log x)^{-\kappa}$$

where $\Phi\left(\frac{\log x - \mu}{\sigma}\right)$ represents the basic lognormal (logs base 10). $H(\dots)$ is the Heaviside function that applies to two filtered Pareto distributions, for positive and negative values



Fractal Geometry in Geosciences, Fig. 21 Spatial statistics for Au mineral occurrences. A. Log-log plot of relationship between $K(r)$ and distance r ; solid lines are LS fits for $\tau(2) = 1.219 \pm 0.037$. B. Log-log plot for relationship between second-order intensity λ_2 and r with theoretical line for $\tau(2) = 1.219$ (Source: Cheng and Agterberg 1995, Fig. 7)

of $(\log x - \mu)$, respectively; it signifies that values at the other side of μ are set equal to zero when this equation is applied to either the upper tail or the lower tail of the Pareto-lognormal distribution. The bridge functions $B_1(\log x)$ and $B_2(\log x)$ span relatively short intervals between the basic lognormal and the Pareto distributions for the largest and smallest values, respectively. They satisfy $\lim_{x \rightarrow \infty} B_1(\log x) = \lim_{x \rightarrow 0} B_2(\log x) = 1$ and $\lim_{x \rightarrow 0} B_1(\log x) = \lim_{x \rightarrow \infty} B_2(\log x) = 0$.

The Pareto-lognormal probability density function $f(\log x)$ corresponding to $F(\log x)$ can be written as

$$f(\log x) \approx \varphi\left(\frac{\log x - \mu}{\sigma}\right) + H(\log x - \mu) \cdot B_1(\log x) \cdot (\log x - \mu)^{-\alpha-1} + H(\mu - \log x) \cdot B_2(\log x) \cdot (\mu - \log x)^{-\kappa-1}$$

It may be useful for prediction of resources to be discovered in the future. The exponents in $(\log x - \mu)^{-\alpha-1}$ and $(\mu - \log x)^{-\kappa-1}$ reflect the fact that the Pareto probability density function remains linear on a plot with logarithmic scales for frequency and deposit size but has a steeper dip than in the corresponding plot for the cumulative frequency distribution.

Figure 22 (from Agterberg 2018b) shows lognormal plots for the worldwide size-frequency distributions of six metals. A lognormal distribution would plot as a straight line on a plot of this type. In each case, two types of departure from lognormality are indicated. There are fewer than expected lognormal frequencies for the largest deposits and more than expected lognormal frequencies for the smallest deposits. A Pareto distribution of the largest deposits has the property that it plots as a straight line on a log-log plot of rank versus size. Figure 23 (also from Agterberg 2018b) shows that the six metals taken, for example, have Pareto upper tails. The method used for fitting these Pareto's is a modification of Quandt (1966)'s least squares method. In this modification the very few largest deposits are not included for estimation because they tend to distort the shape of Pareto distribution (see Agterberg 2018a).

Figure 24 (from Agterberg 2020) shows the Pareto-lognormal distribution for copper together with the observed frequencies as a $Q-Q$ plot (Fig. 24a) and as a log-log plot of cumulative frequency versus copper deposit size (Fig. 24b). It clarifies that relatively many large copper deposits with sizes less than those in the upper Pareto tail are missing. The other departure is that there are more deposits in the lower tail than expected for the central lognormal distribution. An explanation for these two departures from lognormality is that (a) only the very largest deposits in the upper Pareto tail

were discovered and mined because of their stronger geophysical signals during exploration and their greater economic significance and (b) before the nineteenth century only relatively small orebodies could be mined causing a relative increase of frequencies of small and very small deposits in the metal size-frequency distributions. The current model does not only apply to worldwide metal size-frequency distributions but also to Canadian ones (Agterberg 2020) although the land mass in Canada covers only 6.6% of the world's continents.

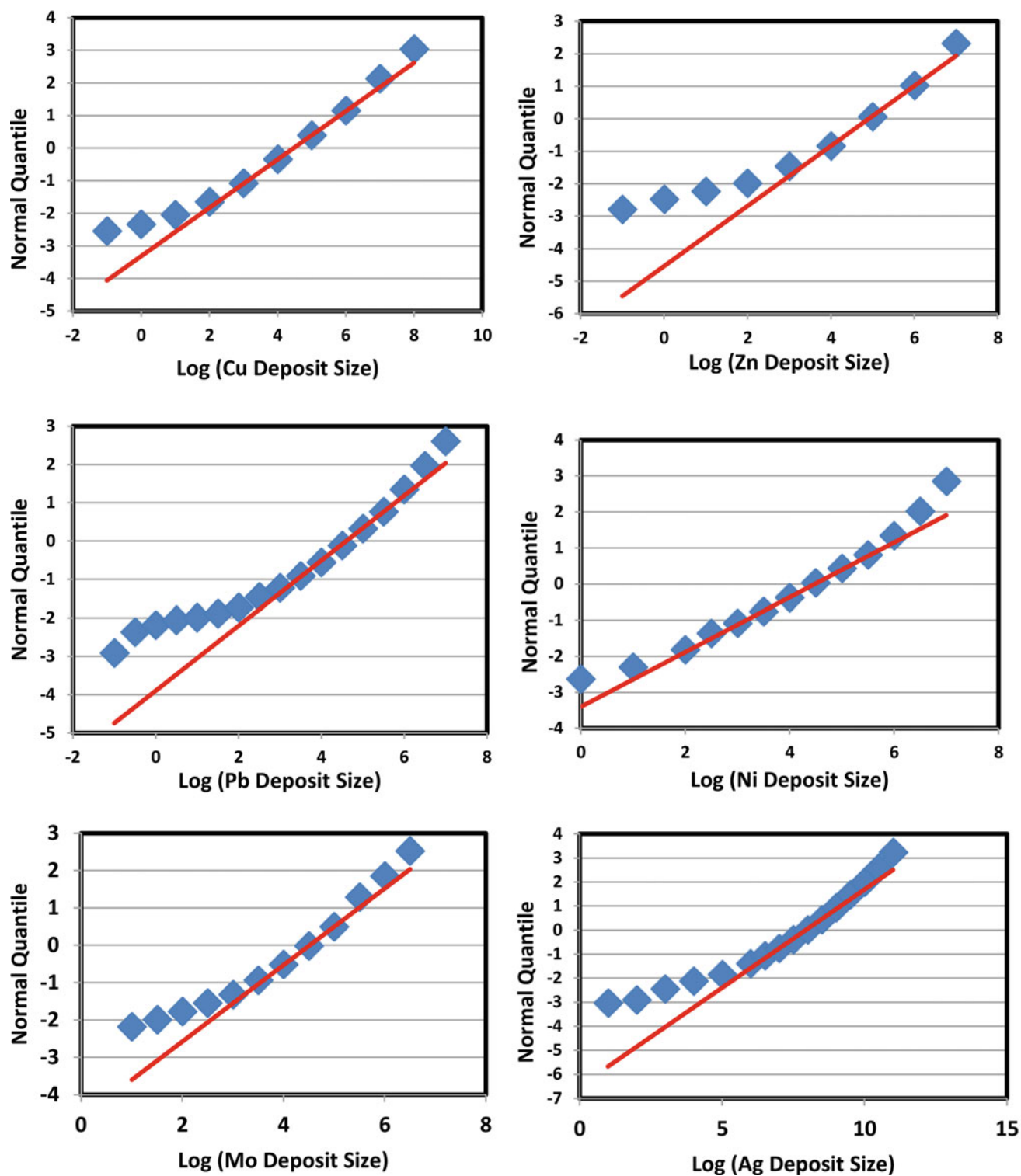
The current Pareto-lognormal was based on the so-called double Pareto-lognormal model originally developed by Reed (2003) and Reed and Jorgensen (2003) with cumulative distribution function:

$$F(\log x) \approx \Phi\left(\frac{\log x - \mu}{\sigma}\right) - \frac{1}{\alpha + \beta} \left[\beta \cdot x^{-\alpha} A(\alpha, \mu, \sigma) \Phi\left(\frac{\log x - \mu - \alpha\sigma^2}{\sigma}\right) + \alpha \cdot x^{\beta} A(-\beta, \mu, \sigma) \Phi\left(\frac{\log x - \mu + \beta\sigma^2}{\sigma}\right) \right]$$

where $A(\vartheta, \sigma, \mu) = \exp.(\vartheta \mu + \vartheta^2 \sigma^2 / 2)$. This model has had various applications in the economic and actuarial sciences (Kleiber and Kotz 2003). The original Reed-Jorgensen model could not be applied to the worldwide or Canadian metal size-frequency model because of the departures from lognormality (a) and (b) mentioned in the preceding paragraph. One of the properties of Reed's double Pareto-lognormal model is that $\beta > 1$ but in all applications to metal size-frequency distribution κ (used as upper tail Pareto parameter instead of β) is significantly less than 1. It is possible that Reed's model is valid for world or large region metal size-frequency distribution but at present it is not possible to estimate β .

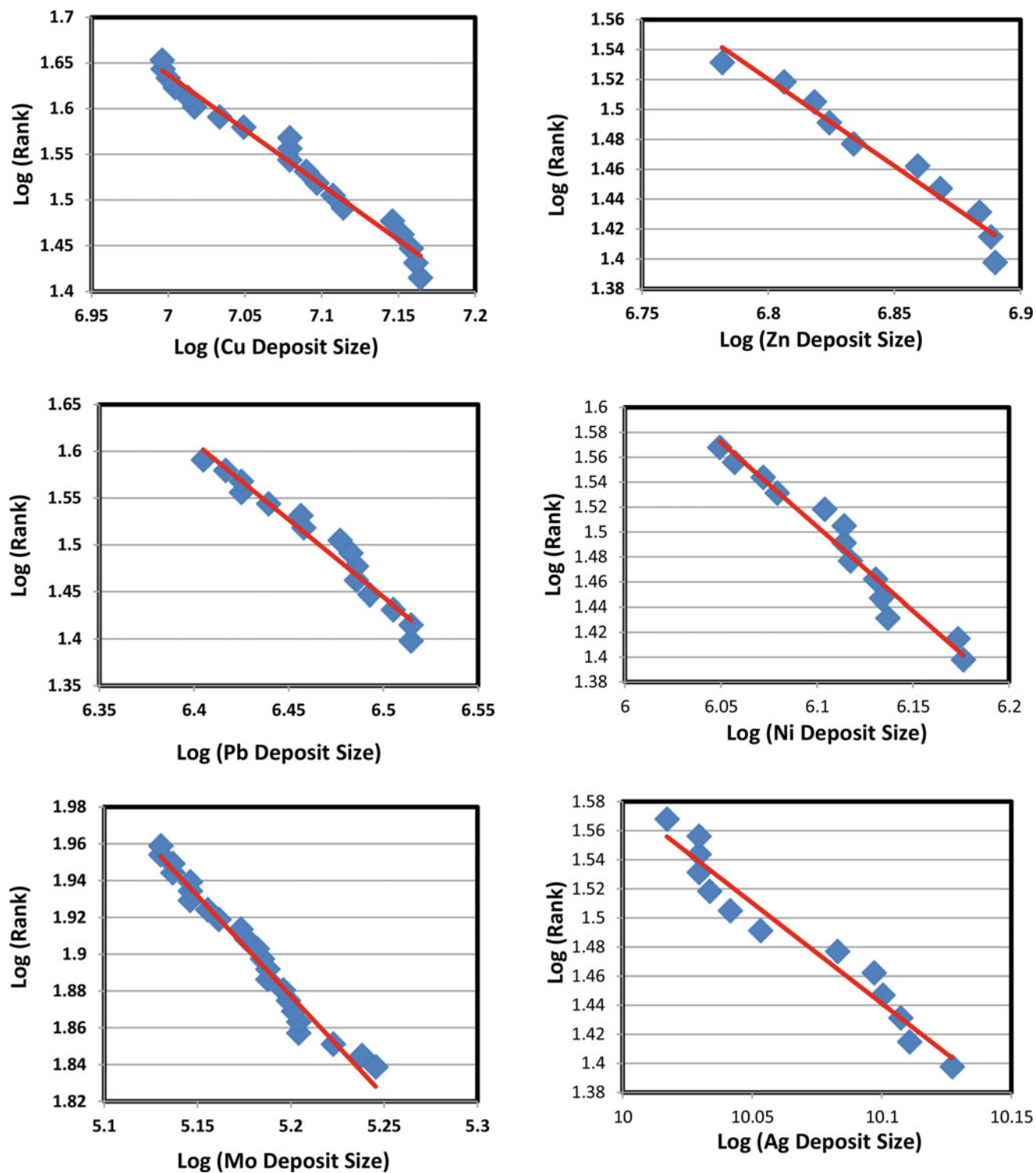
Summary

After their invention by Mandelbrot some 50 years ago, fractals with non-Euclidean dimensions have become widely applied in the geological and other sciences. In this entry, the focus has been on methods to estimate the fractal dimension or multifractal spectrum of geoscientific patterns. New approaches based on fractal geometry include the description of shapes of geological entities and the separation between anomalies and background in geochemistry. Singularity analysis uses local multifractal patterns of change in geological environments. The lognormal and Pareto frequency distributions of geological features including the sizes of mineral deposits can be based on fractal geometry in geoscience.



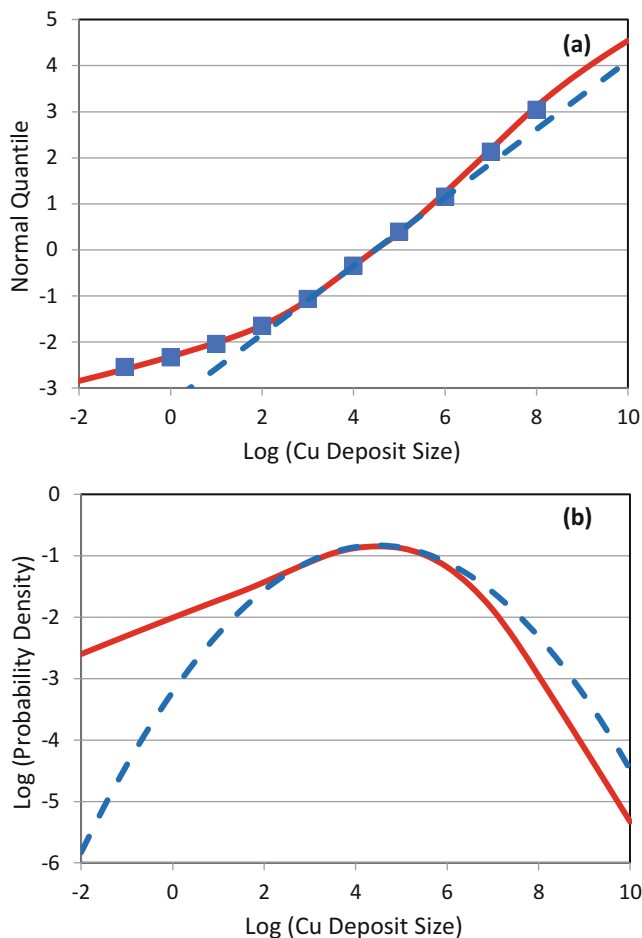
Fractal Geometry in Geosciences, Fig. 22 Lognormal $Q-Q$ plots for six metals. Sample sizes were 2541 (Cu), 1476 (Zn), 1102 (Pb), 464 (Ni) and 343 (Mo). In each case, frequencies for the largest and smallest deposits deviate from the straight-line patterns indicating lower

and higher number frequencies than expected on the basis of the lognormal size-frequency distribution models represented by the straight lines (Source: Agterberg 2018b, Fig. 26.2)



Fractal Geometry in Geosciences, Fig. 23 Sets of log (metal deposit size) values corresponding to estimates of upper tail Pareto coefficients (α) for the six metals of Fig. 22. Corresponding Pareto distributions are

shown as straight lines fitted by least squares method (Source: Agterberg 2018b, Fig. 26.4)



Fractal Geometry in Geosciences, Fig. 24 Two types of plots for worldwide copper deposits. Red lines are for Pareto-lognormal frequency distributions showing good fit to the data in the $Q-Q$ plot of Fig. 22 for copper. Blue lines correspond to straight line shown in Fig. 22a (Source: Agterberg 2020, Fig. 3)

Other developments along the line of multifractals and singularity are referred to other chapters of the current book.

Cross-References

- De Wijs Model
- Frequency Distribution
- Geologic Time Scale Estimation
- Lognormal Distribution
- Multifractals
- Point Pattern Statistics
- Singularity Analysis
- Variogram

References

- Agterberg FP (1974) *Geomathematics*. Elsevier, Amsterdam
- Agterberg FP (1980) Mineral resource estimation and statistical exploration. In: *Facts and principles of world oil occurrence*. Canadian Society of Petroleum Geology Memoir 6, pp 301–318
- Agterberg FP (2007) Mixtures of multiplicative cascade models in geochemistry. *Nonlinear Process Geophys* 14:201–209
- Agterberg FP (2012) Sampling and analysis of element concentration distribution in rock units and orebodies. *Nonlinear Process Geophys* 19:23–44
- Agterberg FP (2013) Fractals and spatial statistics of point patterns. *J Earth Sci* 24(1):1–11
- Agterberg FP (2014) *Geomathematics: theoretical foundations, applications and future developments*. Quantitative geology and geostatistics 18. Springer, Heidelberg
- Agterberg FP (2017a) Pareto-lognormal modeling of known and unknown metal resources. *Nat Resour Res* 26:3–20. (with erratum on p. 21)
- Agterberg FP (2017b) Pareto-lognormal modeling of known and unknown metal resources. II. Method refinement and further applications. *Nat Resour Res* 26(3):265–283
- Agterberg FP (2018a) Can multifractals be used for mineral resource appraisal? *J Geochem Explor* 189:54–63
- Agterberg FP (2018b) Statistical modeling of regional and worldwide size-frequency distributions of metal deposits. In: *Handbook of mathematical geosciences. Fifty years of IAMG*. Springer, Heidelberg, pp 505–527
- Agterberg FP (2018c) New method of fitting Pareto-lognormal size-frequency distributions of metal deposits. *Nat Resour Res* 27(1):265–283
- Agterberg FP (2020) Multifractal modeling of worldwide and Canadian metal size-frequency distributions. *Nat Resour Res* 29(4):539–550
- Agterberg FP, Cheng Q, Wright DF (1993) Fractal modeling of mineral deposits. In: *Application of computers and operations research in the mineral industry*. Canadian Institute of Mining, Metallurgy and Petroleum Engineering, Montreal, pp 43–53
- Arias M, Gumiel P, Martin-Izard A (2012) Multifractal analysis of geochemical anomalies: a tool for assessing prospectivity at the SE border of the Ossa Morena Zone, Variscan Massif (Spain). *J Geochem Explor* 122:101–112
- Burnett AI, Adams KC (1977) A geological, engineering and economic study of a portion of the Lloydminster Sparky Pool, Lloydminster, Alberta. *Bulletin of the Canadian Society of Petroleum Geology* 25(2):341–366
- Carranza EJM (2008) Geochemical anomaly and mineral prospectivity mapping in GIS. In: *Handbook of exploration and environmental geochemistry* 11. Elsevier, Amsterdam
- Chen Z, Cheng Q, Chen J, Xie S (2007) A novel iterative approach for mapping local singularities from geochemical data. *Nonlinear Process Geophys* 14:317–324
- Cheng Q (1994) Multifractal modeling and spatial analysis with GIS: gold mineral potential estimation in the Mitchell-Sulphurets area, northwestern British Columbia. Unpublished doctoral dissertation, University of Ottawa, Canada
- Cheng Q (1999) Spatial and scaling modelling for geochemical anomaly separation. *J Geochem Explor* 65:175–194
- Cheng Q (2006) Multifractal modelling and spectrum analysis: Methods and applications to gamma ray spectrometer data from southwestern Nova Scotia, Canada. *Science in China: Series D Earth Sciences* 49(3):283–294
- Cheng Q (2007) Mapping singularities with stream sediment geochemical data for prediction of undiscovered mineral deposits in Gejiu, Yunnan Province, China. *Ore Geol Rev* 32:314–324
- Cheng Q (2008) Non-linear theory and power-law models for information integration and mineral resources quantitative assessments. *Math Geosci* 40(5):503–532

- Cheng Q (2012) Singularity theory and methods for mapping geochemical anomalies caused by buried sources and for predicting undiscovered mineral deposits in covered areas. *J Geochem Explor* 122:55–70
- Cheng Q (2014) Generalized binomial multiplicative cascade processes and asymmetrical multifractal distributions. *Nonlinear Process Geophys* 21:472–482
- Cheng Q (2016) Fractal density and singularity analysis of heat flow over ocean ridges. *Nat Sci Rep* 6:1–10
- Cheng Q, Agterberg FP (1995) Multifractal modeling and spatial point processes. *Math Geol* 27:831–845
- Cheng Q, Agterberg FP (1996) Multifractal modeling and spatial statistics. *Math Geol* 28(1):1–16
- Cheng Q, Agterberg FP (2009) Singularity analysis of ore-mineral and toxic trace elements in stream sediments. *Comput Geosci* 35: 234–244
- Cheng Q, Agterberg FP, Ballantyne SB (1994) The separation of geochemical anomalies from background by fractal methods. *J Geochem Explor* 51:109–130
- Chhabra A, Jensen RV (1989) Direct determination of the $f(\alpha)$ singularity spectrum. *Phys Rev Lett* 62(12):1327–1330
- Cressie NAC (1991) *Statistics for spatial data*. Wiley, New York
- Crovelli RA (1995) The generalized 20/80 law using probabilistic fractals applied to petroleum field size. *Nonrenewable Resour* 4(3): 233–241
- De Wijs HJ (1951) Statistics of ore distribution, I. *Geol Mijnb* 30: 365–375
- Diggle PJ (1983) *Statistical analysis of spatial point patterns*. Academic Press, London
- Drew LJ, Schuenemeyer JH, Bawie WJ (1982) Estimation of the future rates of oil and gas discoveries in the western Gulf of Mexico. U.S. geological survey professional paper 1252, Washington, DC
- Evertsz CJG, Mandelbrot BB (1992) Multifractal measures. In: *Chaos and fractals*. Springer, New York
- Feder J (1988) *Fractals*. Plenum, New York
- Ford A, Blenkinsop TG (2009) An expanded de Wijs model for multifractal analysis of mineral production data. *Mineral Deposita* 44(2): 233–240
- Gonçalves MA, Mateus A, Oliveira V (2001) Geochemical anomaly separation by multifractal modeling. *J Geochem Explor* 72:91–114
- Herzfeld UC, Overbeck C (1999) Analysis and simulation of scale-dependent fractal surfaces with application to seafloor morphology. *Comput Geosci* 25(9):979–1007
- Herzfeld UC, Kim II, Orcutt JA (1995) Is the ocean floor a fractal? *Math Geol* 27(3):421–442
- Kaye BH (1989) *A random walk through fractal dimensions*. VCH Publishers, New York
- Kleiber C, Kotz S (2003) *Statistical distributions in economics and actuarial sciences*. Wiley, Hoboken
- Korvin G (1992) *Fractal models in the earth sciences*. Elsevier, Amsterdam
- Krige DG (1966) A study of gold and uranium distribution pattern in the Klerksdorp goldfield. *Geoprocessing* 4:43–53
- Lorentz EN (1963) Determinist nonperiodic flow. *J Atmos Sci* 20: 130–141
- Lovejoy S, Schertzer D (2007) Scaling and multifractal fields in the solid Earth and topography. *Nonlinear Process Geophys* 14:465–502
- Lovejoy S, Schertzer D (2018) *The weather and climate: emergent laws and multifractal cascades*. Cambridge University Press, New York
- Mandelbrot BB (1975) *Les Objects Fractals: Forme, Hazard et Dimension*. Flammarion, Paris
- Mandelbrot BB (1977) *Fractals: form, chance, and dimension*. Freeman, San Francisco
- Mandelbrot BB (1983) *The Fractal geometry of nature*. Freeman, San Francisco
- Mandelbrot BB (1989) Multifractal measures, especially for the geophysicist. *Pure Appl Geophys* 131:5–42
- Matheron G (1962) *Traité de Géologie Appliquée*. Mémoire BRGM 14. Éditions Technip, Paris
- Meneveau C, Sreenivasan KR (1987) Simple multifractal cascade model for fully developed turbulence. *Phys Rev Lett* 59(7):1424–1427
- Nickless E, Bloodworth A, Meinert L, Giurco D, Mohr S, Littleboy A (2014) *Resourcing future generations white paper: mineral resources and future supply*. International Union of Geological Sciences. <http://www.americangeosciences.org/community/resourcing-future-generations-white-paper>
- Pahani A, Cheng Q (2004) Multifractality as a measure of spatial distribution of geochemical patterns. *Math Geol* 36:827–846
- Patiño Douce AE (2016a) Metallic mineral resources in the twenty first century. I. Historical extraction trends and expected demand. *Nat Resour Res* 25:71–90
- Patiño Douce AE (2016b) Metallic mineral resources in the twenty first century. II. Constraints on future supply. *Nat Resour Res* 25:97–124
- Patiño Douce AE (2016c) Statistical distribution laws for metallic mineral deposit sizes. *Nat Resour Res* 25:365–387
- Patiño Douce AE (2016d) Loss distribution model for metal discovery probabilities. *Nat Resour Res*. <https://doi.org/10.1007/s11053-016-9325-2>
- Perrin J (1913) *Les Atomes*. NRF-Gallimard, Paris
- Poincaré H (1899) *Les methodes Nouvelles de la mécanique celeste*. Gauthier-Villars, Paris
- Quandt RE (1966) Old and new methods of estimation and the Pareto distribution. *Metrika* 10:55–82
- Reed WJ (2003) The Pareto law of increases: an explanation and an extension. *Physica A* 319:579–597
- Reed WJ, Jorgensen M (2003) The double Pareto-lognormal distribution. A new parametric model for size distributions. *Comput Stat: Theory Methods* 33(8):1733–1753
- Ripley BD (1976) The second-order analysis of stationary point processes. *J Appl Probab* 13:255–266
- Sagar BS, Rangarajan G, Veneziano D (2004) Fractals in geophysics. *Chaos, Solitons Fractals* 10(2):237–239
- Sinclair AJ (1991) A fundamental approach to threshold estimation in exploration geochemistry: probability plots revisited. *J Geochem Explor* 41(1):1–22
- Singer DA, Menzie DW (2010) *Quantitative mineral resource assessments*. Oxford University Press, New York
- Stanley H, Meakin P (1988) Multifractal phenomena in physics and chemistry. *Nature* 335:405–409
- Steinhaus H (1954) Length, shape and area. *Col. Math.*, III: 1–13
- Stoyan D, Stoyan H (1994) *Fractals, random shapes and point fields*. Wiley, Chichester
- Turcotte DL (1997) *Fractals and Chaos in geology and geophysics*. Cambridge University Press, 2nd
- USGS Mineral Commodity Summaries, 2015. <http://minerals.usgs.gov/minerals/pubs/mcs2015.pdf>
- Vening Meinesz FA (1951) A remarkable feature of the Earth's topography. *Proc KNAW Ser B Phys Sci* 54:212–228
- Vening Meinesz FA (1964) *The earth's crust and mantle*. Elsevier, Amsterdam
- Whittle P (1962) Topographic correlation power-law covariance functions and diffusion. *Biometrika* 49:305–314
- Xiao F, Chen J, Zhang Z, Wang C, Wu G, Agterberg FP (2012) Singularity mapping and spatially weighted principal component analysis to identify geochemical anomalies associated with Ag and Pb-Zn polymetallic mineralization in Northwest Zhejiang, China. *J Geochem Explor* 122:90–100
- Xie X, Mu X, Ren T (1997) Geochemical mapping in China. *J Geochem Explor* 60:99–113
- Yu C (2002) Complexity of earth systems – fundamental issues of earth sciences. *J China Univ Geosci* 27:509–519. (in Chinese; English abstract)
- Zuo R, Wang J (2016) Fractal/multifractal modeling of geochemical data: a review. *J Geochem Explor* 164:33–41

Fractal Landscape

Ramendra Sahoo

Department of Earth and Climate Science, IISER Pune, Pune, India

Definition

Fractals are objects with scale-invariant geometry, introduced to describe the non-Euclidian shape of natural phenomena. Fractal dimension, which is proportionate to the degree of space filled by an object withing its embedding space, is generally used to describe the fractal geometry of natural landscapes. Multifractal geometry results, if we have heterogeneity of space filling across the embedding space. Fractal properties of landscape can often be quantitatively related to the spatio-temporal scale of the associated governing natural processes, which provides novel insights into the nonlinear dynamics ubiquitous in nature.

Introduction

Many natural phenomena look similarly complex under different resolutions, a property known as scale-invariance (not to be confused with scale-independence). Without an object with a representative dimension, often referred to as a scale in the geological literature, it is impossible to determine whether the object in a picture covers a few cm or km (e.g., a self-similar fold or a river network). These kinds of scale-invariant patterns are not characterizable with the help of a single scale. It was in this context that Mandelbrot (1967) introduced the concept of fractals in order to describe the complex geometry of a rocky coastline. Subsequently, many natural patterns and processes have been proven to be efficiently modeled by fractals. They include coastline and mountain topography (e.g., Mandelbrot 1982), river networks (e.g., Tarboton et al. 1988), seismicity (e.g., Turcotte 1989), and ocean-floor bathymetry (e.g., Mareschal 1989), just to name a few. The fractal geometry of these natural phenomena can be described with fractal dimension, a number that agrees with our intuitive notion of dimension but need not be an integer.

Fractal dimension of any natural object also highlights the inherent dominant frequency present in its physical attributes. That dominant frequency, in turn, can be attributed to the scale (spatial and/or temporal) of the governing natural processes. For example, fractal dimension of sea level change time series can be related to temporal scale of climate pattern (Indira et al. 1996). Fractal characteristics of climate time series have also been used to find out the periodicity of climate pattern (Gipp 2001) and possible role of chaos in

climate dynamics. Geomorphologists have also related the fractal dimension of the topography to governing controls, such as tectonics, climate, and lithology (Liucci and Melleli 2017). Hence, the application of fractal concepts in natural science is providing novel insights into the underlying process dynamics. This entry focuses on the advancements and applications, along with a brief review, of the concepts of fractal and scaling in the field of geomorphology.

A Review of Fractals: A Unifying Basis to Describe Nature's Complexity

Scale-invariance is a fundamental property of fractal geometry. These geometries cannot be characterized by a single scale. For example, if we consider the famous problem of measuring the length of a coastline, its length keeps on changing with scale. Fractal dimension (d_f) is one of the ways to describe the geometry of such objects. If we define the embedding dimension (d_E) as the minimum Euclidean dimension required to contain the object, the volume $V(\delta)$ of an object can be measured by covering it with $N(\delta)$ balls of linear size δ , and volume δ^{d_E} . So,

$$V(\delta) = N(\delta)\delta^{d_E} \quad (1)$$

One might expect $N(\delta)$ scales as δ^{-d_E} , i.e., $N(\delta) \sim \delta^{-d_E}$ as the volume of the object doesn't change with change in scale. For a fractal object, we have

$$N(\delta) \sim \delta^{-d_f} \quad (2)$$

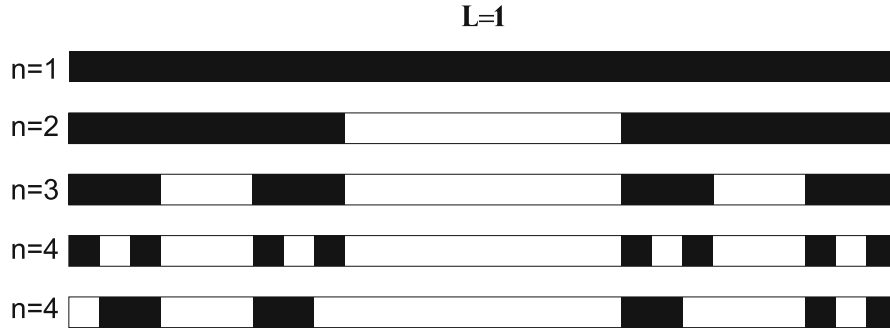
$d_f < d_E$ becomes an important norm to define an object as fractal (Mandelbrot 1982). The above expression can be rearranged as follows to give us an expression for fractal dimension.

$$d_f = \lim_{\delta \rightarrow 0} \frac{\ln N(\delta)}{\ln(1/\delta)} \quad (3)$$

We can use Eq. 3 to estimate the fractal dimension of any natural object (random or statistical fractals). For the same example of coastline length, we can measure the total length, $\delta N(\delta) = \delta^{1-d_f}$, and we would find that it increases as δ decreases. If we plot $N(\delta)$ as a function of δ on a log-log plot, we will get a straight line, whose slope will give an estimate of the fractal dimension of the coastline.

Self-similar Fractals

The scale-invariance nature of fractal objects can be dichotomized into self-similar and self-affine fractals. A self-similar fractal object constitutes of parts that can be rescaled



Fractal Landscape, Fig. 1 Construction of the Cantor set, by repeated removal of the middle third of each interval. The last row depicts a random Cantor set, where the segments are removed from each interval from the previous step ($n = 3$) at random, giving rise to a clumped

isotopically to match the geometry of the original object. Cantor set is a theoretical example of a self-similar fractal, and a drainage network is a natural example for the same. Let us consider the example of a regular deterministic Cantor set (Fig. 1). We start with a line segment of unit length, divide it into three parts, and finally remove the central part. After a number of iterations, the geometry converges to a set of points, known as Cantor set. At stage n of the construction, the set constitutes $N(\delta_n) = 2^n$ intervals of length $\delta_n = 3^{-n}$. The number of segments required to cover the total length is then given by

$$N(\delta_n) = \delta_n^{-d_f} \quad (4)$$

Equation 4 is similar to Eq. 2, and thus, we can use Eq. 3 to calculate the fractal dimension of the Cantor set.

Unlike the mathematical fractals, natural objects do not display exact self-similarity. Instead, they show some degree of statistical self-similarity over a limited spatial or temporal scale. Statistical self-similarity refers to scale-related repetitions of overall complexity but not of the exact pattern. We also observe step-like transition in the power-law relation (Eq. 4), where the environmental properties or constraints in the natural system may significantly change. Hence, the prevalence of power law with respect to the scale of observation within a certain range is commonly used to discern a fractal and, in that range, we can use Eq. 3 to calculate the fractal dimension.

Self-affine Fractals

Fractal objects that need to be rescaled anisotropically are classified as self-affine objects. Fraction Brownian motion (fBm) is a theoretical example of self-affine fractal, while earth's topography is a natural example having self-affine fractal geometry. These fractals can be described using a single-valued self-affine function, $h(x)$, which scales as follows:

distribution. Notably, the last two rows ($n = 4$) have same fractal dimension ($\ln 2 / \ln 3$); however, they are characterized by different set of lacunarity

$$h(x) \sim b^{-H} h(bx) \quad (5)$$

where H is called the Hurst exponent and provides a quantitative measurement of the roughness of the function $h(x)$. Eq. 5 explains that if we scale the function with a factor b horizontally ($x \rightarrow bx$), it must be scaled with a factor b^H vertically ($h \rightarrow b^H h$) in order that the resulting function overlaps the original function. For the special case $H = 1$, the transformation is isotropic and the system is self-similar. Using the roughness exponent, we also may further define a fractal dimension for the self-affine function using the following expression (Seuront 2009):

$$d_f = d_E - H \quad (6)$$

One corollary of the scaling relation of the self-affine function (Eq. 5) is

$$\Delta(\delta) \sim \delta^H \quad (7)$$

where $\Delta(\delta) = |h(x_1) - h(x_2)|$, for $\delta = |x_1 - x_2|$. For a self-affine function similar to Brownian motion, the standard deviation is defined as

$$\left\langle [h(x_1) - h(x_2)]^2 \right\rangle^{1/2} \sim |x_1 - x_2|^{1/2} \quad (8)$$

Equation 8 is similar to Eq. 7, giving $H = 1/2$. Consequently, the fractal dimension of a two-dimensional $h(x)$ is 1.5. In general, it is possible to generate a function, such as fractional Brownian motion, where the standard deviation scales in a more general manner as follows:

$$\left\langle [h(x_1) - h(x_2)]^2 \right\rangle^{1/2} \sim |x_1 - x_2|^H \quad (9)$$

While $H = 0.5$ recovers a Brownian motion, where direction of displacements is equally likely at each step, for

$H > 0.5$ and $H < 0.5$, the increments are positively (persistent) and negatively (antipersistent) correlated, respectively. Persistence and anti-persistence provide an intuitive sense of the governing physical processes of the natural system, such as climate dynamics (Xu et al. 2017).

Multifractals

Previous sections establish that natural patterns are adequately described by fractal geometry and fractal dimension irrespective of the measurement scale. Nevertheless, many natural phenomena are characterized by highly intermittent patterns (a few hotspots dispersed over a wide range of low-density areas). Mandelbrot (1982) recognized that objects with identical fractal dimension (e.g., deterministic and random Cantor set in Fig. 1) can have greatly different appearances owing to the spatial distribution of the gaps. Hence, he introduced lacunarity index, which uses higher order moments of height-height distribution to uniquely describe objects with intermittent patterns. The utilization of higher order moments was subsequently generalized through the concept of multifractals.

Multifractals are defined as union of subsets with different scaling factors. Morphologically, a multifractal object is described as having a spatially heterogeneous roughness exponent. The concept gets more intuitive if one thinks of fractal dimension in terms of probability of an object filling up its embedding space (analogous to counting the fraction of embedding space occupied by an object). A probability distribution can be completely defined by its moments. In case of multifractals, characterized by non-Gaussian underlying distribution, we need multiple moments to exhaustively describe their pattern. This can be achieved using the q^{th} -order structure functions that are an empirical generalization to higher orders of moments. We can define the q^{th} -order moment of the probability distribution, also known as the partition function (Halsey et al. 1986), as follows:

$$M_q(\delta) = \sum_{i=1}^{N(\delta)} f_i(\delta)^q \quad (10)$$

where $f_i(\delta)$ denotes the probability of finding the object in the segment i with a linear dimension δ , and $N(\delta)$ is the number of discrete segments. The generalized information dimension and the generalized correlation dimension are then expressed as (Seuront 2009)

$$I_q(\delta) = \frac{1}{q-1} \ln M_q \quad (11)$$

$$D(q) = \lim_{\delta \rightarrow 0} \frac{I_q(\delta)}{\ln(1/\delta)} \quad (12)$$

This generalized correlation dimension, $D(q)$, is defined for all real values of $q (\neq 1)$ and is estimated as the slope of the log-log plot of $I_q(\delta)$ vs δ . For $q = 1$, the generalized dimension is given by

$$D(q) = \lim_{\delta \rightarrow 0} \frac{1}{q-1} \frac{\ln \sum_{i=1}^{N(\delta)} f_i(\delta) \ln f_i(\delta)}{\ln(1/\delta)} \quad (13)$$

The set of points $(q, D(q))$ defines a curve that depicts the generalized spectrum of the measure $f(\delta)$. This spectrum is a decreasing function of q with a sigmoidal shape for a multifractal object. The partition function, $M_q(\delta)$, can also be expressed as a power-function of length scale as follows (Seuront 2009):

$$M_q(\delta) \propto \delta^{-\tau(q)} \quad (14)$$

where $\tau(q)$ is defined as a mass exponent function, which is linear for a monofractal object, that is, $\tau(q) = qH - 1$. For a multifractal object, if we combine Eqs. 11, 12, and 14, we get the following functional form for the mass exponent.

$$\tau(q) = (q-1)D(q) \quad (15)$$

The characterization of the measure $f(\delta)$ can also be done in terms of its multifractal spectrum, $f(\alpha)$, which is related to the mass exponent through a Legendre transformation (Evertsz and Mandelbrot 1992) as follows:

$$\alpha(q) = \frac{d\tau(q)}{dq} \quad (16)$$

$$f(\alpha(q)) = q\alpha(q) - \tau(q) \quad (17)$$

where $\alpha(q)$ is the local Holder's exponent. $\Delta\alpha$ is widely used in the literature as a measure of multifractality (e.g., Dutta 2017).

Scaling of Earth's Landscape

The shape of Earth's landscape is determined by numerous competing processes that act to either roughen or smoothen the surface. Fractal analysis provides a scale-independent robust measure of roughness observed in natural landscape. This section succinctly describes a few applications of fractal and multifractal concepts in the field of geomorphology.

Vening Meinesz (1951) was the first quantitative analysis of scaling in landscape topography and showed that the power spectrum $E(k)$ of topography (where k is a wavenumber) roughly scales as $k^{-\beta}$ with a spectral exponent $\beta \approx 2$. Later

results from subsequent bathymetric studies reported a range of values for the exponent ($\beta \sim 1.6\text{--}2.3$) across a wide range of scales. Accordingly, the isolines (intersection of topography with a horizontal plane, such as coastlines and contour lines) can also be characterized by power-law spectrum. Richardson (1961) found that the length of various coastlines varies in a power-law way with the length of the rulers used to measure them. Mandelbrot (1967) interpreted these scaling exponents in terms of fractal dimensions. Subsequently, fractal geometry (Mandelbrot 1982) has been proven to be a successful mathematical model for describing real landscapes. Fractal parameters, particularly the fractal dimension (FD), have been shown to be a good morphometric descriptor across the domains of geomorphology.

Surface roughness not only aids in characterization of landscape texture, but also is an important parameter in the evolution of landscape topography. This fact has led to an in-depth process-response investigation and association between the fractal nature of the topography and the governing controls in fluvial landscapes (e.g., Sahoo et al., 2020), fractal analysis of ice sheet surface and isolines (e.g., Rezvanbehbahani et al., 2019), and fractal analyses of coastal (Donadio et al., 2020) and aeolian landscape (Zgheib et al., 2018).

The fractal properties of natural surfaces (and thus their FD) change with the scaling range. A single FD may not be a good descriptor of terrain complexity for a large study area. Lovejoy and Schertzer (1990) argued that we need multifractal measures (rather than a unique scaling exponent, such as the FD) to describe any natural surface. A similar line of discussion arises in the analyses of fractures and other artificial surfaces (e.g., Morel et al., 2002). Multifractalism also aided in inferring process complexity based on the values of Holder's exponent (α) and generalized fractal dimension spectrum (D_q) of the topography (e.g., Dutta, 2017).

Drainage networks exhibit different drainage patterns depending on the hydrogeological properties of the underlying materials. First scaling relations for the drainage network patterns were proposed by Horton (1945) and Strahler (1957). Subsequent research on these scaling properties led geologists and hydrologists to the recognition of the fractal and multifractal nature of drainage networks (e.g., Tarboton et al. 1988; Camara et al., 2016). Indeed, the ubiquity of scaling relations in surface hydrology would be difficult to comprehend without wide range scale-invariance of the topography.

These scaling laws have been found to be valid for the topography and drainage networks of other planets as well (e.g., Gou et al., 2019). This undeniably shows the universality of the fractal nature of landscape geometry and establishes the fractal dimension (or set of scaling exponents in case of multifractal) to be a fundamental scale-independent property of the landscape.

Proposed Dynamical Models for Fractal Landscapes

Beyond describing the non-Euclidian shape of natural objects, fractal geometry also demands that we look into the possible process dynamics responsible for such forms of natural landscapes. This section describes a few interesting mathematical concepts proposed for the fractal nature of fluvial landscapes.

A geometry of landscape is fractal if there is no characteristic length scale. Ergo, one would expect the landscaping processes also to be devoid of any characteristic length. Fluvial landscapes, in particular bedrock rivers, are primarily dictated by hillslope smoothing (a diffusive process) and channel erosion (an advective process) (Tucker and Hancock 2010). The advection and diffusion processes operate at different timescales, which are defined as follows (Perron et al. 2008):

$$T_{adv} = \frac{L^{1-2m}}{K} \quad (18)$$

$$T_{dif} = \frac{L^2}{\kappa} \quad (19)$$

where L is the length scale of a landscape, and K and κ are channel erodibility and hillslope diffusivity, respectively. It is quite apparent from Eqs. 18 and 19 that diffusion process is scale-dependent, while the advection process is scale-independent for $m = 0.5$, which is an acceptable value for the exponent (Whipple and Tucker 1999). This is in good agreement with the observation in real topography (Pelletier 2004), where at large length scales, the model is self-affine and closely follows the power-law dependence and at smaller length scales, where hillslope processes are prevalent, the power spectrum drops off below the power-law trend. Channel erosion, therefore, is regarded as the primary surface process to produce fractal topography.

Another model proposed in the literature are Kardar-Parisi-Zhang (KPZ) equation, combining linear diffusion (for hillslope process) and nonlinear advection (for channel erosion) (Pelletier 2007; Duclut and Delamotte 2017).

$$\frac{\partial h}{\partial t} = \kappa \frac{\partial^2 h}{\partial x^2} + \frac{\lambda}{2} \left(\frac{\partial h}{\partial x} \right)^2 + \eta(x, y, t) \quad (20)$$

The first term on the right-hand side is the linear diffusion term. The second term is the nonlinear advection term and the third term is a random noise. KPZ equation in 3D generates a self-affine fractal surface under the combined action of smoothing by the diffusion and roughening by the nonlinear advection and the random noise. A deterministic KPZ model could also capture the drainage migration process (Pelletier

2007), which is credited as a major contributor to the dynamic complexity of the landscape (Chin 2002).

Other models suggested for the fractal nature of landscapes include diffusion limited aggregation (DLA) model, which has been proposed for generation of self-similar network pattern in dendritic rivers (Ossadnik 1992). DLA clusters have also been found to strongly resemble other natural network patterns, such as fault and joint network (Sornette et al., 1990) and mineral growth pattern (Fowler 1995). Least energy dissipation principle is another concept which has also been put forth as a possible mechanism for fractal structures of river network (Rodriguez-Iturbe et al. 1992).

A broad class of nonlinear physical problems, primarily consisting of complex geological processes or self-organized critical behavior, also invariably yields fractal behavior for natural landscape (Phillips, 2016). Nonlinearity in a dynamical system primarily arises from self-organized criticality or complexity. Multiple geomorphic controls and their associated dynamics give rise to complex geomorphic systems (e.g., Keiler, 2011). When there exists a strong feedback between the geomorphic components, such as climate-tectonics feedback (Molnar and England 1990), the dynamical system can move toward a chaotic state, giving rise to strange attractor with fractal scaling (e.g., Baas 2002). Conversely, a dominant effect of any autogenic control, such as lithology or catchment shape, in landscape evolution can result in a self-organized critical system (e.g., Phillips, 2016). In such scenarios, the system always evolves toward a finite number of critically stable state(s) and is primarily controlled by some form of threshold. Recent experimental or field-based studies also drew attention to the self-organized critical nature of fluvial geomorphic systems (e.g., Baynes et al., 2018). Resulting nonlinearity in fluvial landscapes can lead to complex behavior only at one scale or across multiple scales, and generation of fractal patterns with single or multiple scaling exponents.

Conclusion

Fractal concepts provide important tools to study the rough, crenulated, and crumpled 3-D surface of natural landscapes. Owing to its scale-independent property, fractal dimension captures the intricate complexity of landscape pattern that were not otherwise discernible using traditional morphometric parameters. The topographic characterization through the fractal patterns also presents an opportunity to analyze the innate nonlinearity in the geomorphic system. Fractal properties are useful to detect the heterogeneity in the degree of complexity in landscape pattern as well. Fractals have brought a significant change in the conventional ways of thinking about spatial forms, and complex natural landscapes are inherently amenable to these types of investigations.

Bibliography

- Baas AC (2002) Chaos, fractals and self-organization in coastal geomorphology: simulating dune landscapes in vegetated environments. *Geomorphology* 48(1–3):309–328
- Baynes ER, Lague D, Attal M, Gangloff A, Kirstein LA, Dugmore AJ (2018) River self-organisation inhibits discharge control on waterfall migration. *Sci Rep* 8(1):1–8
- Camara J, Gómez-Miguel V, Martín MÁ (2016) Identification of bed-rock lithology using fractal dimensions of drainage networks extracted from medium resolution LiDAR digital terrain models. *Pure Appl Geophys* 173(3):945–961
- Chin CS (2002) Passive random walkers and riverlike networks on growing surfaces. *Phys Rev E* 66(2):021104
- Donadio C, Paliaga G, Radke JD (2020) Tsunamis and rapid coastal remodeling: linking energy and fractal dimension. *Prog Phys Geogr Earth Environ* 44(4):550–571
- Duclut C, Delamotte B (2017) Nonuniversality in the erosion of tilted landscapes. *Physical Review E* 96(1):012149(1–10)
- Dutta S (2017) Decoding the morphological differences between himalayan glacial and fluvial landscapes using multifractal analysis. *Sci Rep* 7(1):1–8
- Evertsz CJG, Mandelbrot BB (1992) Multifractal measures. In: Peitgen HO, Hartmut J, Saupe D (eds) *Chaos and fractals: new Frontiers of science*. Springer, New York, pp 849–881
- Fowler AD (1995) Mineral crystallinity in igneous rocks. In: *Fractals in the earth sciences*. Springer, Boston, pp 237–249
- Gipp MR (2001) Interpretation of climate dynamics from phase space portraits: is the climate system strange or just different? *Paleoclimatology* 16(4):335–351
- Gou S, Yue Z, Di K, Xu Y (2019) Comparative study between rivers in Tarim Basin in Northwest China and Evros Vallis on Mars. *Icarus* 328:127–140
- Halsey TC, Jensen MH, Kadanoff LP, Procaccia I, Shraiman BI (1986) Fractal measures and their singularities: the characterization of strange sets. *Phys Rev A* 33(2):1141
- Horton RE (1945) Erosional development of streams and their drainage basins; hydrophysical approach to quantitative morphology. *Geol Soc Am Bull* 56(3):275–370
- Indira NK, Singh RN, Yajnik KS (1996) Fractal analysis of sea level variations in coastal regions of India. *Curr Sci* 70(8):719–723
- Keiler M (2011) Geomorphology and complexity-inseparably connected? *Zeitschrift für Geomorphologie-Supplementband* 55(3): 233–257
- Liucci L, Melelli L (2017) The fractal properties of topography as controlled by the interactions of tectonic, lithological, and geomorphological processes. *Earth Surf Process Landf* 42(15):2585–2598
- Lovejoy S, Schertzer D (1990) Multifractals, universality classes and satellite and radar measurements of cloud and rain fields. *J Geophys Res Atmos* 95(D3):2021–2034
- Mandelbrot B (1967) How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science* 156(3775):636–638
- Mandelbrot BB (1982) *The fractal geometry of nature*, vol 1. WH Freeman, New York
- Mareschal JC (1989) Fractal reconstruction of sea-floor topography. In: *Fractals in geophysics*. Birkhäuser, Basel, pp 197–210
- Molnar P, England P (1990) Late Cenozoic uplift of mountain ranges and global climate change: chicken or egg? *Nature* 346(6279):29–34
- Morel S, Bouchaud E, Schmittbuhl J, Valentin G (2002) R-curve behavior and roughness development of fracture surfaces. *Int J Fract* 114(4):307–325
- Ossadnik P (1992) Branch order and ramification analysis of large diffusion-limited-aggregation clusters. *Phys Rev A* 45(2):1058–1066
- Pelletier JD (2004) Persistent drainage migration in a numerical landscape evolution model. *Geophys Res Lett* 31(20)

- Pelletier JD (2007) Fractal behavior in space and time in a simplified model of fluvial landform evolution. *Geomorphology* 91(3–4):291–301
- Perron JT, Dietrich WE, Kirchner JW (2008) Controls on the spacing of first-order valleys. *J Geophys Res Earth* 113(F4)
- Phillips JD (2016) Vanishing point: scale independence in geomorphological hierarchies. *Geomorphology* 266:66–74
- Rezvanehbehbahani S, van der Veen CJ, Stearns LA (2019) Fractal properties of Greenland isolines. *Math Geosci* 51(8):1075–1090
- Richardson LF (1961) The problem of contiguity: an appendix of statistics of deadly quarrels. *General Syst Yearbook* 6:139–187
- Rodriguez-Iturbe I, Rinaldo A, Rigon R, Bras RL, Ijjasz-Vasquez E, Marani A (1992) Fractal structures as least energy patterns: the case of river networks. *Geophys Res Lett* 19(9):889–892
- Sahoo R, Singh RN, Jain V (2020) Process inference from topographic fractal characteristics in the tectonically active Northwest Himalaya, India. *Earth Surf Process Landforms* 45(14):3572–3591
- Seuront L (2009) *Fractals and multifractals in ecology and aquatic science*. CRC Press
- Sornette A, Davy P, Sornette D (1990) Growth of fractal fault patterns. *Phys Rev Lett* 65(18):2266–2269
- Strahler AN (1957) Quantitative analysis of watershed geomorphology. *EOS Trans Am Geophys Union* 38(6):913–920
- Tarboton DG, Bras RL, Rodriguez-Iturbe I (1988) The fractal nature of river networks. *Water Resour Res* 24(8):1317–1322
- Tucker GE, Hancock GR (2010) Modelling landscape evolution. *Earth Surf Process Landf* 35(1):28–50
- Turcotte DL (1989) A fractal approach to probabilistic seismic hazard assessment. *Tectonophysics* 167(2–4):171–177
- Vening Meinesz FA (1951) A remarkable feature of the Earth's topography, origin of continents and oceans. *Proc Koninkl Ned Akad Wetensch ser B* 55:212–228
- Whipple KX, Tucker GE (1999) Dynamics of the stream-power river incision model: implications for height limits of mountain ranges, landscape response timescales, and research needs. *J Geophys Res Solid Earth* 104(B8):17661–17674
- Xu Z, Tang Y, Connor T, Li D, Li Y, Liu J (2017) Climate variability and trends at a national scale. *Sci Rep* 7(1):1–10
- Zgheib N, Fedele JJ, Hoyal DCJD, Perillo MM, Balachandar S (2018) Direct numerical simulation of transverse ripples: 1. Pattern initiation and bedform interactions. *J Geophys Res Earth* 123(3):448–477

Frequency Distribution

Ricardo A. Olea
Geology, Energy and Minerals Science Center,
U.S. Geological Survey, Reston, VA, USA

Definition

Given a numerical dataset, a frequency distribution is a summary displaying fluctuations of an attribute within the range of values. In contrast to an analytical probability distribution, a frequency distribution always deals with empirically observed values (Everitt and Skondall 2010). In general, the larger the number of values, the more useful the frequency distribution is relative to listing all values. Today multiple software packages allow easy display of a frequency

distribution. The dataset may be univariate or multivariate. For simplicity, this entry deals with univariate datasets.

Different Forms

In the broadest sense, a frequency distribution may be tabular or graphical. In either case, the most common practice is to classify the data into classes with the same interval length, also called bins. Table 1 and Fig. 1 provide an example comprising actual thickness measurements of a coal bed. The width of all classes in this illustrative example is 1 m. When the height of the class bars is proportional to the frequency, the graph is called a histogram. A less common practice is to display the same classes as a pie chart. Typically, the histogram is supplemented with a list of frequency distribution properties explained below.

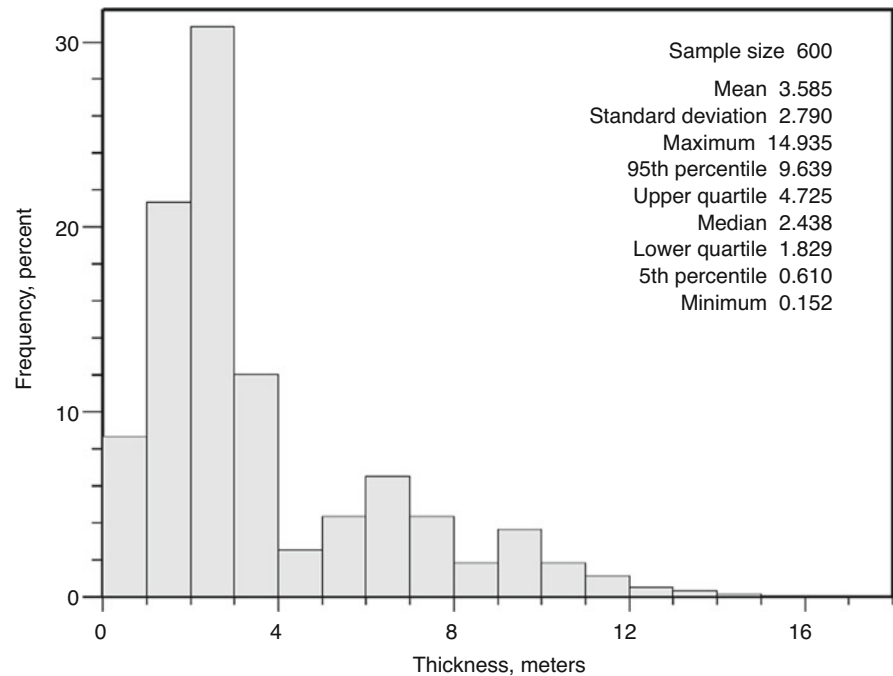
The frequency table and the histogram in particular allow to quickly capture the main features in the style of the fluctuations. In the example, the values increase rapidly from close to zero to 3 m. Approximately 60% of the values are less than 3 m, which is close to only 20% of the range in thickness. An undesirable aspect of the histogram and the associated frequency table is that their appearance changes depending on the selection of the class interval. Although the details vary, the general trend may remain the same. For example, in the case of Fig. 1, by selecting five class intervals that are each 3 m wide, the first bar would be the tallest and the other bars would decrease in height systematically. In this example, details such as the initial increase in frequency or the local

Frequency Distribution, Table 1 Thickness of the Cherokee main bench at the Little Snake River coal field and Red Desert assessment area, Greater Green River Basin, Wyoming (Scott et al. 2019). The notation n^+ denotes a small number above n . The percentages are rounded to two decimal places.

Class interval Meters	Frequency Count	Frequency Percent
0–1	52	8.67
1 ⁺ –2	128	21.33
2 ⁺ –3	185	30.83
3 ⁺ –4	72	12.00
4 ⁺ –5	15	2.50
5 ⁺ –6	26	4.33
6 ⁺ –7	39	6.50
7 ⁺ –8	26	4.33
8 ⁺ –9	11	1.83
9 ⁺ –10	22	3.67
10 ⁺ –11	11	1.83
11 ⁺ –12	7	1.17
12 ⁺ –13	3	0.50
13 ⁺ –14	2	0.33
14 ⁺ –15	1	0.17
Total	600	100.00

Frequency Distribution,

Fig. 1 Thickness histogram of the main bench of the Cherokee coal bed, Wyoming (after Scott et al. 2019). All listed values are in the same units as the horizontal axis: meters



minimum between 4–5 m disappear. On the contrary, if there are too many bars as a result of taking too many classes, the histogram becomes noisy and it will be difficult to observe the relevant information, thus eliminating the main appeal of frequency distributions. There are no definitive rules concerning the optimal number of classes, but usually between 5 and 20 classes is an adequate number.

Statistics

The values listed in the histogram of Fig. 1 are numbers provided with the intention of further summarizing the data – statistics. The term sample size is just a different name for the number of data.

The rest of the statistics are primarily measures of location along the thickness axis. For example, the sum of all 600 thickness values is 2,150.968, which when divided by the total number of observations gives a value of 3.585 m. This value is called the mean, which can be interpreted as the central tendency of the dataset. Another central tendency value is the median, which is simply the value separating the half of lowest values from the half of highest values. In our illustrative example, this is the value between the 300th and 301st observations when the dataset is sorted, which is 2.438 m. One idea, but two clearly different values for the central tendency. Neither is a better indicator than the other, but the median is less sensitive to errors – more robust – especially for small sample sizes.

The idea of sorting the data and then dividing them into groups of equal sizes can be generalized from two to any other practical number of dividers (quantiles). Two of the most widely used number of classes are represented in the listing in Fig. 1: 4 and 100. Note that there is always one less divider than the number of groups. So, for four groups, there are three dividers. The divider separating the group of the smallest values from the group of the second smallest values is the lower quartile, which in this example is the value between the 150th and 151st values when sorted in increasing order; here it is 1.829 m. The divider between the two groups with the largest values is the upper quartile, which in this case is 4.725 m. The divider between the second and the third largest groups is still the median.

The dividers in a dataset sorted by increasing numbers and broken into 100 groups are called percentiles. The divider between the 5th lowest group and the next highest one is the 5th percentile. The 95th percentile is the divider between the 5th and the 6th highest groups. In terms of percentiles, the lower quartile and the upper quartile are the 25th and the 75th percentiles, respectively.

Next, we have the summary statistics of dispersion. The standard deviation, sd , is the only one of those statistics listed on Fig. 1. The calculation is more elaborate than those for the other statistics discussed so far:

$$sd = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^n (z_i - m)^2}, \quad (1)$$

where n is the sample size, z_i is each one of the measurements, and m is the mean. The further apart are the measurements from the mean, the larger the value of the standard deviation is. The difference between the lower and the upper quartiles (interquartile range), and the difference between the 5th and the 95th percentile are two other measures of dispersion. In contrast to the standard deviation, these two intervals have an associated confidence interval: 50% of the values are between the lower and upper quartiles and 90% of them are between the 5th and the 95th percentiles. Hence, their information content is higher than that of the standard deviation.

Finally, the statistics in Fig. 1 can be used to describe the degree of symmetry of the frequency distribution with one value, which is calculated by subtracting the median from the mean. In this case, the result is 1.087 m. When the difference is positive, the frequency distribution tapers more slowly to the right; it is said to exhibit positive skewness. Positive skewness is typical in earth science datasets, particularly for geochemistry data. If the frequency distribution is symmetric, then the mean is equal to the median, but the converse is not necessarily true. Negative skewness implies a median larger than the mean.

Cautions and Warnings

Care must be taken to make sure that data have been collected properly so as to have representative samples (de Gruijter et al. 2006). In the case of the data in Fig. 1, more drilling was done where the coal bed was thicker, consequently the sampling is biased and the bias needs to be removed previous to any inference. Solutions in the presence of spatial correlation – which is the case of the example in Fig. 1 – involve weighting the data (Deutsch 1989) or selective discarding (Olea 2021).

In statistics, the collection of all possible measurements within a sampling domain is called a population sample, something close to impossible to attain in practice. A dataset such as the one in the example represents a minimal fraction of the population sample. However, ordinarily there is the interest and the need to use a partial sample to characterize the population sample. Several considerations are necessary for an adequate characterization.

Any parameter of the population frequency distribution is likely to be different from that of a partial sample; thus it is subject to uncertainty. The most convenient way to characterize the parameters of a population frequency distribution such as the median, the standard deviation, or any other parameter is to use the bootstrap method. The method provides the likely values for any of those parameters, thus allowing to see how low or how high the parameter actually could be.

If the partial sampling does not cover the entire area of interest, there is the possible complication that the frequency distribution is not going to be stationary: the frequency

distribution is different in different areas of the sampling domain. The only way to investigate or remedy this problem is with additional sampling.

Summary and Conclusions

The frequency distribution and related statistics are a straightforward way to have a dataset in a more understandable form than simply a listing of all measurements.

Cross-References

- [Bootstrap](#)
- [Interquartile Range](#)
- [Robust Statistics](#)
- [Spatial Autocorrelation](#)
- [Stationarity](#)
- [Uncertainty Quantification](#)
- [Univariate](#)

Bibliography

- de Gruijter J, Brus DJ, Bierkens MFP, Knotters M (2006) Sampling for natural resource monitoring: Statistics and methodology of sampling and data analysis. Springer, 348 pp
- Deutsch CV (1989) DECLUS: a FORTRAN 77 program for determining optimum spatial declustering weights. *Comput Geosci* 15(3): 325–332
- Everitt BS, Skondall A (2010) The Cambridge dictionary of statistics, 4th edn. Cambridge University Press, 431 pp
- Olea RA (2021) Revisiting the declustering of spatial data with preferential sampling. *Comput Geosci* 157, 12 pp. <https://doi.org/10.1016/j.cageo.2021.104946>
- Scott DC, Shaffer BN, Haacke, JE, Pierce PE, Kinney SA (2019) Coal geology and assessment of coal resources and reserves in the Little Snake River coal field and Red Desert assessment area, Greater Green River Basin, Wyoming. U.S. Geological Survey Professional Paper 1836, 169 pp. <https://doi.org/10.3133/pp1836>. Accessed 15 Sept 2019

Frequency-Wavenumber Analysis

Frits Agterberg
Central Canada Division, Geomathematics Section,
Geological Survey of Canada, Ottawa, ON, Canada

Definition

Frequency-wavenumber analysis, also known as frequency-wavenumber spectral analysis, originated in atmospheric

physics where it is used to partition waves into standing and traveling varieties, the advantage of which is extraction of the different velocities of the traveling waves. This approach uses classical Fourier-series based methods commonly used in other branches of science. The common output of frequency-wavenumber analysis is a diagram in which frequency is plotted horizontally and a measure of spectral density along the vertical axis. Diagrams of this type are useful in geoscience, especially in multifractal modeling and for the study of data obtained from homogeneous geologic entities sampled equidistantly along straight lines; e.g., by drilling bore-holes.

Introduction

Cressie and Wikle (2011) have pointed out that statistics for spatiotemporal data constitute the next frontier in applications of mathematical statistics to geoscientific data. Much progress in this field has been made in atmospheric science (cf. Geoga et al. 2018; Simons and Wang 2011). In this chapter, the emphasis will be on frequency-wavenumber analysis as applied to equidistant solid Earth samples for which exact time measurements are not generally available. Time in solid Earth geoscience applications covers a broad spectrum ranging from nearly instantaneous events (e.g., earthquakes, volcanic eruptions, turbidity currents) to different types of long-term processes measured in millions of years with (or without) age determinations that are accompanied by relatively wide error bars. However, it is possible to apply methods of frequency-wave number analysis to sequences of chemical composition data and other measurements on rock samples. Spectral analysis of such data can provide valuable new insights. In addition to normal spectral analysis, attention in this entry will also be paid to applications of multifractal theory including frequency-wavenumber analysis without time measurements as originally derived in atmospheric physics (cf. Lovejoy and Schertzer 2013).

Sampling of Rocks: Pulacayo Zinc Deposit Example

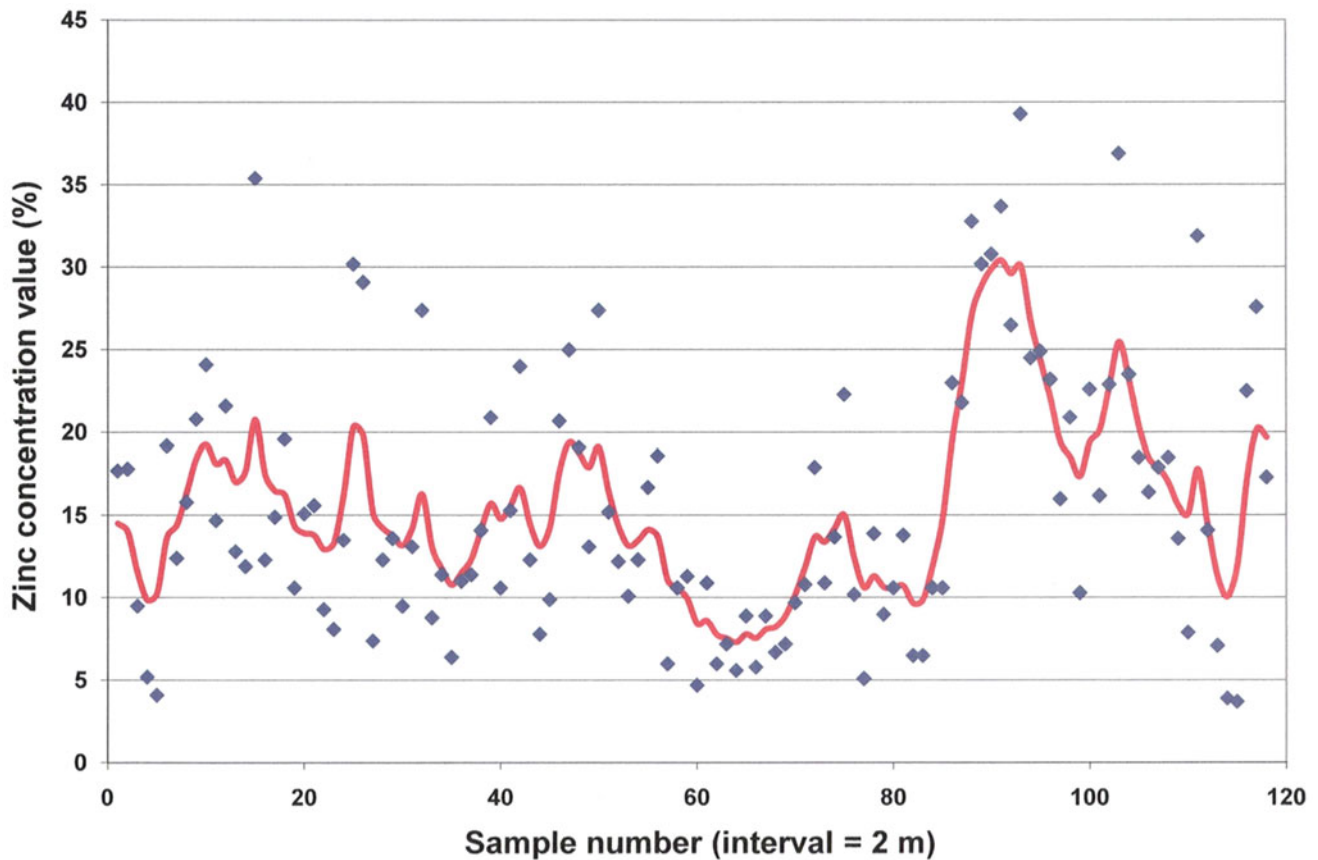
Sampling of rocks at the surface of the Earth and performing measurements on them can be difficult because of limited exposure of bedrock but drilling often provides continuous sequences of measurements. At the microscopic scale, most rocks are heterogeneous consisting of different substances (e.g., different kinds of crystals) and volume of rock samples should not be too small. The purpose of the example given in this entry is to illustrate several of the problems encountered

when rock samples are collected for the purpose of chemical analysis. Normally, rocks are crushed before their chemical composition is determined on samples taken from the resulting powder. This aspect of rock sampling has been studied in detail by geochemists, chemists, and mineral engineers (see, e.g., Gy 2004). The primary example used in this entry concerns the spatial distribution of sphalerite crystals in relatively large “channel” samples from the Pulacayo mine in Bolivia.

De Wijs (1951) used a series of 118 zinc concentration values averaging 15.61% Zn from samples taken at a regular (2 m) interval along a horizontal drift on the 446-m mining level of the Pulacayo Mine in Bolivia. Level depth was measured downward from elevation of a tunnel along which the 2.7 km wide Tajo vein was discovered in 1883 and mined until 1956. This series shown graphically in Fig. 1 has been used extensively for later studies by many authors including Matheron (1962), Agterberg (1961, 1974, 2014), and Lovejoy and Schertzer (2007). The zinc values in Fig. 1 are for 1.3-m wide channel samples cut across this steeply dipping vein-type silver-zinc ore deposit.

Average sphalerite (zinc sulfide) content in the Tajo vein increased downward. Maximum possible zinc content is about 66% which is greater than the largest value of 39.3% zinc shown in Fig. 1. This is because the sampled material consisted not only of massive sulfide but also included mineralized wall rock, so that the largest possible value is considerably less than 66%. On the 446-m level, average thickness of massive vein filling averaged only 0.50 m but wall rocks on both sides of the vein contained disseminated sphalerite, partly occurring in subparallel stringers. The channel samples were cut over the standard width of 1.30 m that was the expected mining width. In size, they significantly exceed that of bore-hole samples obtained by drilling, thus reducing chemical measurement errors.

Figure 1 also shows a smoothed version of the 118 zinc values. It was based on the autocorrelation function shown in Fig. 2. The signal-plus-noise method used for the smoothing assumes that each zinc value is the sum of a “signal” component with continuous change along the series plus a random white-noise component. Together these two components produced an autocorrelation function of the type $\rho_h = c \cdot \exp(-ah)$ where h represents distance along the drift. Several other statistical methods such as simple moving averaging, ordinary kriging, or inverse distance weighting could be used to produce smoothed patterns similar to that shown in Fig. 1. Agterberg (1961) initially fitted a harmonic trend function to the zinc values using ordinary Fourier expansion that resulted in a curve which was smoother than those obtained by the other methods.



Frequency-Wavenumber Analysis, Fig. 1 Pulacayo Mine zinc concentration values for 118 2-m channel samples along horizontal drift samples (original data from De Wijs 1951). Sampling interval is 2 m. “Signal” retained after filtering out “Noise.” (Source: Agterberg 2012, Fig. 1)

Short-Distance Nugget Effect Modeling

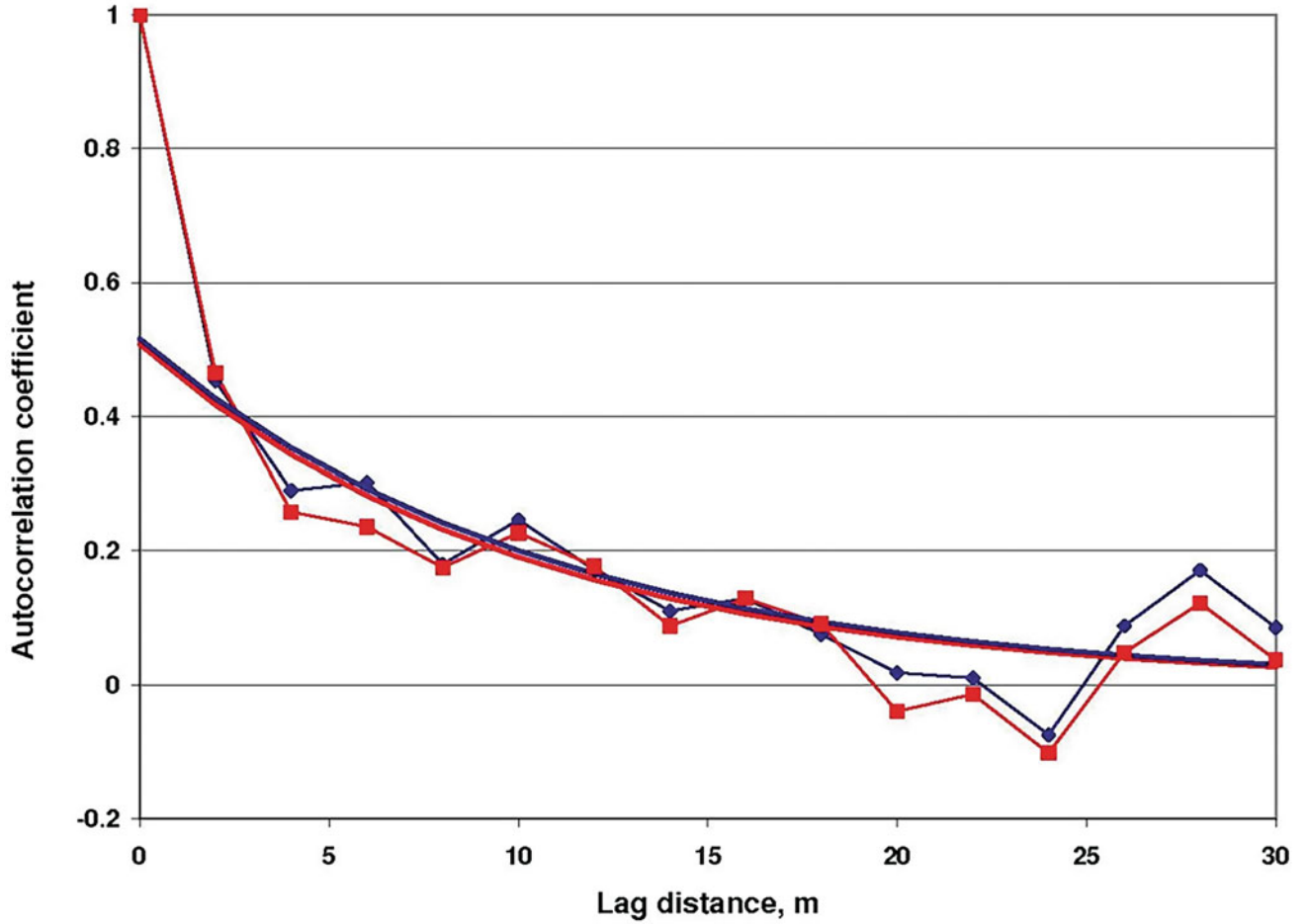
In the preceding section, it was pointed out that there is uncertainty associated with the definition of “effective” channel sample length (L) in the Pulacayo Mine. Matheron (1963) set $L = 0.5$ because the Tajo vein has (horizontally measured) thickness of only 0.50 m on the 446-m level. This particular choice of L resulted in average zinc concentration estimates for lag distances greater than 2 m (up to 400 m). For shorter lag distances, however, it is useful to generalize Matheron’s concept by defining $A(L)$, which depends on the value of L .

In Agterberg (2012, Fig. 10) estimates of $A(L)$ are shown for effective channel sample lengths less than 0.5 m. For $L > 3$ cm, $A(L)$ increases slightly from about 0.01 to 0.0165 at $L = 0.5$ m; for $L < 0.03$ m, there is rapid decrease to $A(L) = 0$. Sums of squared deviations from lines of best fit for different values of L showed that the optimum solution $\alpha(L) = 0.021$ is obtained at $L = 13$ cm (Agterberg 2012). This implies that the negative exponential autocorrelation function previously used (see, e.g., Fig. 2) is too simple for short

distances ($h < 2$ m). The true pattern is closer to that shown in Fig. 3a, which differs from the earlier model in that strong autocorrelation is assumed to exist over very short distances. Rapid increase in autocorrelation near the origin is probably caused by localized clustering of ore crystals, although at the microscopic scale there remains rapid decorrelation related to crystal shape and measurement error. The model of Fig. 3 is an example of a nested semi-exponential autocorrelation function as previously used Serra (1966). It satisfies the equation:

$$\rho(h) = c_1 e^{-a_1 h} + c_2 e^{-a_2 h}$$

The coefficients in the first term are $c_1 = 0.5157$ and $a_1 = 0.1892$ as in Fig. 1. The second term represents the strong autocorrelation due to clustering over very short distances. Decorrelation at microscopic scale can be represented by a small white noise component (probably augmented with a random measurement error) with variance equal to



Frequency-Wavenumber Analysis, Fig. 2 Estimated autocorrelation coefficients for original data (in red) and logarithmically transformed Pulacayo zinc values (in blue), shown together with best-fitting negative exponential autocorrelation functions. The red and blue curves for

original and transformed data coincide approximately illustrating that logarithmic transformation of the original data does not significantly affect autocorrelation in this application. (Source: Agterberg 2012, Fig. 5)

$c_0 = 0.0208$ (Agterberg 2014, Section 11.3.1). The coefficient c_2 in the second term on the right side of the preceding equation satisfies $c_2 = 1 - c_0 - c_1 = 0.4635$. Choosing $a_2 = 2$ provides a good fit over the entire observed correlogram (Fig. 1). It affects extrapolation toward the origin with $h < 2$ m only. Figure 3b shows that the second term on the right side of the preceding nested design equation could not be detected in the correlogram for sampling intervals greater than 2 m.

The normalized power spectrum corresponding to the nested design of Fig. 3 is:

$$P(f) = \sum_{i=1}^2 \left[1 - c_i + \frac{c_i / \pi f_{ci}}{1 + (f/f_{ci})^2} \right]$$

where $f_{ci} = a_i / 2\pi$. A log-log plot of this spectrum is shown in Fig. 4 adopting the coefficients previously used for Fig. 3. The curve in Fig. 4 is approximately a straight line for lower

frequencies, but for high frequencies, there is a marked decrease of slope reflecting the nugget effect.

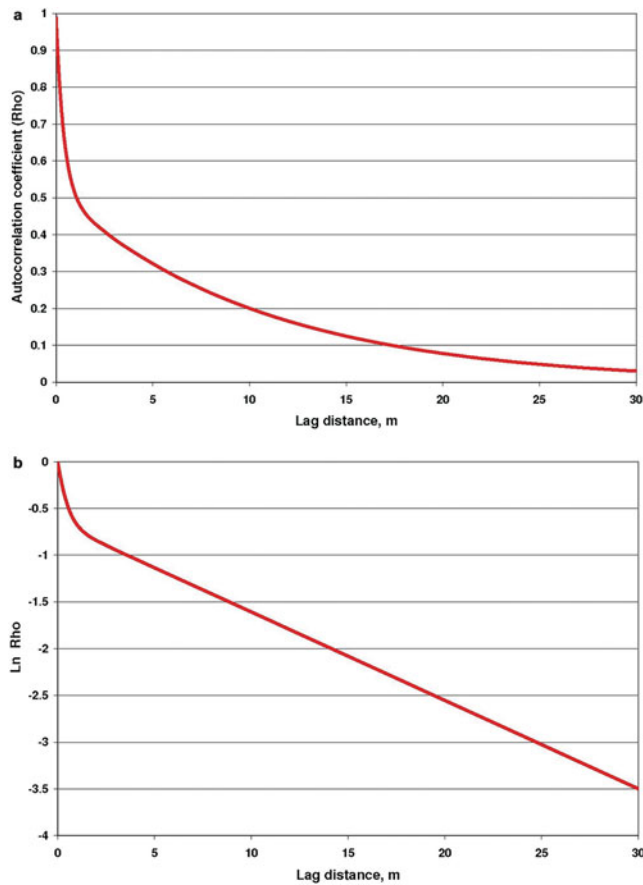
If n is even and the origin is chosen in the center of a series y_k ($k = 1, 2, \dots, n$), it can be written as:

$$y_k = a_0 + 2 \sum_{i=1}^{1/2n-1} \left[a_i \cos\left(\frac{2\pi i k}{n}\right) + b_i \sin\left(\frac{2\pi i k}{n}\right) \right] + a_{1/2n} \cos(\pi k)$$

Using amplitude and phase representation, this becomes:

$$y_k = R_0 + 2 \sum_{i=1}^{1/2n-1} \left[R_i \cos\left(\frac{2\pi i k}{n} + \varphi_i\right) + R_{1/2n} \cos(\pi k) \right]$$

where $R_i = \sqrt{a_i^2 + b_i^2}$ and $\varphi_i = \arctan\left(\frac{b_i}{a_i}\right)$. The periodogram is a plot of $P_i = R_i^2$ against I that is distributed as χ^2 with 2 degrees of freedom. The average of q consecutive



Frequency-Wavenumber Analysis, Fig. 3 Hypothetical autocorrelation function consisting of two negative exponentials to incorporate nugget effect over short lag distances h with $h < 2$ m. Autocorrelation function for nugget effect is superimposed on negative exponential curve for distances $h \geq 2$ m. (a) Vertical scale is linear; (b) vertical scale is logarithmic. (Source: Agterberg 2012, Fig. 12)

periodogram values provides an estimate of the power spectrum $E(P_i)$ that is distributed as χ^2 with approximately $2q$ degrees of freedom. The power spectrum of a continuous series is:

$$P_f = s^2(x) \int_{-\infty}^{\infty} r_h \cos(2\pi fh) dh$$

where r_h is the autocorrelation function. For example, if $r_h = c \cdot \exp.(-a \cdot |h|)$:

$$P(f) = (1 - c)s^2(x) + \frac{c \cdot s^2(x)/\pi f_c}{1 + (f/f_c)^2}.$$

Figure 5 shows the power spectrum $E(P_i)$ estimated by averaging pairs of consecutive values of the periodogram for the 118 zinc values together with a quadratic curve fitted by least squares. A best-fitting straight line for the same values results in $\beta = 0.72$ that can be shown to be significantly better

than the linear fit (for level of significance $\alpha = 0.01$). The slope of the curve at the origin in Fig. 5 gives $\beta = 1.18$ with gradually decrease to 0.49 at maximum log wave number on the right.

Universal Multifractals

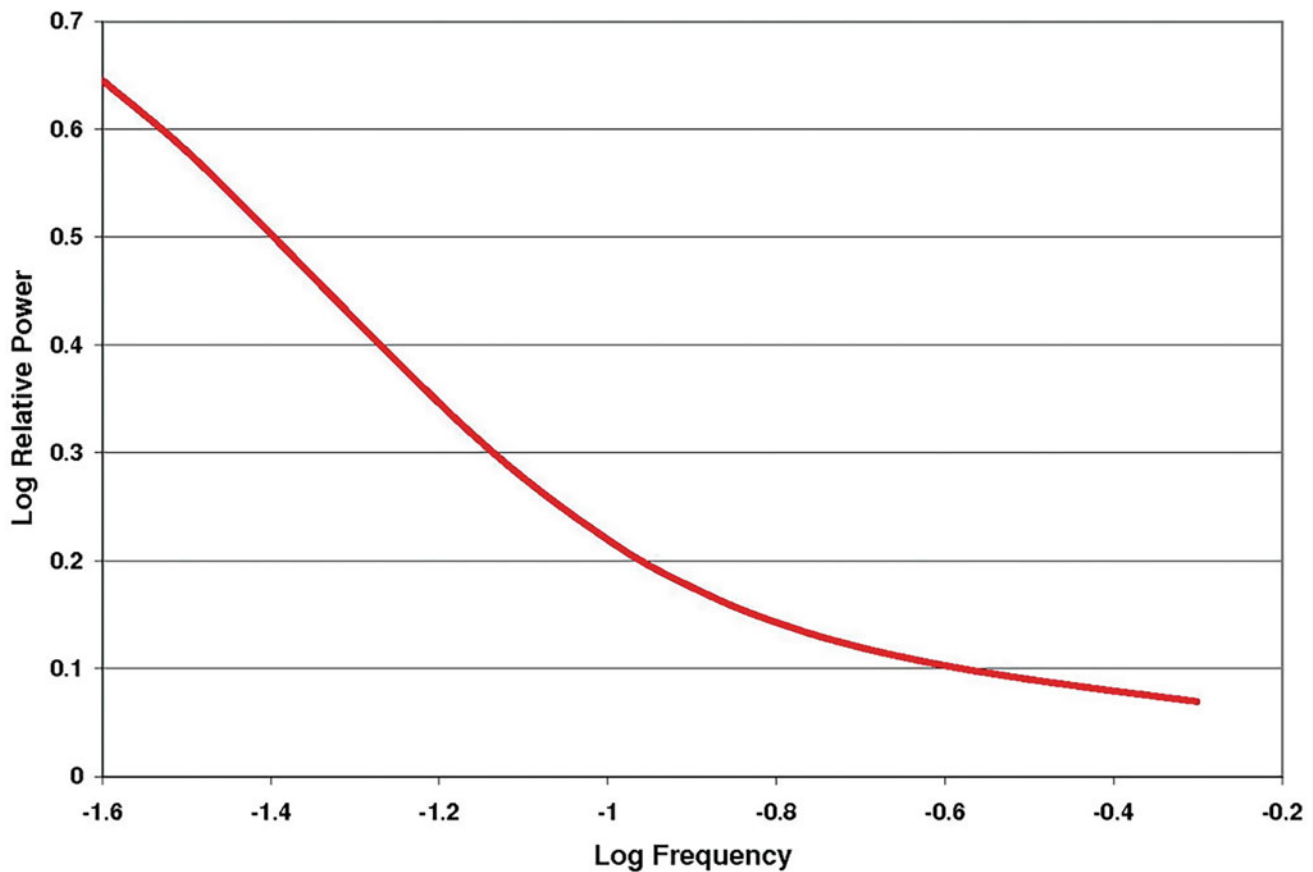
Already in the 1980s, Schertzer and Lovejoy (1985) pointed out that the so-called binomial/ p model (also known as the model of de Wijs) can be regarded as a “micro-canonical” version of their α -model in which the strict condition of local preservation of mass is replaced by the more general condition of preservation of mass within larger neighborhoods (preservation of ensemble averages). Cascades of this type can result in lognormal distributions with Pareto tails as are encountered in the geosciences (cf. Agterberg 2021). In the three-parameter Lovejoy-Schertzer α -model, α (bold alpha) does not represent the singularity α but represents the so-called Lévy index that, together with the “codimension” C_1 and the “deviation from conservation” H , characterizes a universal multifractal field. Examples of applicability of universal multifractals to geological processes that took place in the Earth’s crust have been given by Lovejoy and Schertzer (2007).

The model originally developed by Schertzer and Lovejoy (1991) is based on the concept of a multifractal field ρ_λ with codimension $C_1(\gamma)$ where $\lambda = L/\epsilon$ is the scale ratio with L representing the largest scale that can be set equal to 1 without loss of generality (Cheng and Agterberg 1996). The field ρ_λ can be a flux or density in physics. In geochemical applications, it is the element concentration value of a chemical element. It is characterized by its probability distribution $P(\rho_\lambda \geq \lambda^\gamma \propto \lambda^{-C_1(\gamma)})$ with statistical moments $E(\rho_\lambda^q) \propto \lambda^{K(q)}$. The relations between $K(q)$, $C_1(\gamma)$, and the field order γ are:

$$K(q) = \max_{\gamma} \{q \cdot \gamma - C_1(\gamma)\}; C_1(\gamma) = \max_q \{q \cdot \gamma - K(q)\}$$

Cheng and Agterberg (1996) have made a systematic comparison of the binomial/ p model with the universal multifractal model of Schertzer and Lovejoy (1991). These authors have successfully applied their model to the 118 Pulacayo zinc values as will be discussed in the next.

The three-parameter α -model is a generalization of the conservative two-parameter $(0, C_1, \alpha)$ universal canonical multifractal (Schertzer and Lovejoy 1991, p. 59) model. So-called extremal Lévy variables play an important role in a two-parameter approach. In their large-value tail, these random variables have probability distributions of the Pareto form $P(X \geq x) \propto |x|^{-\alpha}$ ($\alpha < 2$). They can be used to generate multiplicative cascades starting from uniform random variables as follows. Suppose W is a uniform random variable within the interval $[0, 1]$ and $V = W^{-1/\alpha}$; so that the probability $P(V = v) = \alpha \cdot v^{-1-\alpha}$ if $v \geq 1$, and $P(V = v) = 0$ if $v < 1$ (cf. Wilson et al. 1991, APPENDIX B). Then, $\sum_{i=1}^n V_i$ representing the sum of n such random variables V_i ($i = 1,$



Frequency-Wavenumber Analysis, Fig. 4 Relative power spectrum for autocorrelation function of Fig. 3. Decrease in slope at higher frequency side is caused by the superimposed nugget effect. (Source: Agterberg 2012, Fig. 24)

2, ..., n) has a Lévy limit distribution with high-value tails. The Lévy index α defines another index α' by means of the relation $1/\alpha + 1/\alpha' = 1$. Suppose that μ and c are two constants and a new random variable is defined as $Y = c \cdot (\mu - V)$. Then, for large n : $Y = \frac{1}{n^{1/\alpha}} \sum_{i=1}^n c \cdot \left(\frac{\alpha}{\alpha-1} \cdot w_i^{-1/\alpha} \right)$ with Laplacian characteristic function $E(e^{Y \cdot q}) = \exp \{ \alpha \cdot \Gamma(-\alpha) \cdot (cq)^\alpha \}$. It follows, after some manipulation, that $c = \left(\frac{C_1}{\Gamma(2-\alpha)} \right)^{1/\alpha}$ resulting in the two-parameter universal form:

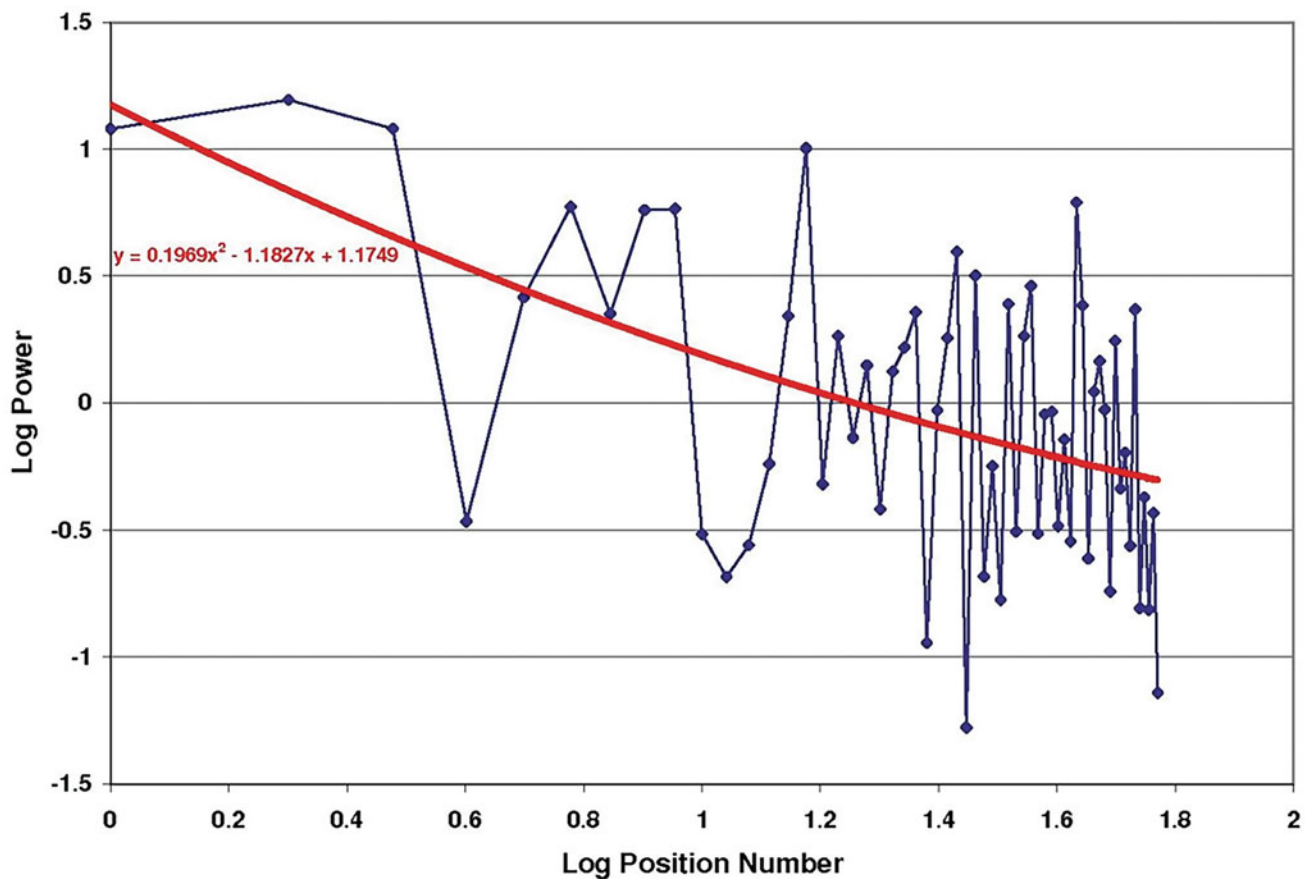
$$K(q) = \frac{C_1 \cdot \alpha'}{\alpha} (q^\alpha - q); \quad 0 \leq \alpha \leq 2; 0 < C_1 < E.$$

The Lévy index α characterizes the degree of multifractality and the codimension C_1 characterizes the sparseness of the mean field (Lovejoy and Schertzer 2010). In the three-parameter Lovejoy-Schertzer α -model, a third parameter H is added that characterizes the deviation from conservation. A space-time universal multifractal field for element concentration can be subject to change of its mean value in the course of time. Even in static 3D situations, incorporation of H into the model provides increased flexibility. Lovejoy and

Schertzer (2007) provide examples of geophysical data for which the universal multifractal approach is applicable.

Figure 6 taken from Lovejoy and Schertzer (2007) shows a realistic universal multifractal simulation for the Pulacayo orebody using the following three parameters: Lévy index $\alpha = 1.8$, codimension $C_1 = 0.03$, and deviation from conservation $H = 0.090$. This approach is explained in detail and illustrated by means of other applications in a large number of publications including Lovejoy and Schertzer (2013). The codimension C_1 , which characterizes sparseness of mean field, and deviation from conservation H can be derived as follows.

First a log-log plot of the so-called “first order structure function” is constructed. Successive moments are obtained for absolute values of differences between concentration values for points that are distance h apart by raising them to the powers q ($= 0.25, 0.5, \dots, 3$ for the 118 zinc values). The resulting pattern for $q = 2$ represents the variogram and the first point on the pattern for $q = 0$ is the de Wijs index of dispersion d . Straight lines are fitted to all patterns and a new diagram is constructed with the slopes of the lines (ξ_q) plotted against q . Slope and value of this new line near $q = 1$ yielded $H = 0.090$ and $C_1 = 0.02$ for the Pulacayo orebody because



Frequency-Wavenumber Analysis, Fig. 5 Periodogram of 118 zinc values with quadratic curve fitted by least squares (Logarithms base 10). The flattening of the curve toward higher frequencies is believed to be due to the nugget effect. (Source: Agterberg 2012, Fig. 25)

$H = \xi_l$ and $C_1 = H - \xi'_1$ where ξ'_1 is the first derivative of ξ_q with respect to q (Fig. 7).

Use of the so-called “double trace moment” method yielded estimates of the Lévy index equal to $\alpha = 1.76$ and $\alpha = 1.78$, and codimension $C_1 = 0.023, 0.022$, respectively (Fig. 8). In general, a relatively small value of C_1 with respect to H indicates that the multifractality is so weak that deviation from conservation (H) will be dominant except for quite large moments (Lovejoy and Schertzer 2007, p. 491). In the Pulacayo application, the estimated value of H is small and the two-parameter model of de Wijs would be approximately satisfied. As outlined before, the binomial/ p model produced inconsistencies between results for very small and very large moments. Universal multifractal modeling is more flexible and produces realistic zinc concentration variability. On the other hand, the estimate for the second-order moment obtained by the method of moments $\tau(2) = 0.979 \pm 0.038$ produced a realistic autocorrelation function including the nugget effect.

Frequency-wavenumber analysis is an important tool in universal multifractal modeling. Theoretically, the Lovejoy-Schertzer approach results in a spectrum consisting of a

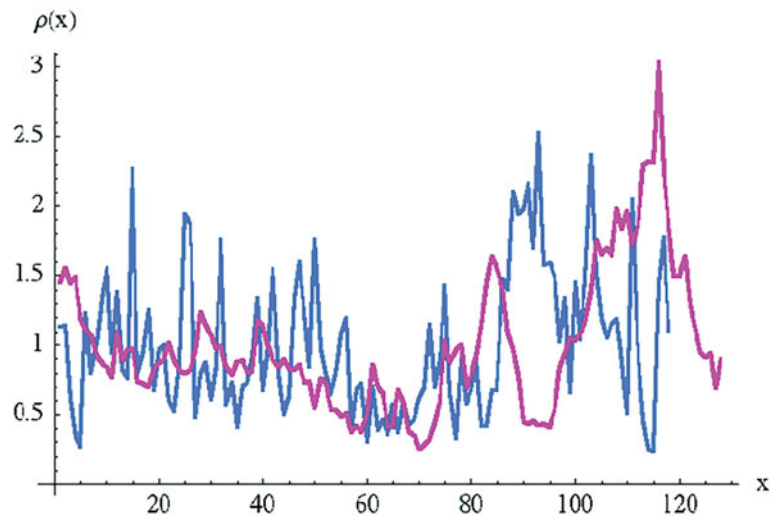
straight line with slope $-\beta$. This parameter can either be estimated directly or indirectly using $\beta = 1 - K_2 + 2H$ where K_2 representing the “second characteristic function.” Lovejoy and Schertzer (2007) estimated $K_2 = 0.05$ by double trace moment analysis. With the previously mentioned estimate $H = 0.090$, this yielded $\beta \approx 1.12$ in good agreement with the experimental spectrum for the 118 zinc values previously shown in Fig. 5.

Summary and Conclusions

Frequency-wavenumber analysis is used in atmospheric physics to partition waves into standing and traveling varieties. The approach uses classical Fourier-series based methods. The common output of frequency-wavenumber analysis is a diagram in which frequency is plotted horizontally and a measure of spectral density along the vertical axis. This entry was concerned with applications of frequency-wavenumber analysis in solid Earth geoscience without use of the dimension of time for which sufficiently precise data are not available. Spectral analysis and multifractal modeling

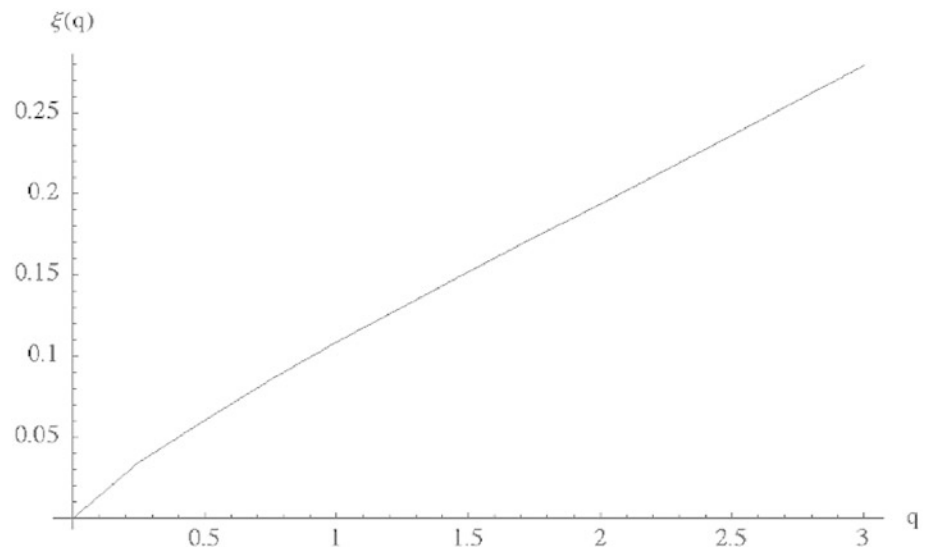
Frequency-Wavenumber Analysis, Fig. 6

Result of experimental multifractal experiment by Lovejoy and Schertzer (2007, Fig. 3a). Blue: original Pulacayo zinc values, with x representing horizontal distance in units of 2 m. Pink: universal multifractal simulation results for Levy index $\alpha = 1.8$; codimension $C_1 = 0.03$ characterizing the “sparseness” of the mean field; and deviation from conservation $H = 0.090$. Both patterns were normalized to unity (mean = 15.6% Zn). (Source: Agterberg 2014, Fig. 12.33)



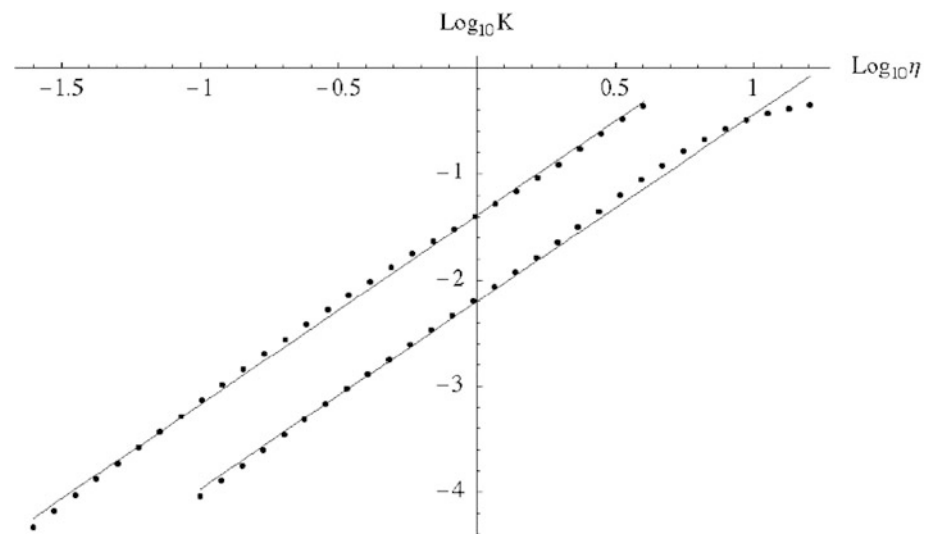
Frequency-Wavenumber Analysis, Fig. 7

Slopes $\xi(q)$ according to Lovejoy and Schertzer (2007, Fig. 26a). Slope and value near $q = 1$ yield $H = 0.090$, $C_1 = 0.018$, because $H = \xi(1)$ and $C_1 = H - \xi'(1)$; H represents deviation from conservation of pure multiplicative process for which $H = 0$. (Source: Agterberg 2014, Fig. 12.35)



Frequency-Wavenumber Analysis, Fig. 8

Lovejoy and Schertzer (2007, Fig. 27)'s double trace moment analysis of the Pulacayo zinc values with $q = 2$ (left), $q = 0.5$ (right). Slopes yield Levy index $\alpha = 1.76, 1.78$; and codimension $C_1 = 0.023, 0.022$, respectively. (Source: Agterberg 2014, Fig. 12.36)



were discussed and applied to the spatial distribution of zinc concentration values for equidistant elongated samples from the Pulacayo mine in Bolivia that previously have been studied by relatively many authors who took different points of view but reached similar results.

Cross-References

- Autocorrelation
- Fast Fourier Transform
- Moments
- Multifractals
- Signal Analysis
- Spectral Analysis

Bibliography

- Agterberg FP (1961) The skew frequency curve of some ore minerals. *Geol Mijnb* 40:149–162
- Agterberg FP (1974) *Geomathematics*. Elsevier, Amsterdam
- Agterberg FP (2012) Sampling and analysis of element concentration distribution in rock units and orebodies. *Nonlinear Process Geophys* 19:23–44
- Agterberg FP (2014) *Geomathematics: theoretical foundations, applications and future developments*. Springer, Dordrecht
- Agterberg F (2021) Aspects of regional and worldwide mineral resource prediction. *Jour Earth Science*. <https://doi.org/10.1007/312583-020-1397-4>
- Cheng Q, Agterberg FP (1996) Comparison between two types of multifractal modeling. *Math Geol* 28:1001–1015
- Cressie N, Wikle CK (2011) *Statistics for spatio-temporal data*. Wiley, New York
- De Wijs HJ (1951) Statistics of ore distribution I. *Geol Mijnb* 13: 365–375
- Geoga CJ, Haley CL, Siegel A, Anitescu M (2018) Frequency-wavenumber spectral analysis of spatiotemporal flows. *J Fluid Mech* 848:545–559
- Gy P (2004) 50 years of Pierre Gy's "Theory of sampling – a tribute". *Chemom Intell Lab* 74:61–70
- Lovejoy S, Schertzer D (2007) Scaling and multifractal fields in the solid Earth and topography. *Nonlinear Process Geophys* 14:465–502
- Lovejoy S, Schertzer D (2010) On the simulation of continuous in scale universal multifractals, Part I: spatially continuous processes. *Comput Geosci* 36:1393–1403
- Lovejoy S, Schertzer D (2013) *The weather and climate*. Cambridge University Press
- Matheron G (1962) *Traité de géostatistique appliquée*, vol 14. Mémoires BRGM, Paris
- Schertzer D, Lovejoy S (1985) The dimension and intermittency of atmospheric dynamics multifractal cascade dynamics and turbulent intermittency. In: Launder B (ed) *Turbulent shear flow*. Springer-Verlag, New York, pp 7–33
- Schertzer D, Lovejoy S (eds) (1991) *Non-linear variability in geophysics*. Kluwer, Dordrecht
- Serra J (1966) Remarques sur une lame mince de minerai lorrain. *Bulletin BRGM* 6:1–36
- Simons FJ, Wang DV (2011) Spatospectral concentration in the Cartesian plane. *Int J Geomath* 2(1):1–36
- Wilson J, Schertzer D, Lovejoy S (1991) Continuous multiplicative cascade models of rain and clouds. In: Schertzer D, Lovejoy S (eds) *Non-linear variability in geophysics*. Kluwer, Dordrecht, pp 185–207

Full Normal Plot

Gianna Serafina Monti¹ and Glòria Mateu-Figueras^{2,3}

¹Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy

²Department of Informatics, Applied Mathematics and Statistics, University of Girona, Girona, Spain

³Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Girona, Spain

Definition

The Ful normal plot (FUNOP) is a graphical technique proposed by Tukey (1962) in his lead article *The Future of Data Analysis* useful in delineating outliers or other singularity in the data, in an automatic way. The FUNOP is often used as a starting point for a “vacuum cleaning” process, namely, a regression procedure for cleaning residuals (Anscombe and Tukey 1963).

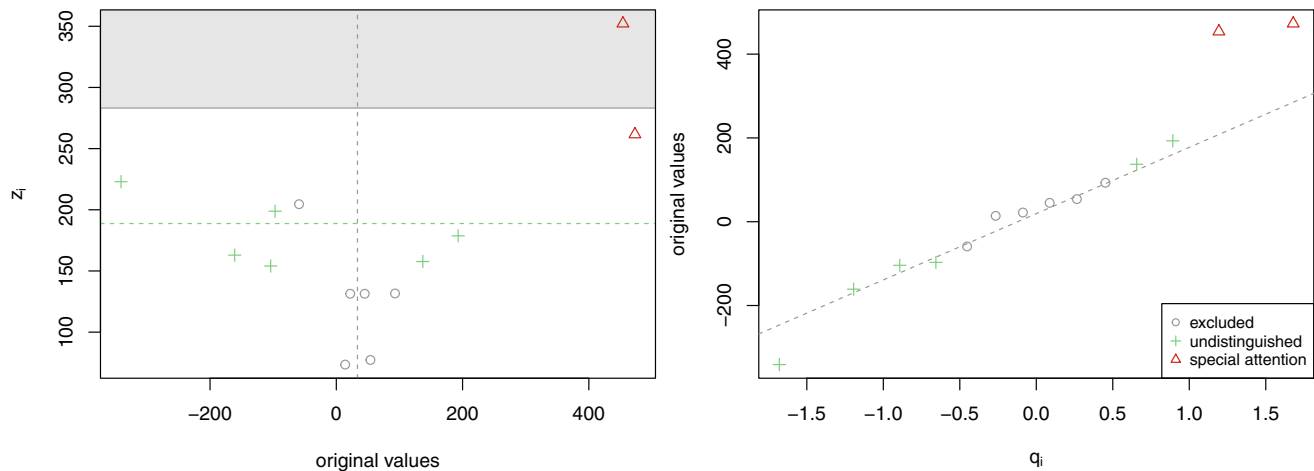
The Method

Let $y_1, \dots, y_n$ be the n observed ordered values of a variable of interest. Let $\tilde{y}$ be their median (or let $\check{y}$, the 33% trimmed mean of the y_i). For $i \leq \frac{n}{3}$ or $i > \frac{2n}{3}$, let's define the Judd values as the deviation of the observations from the median divided by the standard normal deviate corresponding to the plotting-point percentage, thus

$$z_i = \frac{y_i - \tilde{y}}{q_i}, \quad (1)$$

or $z_i = \frac{y_i - \check{y}}{q_i}$, where q_i is the quantile of the standard normal distribution of order $i/(n+1)$, i.e., $q_i = \Phi^{-1}(i/(n+1))$, where Φ is the cumulative distribution function of a standard normal distribution or more conveniently $q_i = \Phi^{-1}((3i-1)/(3n+1))$ (Blom 1958, for numerical details). Let $\tilde{z}$ be the median of the Judds thus obtained (about two-thirds of the original sample size).

Note that each Judd value could be interpreted as the slope of the secant leading from the midpoint $(0, \tilde{y})$ to (q_i, y_i) and they are worthy of attention the more they deviate from the median point. By inspecting the Judds associated to the top and bottom thirds of the sorted population only, one can see if



Full Normal Plot, Fig. 1 Judds plotted against sorted original values (left) and sorted original values plotted against the theoretical distribution q_i (right). Tukey's example data

any of the observations seem to be unusually large or small, in particular Tukey recommends to give special attention to points whose slope is 1.5 or 2.0 times larger than the median (i.e., z_i for which $z_i \geq B_z$ with B equals to 1.5 or 2.0) (Koch and Link 1970–71). For the choice of B value, Tukey (1962) recognizes that it is going to be a matter of judgement and other values could be explored.

For data sets with small n , Tukey also recommend to give “special attention” to z_j with j more extreme than i for which z_i is selected following the previous criteria.

Application

Tukey's example (Tukey 1962) is used here to illustrate in practice the FUNOP technique. Figure 1 (left) shows Judds plotted against sorted original values. Grey dotted points refer to the middle-third of original values excluded from the analysis, and their median is indicated by the gray dotted line, green crosses represent non-suspicious data which belong to the upper and lower third of the sorted values according to Eq. (1), whereas red triangles mark special points which should deserve special attention. The dashed gray area in Fig. 1 (left) represents the region of points which exceeded 1.5 times the median z , indicated by the green horizontal dotted line. The first red triangle lies in this dashed gray area, whereas the second one is highlighted as it is more extreme than the ones already identified.

Figure 1 (right) shows sorted original values plotted against the theoretical distribution q_i . It is the standard Q-Q plot for normality, and ideally, the points should lie near the straight line which represents percentiles for the standard normal distribution. The red triangles, automatically marked by the FUNOP procedure as suspicious to be outliers, deviate

sharply from this line. FUNOP technique is successfully implemented in the R package vacuum Sielinski (2020).

The FUNOP to identify outliers is used by the FUNOR (full normal rejection) and FUNOM (full normal modification) procedures to treat them in the “vacuum cleaner” process (Tukey 1962). The usefulness of this procedure in an earth science context with geological data is discussed in Koch and Link (1970–71).

Summary

In this chapter the FUNOP, an automated procedures for detecting outliers, was presented as both general and accessible graphical technique in geosciences. The pattern of points in the plot suggest which observations should deserve “special attention.” It is noteworthy that the efforts of Tukey in detecting and dealing with outliers influenced succeeding Huber's work on robust estimation (Huber 1964; Hampel 1991).

However, we observe that, although presented in a fundamental article for the future of statistics, FUNOP is not very used in practice.

Cross-References

- Normal Distribution
- Robust Statistics
- Statistical Outliers

References

- Anscombe FJ, Tukey JW (1963) The examination and analysis of residuals. *Technometrics* 5(2):141–160

- Blom G (1958) Statistical estimates and transformed Beta variables. Wiley, New York
- Hampel F (1991) Introduction to P.J. Huber (1964): robust estimation of a location parameter. In: Kotz S, Johnson NL (eds) Breakthroughs in statistics, vol 2. Springer-Verlag, New York
- Huber PJ (1964) Robust estimation of a location parameter. Ann Math Stat 35(1):73–101
- Koch GS, Link RF (1970–71) Statistical analysis of geological data, vol 2. Wiley, New York
- Sielinski R (2020) vacuum: Tukey's Vacuum Cleaner. URL <https://CRAN.R-project.org/package=vacuum>, r package version 0.1.0
- Tukey JW (1962) The future of data analysis. Ann Math Stat 33(1):1–67

Fuzzy C-Means Clustering

Jaya Sreevalsan-Nair

Graphics-Visualization-Computing Lab, International
Institute of Information Technology Bangalore, Electronic
City, Bangalore, Karnataka, India

Definition

Clustering is a classical data mining method applied on a set of items to group like items, thus separating unlike items. Most clustering algorithms give a definite mapping of an item to a cluster, referred to as *hard clustering*. But in real world, an item can belong to multiple clusters with varying degrees of affinity to the cluster, given by a membership score. Consider that n items x_i , where $i = 0, 1, \dots, (n-1)$, in a dataset X are grouped in c clusters, given by C_j for $j = 0, 1, \dots, (c-1)$. Then, each item x_i of the finite set X has the flexibility of mapping to one or more clusters. Hence, a membership or affinity score is introduced here, u_{ij} , which is for item x_i with respect to C_j . This type of clustering is called *soft clustering*. Since it brings in an element of uncertainty by assigning probability values, i.e., u_{ij} , for mapping an item to a cluster, soft clustering is also called fuzzy clustering. An item must belong to one or more clusters, discarding the possibility of an item not belonging to any cluster, thus giving $u_{ij} > 0, \forall i \in [0, n-1]$ and $\forall j \in [0, c-1]$. This specific requirement along with the rule of sum of probabilities of all possibilities of the event, i.e., the item to a cluster mapping, also gives $\sum_{j=0}^{c-1} u_{ij} = 1.0, \forall i \in [0, n-1]$. Thus, the membership matrix $U = \{u_{ij}\}$ of size $n \times c$ provides the fuzzy c-partition (Bezdek 1981). Fuzzy c-means (FCM) clustering is an algorithm that provides the fuzzy c-partition.

Hard clustering can be derived from soft clustering by assigning rules, such as binarizing the membership score using a threshold, that will constraint the mapping of an

item to a single cluster. The binarization of the membership matrix leads to the outcome of compact well-separated clusters (Dunn 1973). If the constraints on membership scores are relaxed, such that $0 < u_{ij} \leq 1$ and $0 < \sum_{j=0}^{c-1} u_{ij} \leq c$, then the partition outcome is that of *possiblistic c-means*.

Overview

The application of the theory of fuzzy sets to cluster analysis by Ruspini in 1969 bore the earliest resemblance to the FCM algorithm (Bezdek 1981), the theory of fuzzy sets. The fuzzy set theory was developed in the 1960s, by several researchers independently as an alternative to the classical set theory. Particularly, Zadeh introduced an axiomatic structure, referred to as *fuzzy set*, in 1965, to combine the concepts of *imprecision* and *information* in mathematical models of grouping of entities. The aforementioned membership matrices with respect to soft and hard clustering correspond to the fuzzy and classical set theories, respectively. The original version of the FCM algorithm is attributed to the introduction of the fuzzy versions of the ISODATA clustering algorithm, which is a variant of K-means partitioning (Dunn 1973).

FCM is an iterative algorithm that uses optimization for finding the solution of the membership matrix U . Thus, FCM introduces a cost function $J_m(U, \mathbf{v})$ for the membership matrix U and the vector of cluster centers $\mathbf{v} = \{c_0, c_1, \dots, c_{(c-1)}\}$ in the m^{th} iteration. J_m has to be minimized during the algorithm, similar to K-means algorithm (see the article on “K-Means algorithm” for more details). Referred to as the fuzzy c-means functional, $J_m(U, \mathbf{v})$ is given by:

$$J_m(U, \mathbf{v}) = \sum_{i=0}^{(n-1)} \sum_{j=0}^{(c-1)} u_{ij}^m d_{ij}^2, \text{ where } d_{ij}^2 = \|x_i - c_j\|_2^2, \text{ and}$$

$$u_{ij} = \left(\sum_{k=0}^{(c-1)} \left(\frac{d_{ij}}{d_{ik}} \right)^{\frac{2}{m-1}} \right)^{-1}.$$

In later variants of the FCM algorithm, the distance d_{ij}^m can flexibly use diagonal norm or Mahalanobis distance, etc. in addition to the Euclidean or L_2 norm. The pairwise distance of items is computed between s -dimensional feature vector of the items. The FCM algorithm, as originally specified in (Bezdek 1973), is similar to the K-means algorithm. The FCM algorithm has two main steps of initialization and update. In the initialization step, c is fixed, such that $1 < c < n$, and the cluster centers in $\mathbf{v}$ is initialized. In the update step, U is computed using the current value of $\mathbf{v}$, followed by computation of $J_m(U, \mathbf{v})$, and then update $\mathbf{v}$. The computation

of the cluster centers is the weighted average of the membership values. Repeat the update step iteratively, and terminate when the matrix norm of the difference between the value of U in consecutive iterations is less than a threshold, ϵ_L , which is an infinitesimal value.

Since the introduction of FCM algorithm, it has been constantly improvised owing to its popularity as a probabilistic counterpart of the K-means clustering algorithm. One of the early improvisations was in using FCM with incomplete data, where some of the s -dimensional feature vector of the items has missing values (Hathaway and Bezdek 2001). If the amount of missing data is substantial, then one of the strategies is to compute partial distance, i.e., compute distances between only part of the vector which is complete. A second strategy for modifying FCM is in determining the optimal values for the missing values, such that the FCM functional $J_m(U, \nu)$, is minimized. A third strategy is in identifying the nearest neighbor using partial strategy, and use the values in the neighbor to fill out the missing values in the current item. Overall, the optimal completion strategy has given the best results.

FCM is sensitive to the distance metric used in the computation of U that also affect the computations of ν and J_m (Wang et al. 2004). This distance measure can be improved by using the weighted Euclidean distances where the weights are learnt using a gradient descent method. This method introduces a measure of fuzziness in the similarity between two items, which are represented using s -dimensional feature vector. If the similarity value is 0 or 1, then it is discernable, but if the similarity value is close to 0.5, then the fuzziness is increased. The fuzziness of an iteration of the FCM is the aggregate sum of the fuzziness across all elements. The fuzziness is further encapsulated in an evaluation function, which is minimized. The argument of the minimum value of the evaluation functions provides the weights for the feature-weight learning process. While the clustering outcomes have improved, the tradeoff with the original FCM algorithm is that the time complexity has increased to $O(cn^2)$.

For evaluating the clustering outcomes, cluster validity measures are required for fuzzy clusters similar to the clear well-separated clusters from hard clustering (Pakhira et al. 2004). The latter is also referred to as crisp clusters. While the Dunn's index and Davies-Bouldin index are exclusively for crisp clusters, the Xie-Beni (XB) index has been used for fuzzy clusters. The XB index is given as a ratio of the intra-cluster sum of squared distances to the minimum inter-cluster, and the ratio is normalized by the number of items. Various new validity indices for fuzzy clusters have been proposed in the 2000s and later, which improve upon the XB index. For instance, a PBM index has been proposed which does better for both elliptical and circular shaped clusters, unlike the XB index that fails for the elliptical shapes.

Applications with a Focus on Image Processing

FCM can find several uses in geosciences including clustering a certain set of objects as well as clustering within a data item with spatial context, e.g., images. Here, the applications of FCM in image processing problems are discussed. FCM itself can be improved to work with spatial datasets, such as images, where spatial contiguity must be preserved in the clustering outcome (Chuang et al. 2006). This application of clustering is on pixels, thus leading to image segmentation. For spatial FCM, the membership score is computed using weights which use the results of convolution of specific kernels in the $k \times k$ pixel neighborhood of each pixel in the image. Thus, introducing the spatial weights in the computation of the membership matrix facilitates preserving spatial contiguity in the FCM clustering outcomes. FCM clustering is especially useful in segmentation of multispectral images (Tran et al. 2005). Even with high resolution scanners, the pixels in multispectral images tend to contain the spectral response of several components. Thus, overlapping clusters naturally exist in such images. Thus, FCM is a better fit for mapping pixels to components in such an application instead of the K-means algorithm.

A similar segmentation with the use of FCM has been used for change detection in landslide mapping (Lei et al. 2018). In this application, firstly, a fast implementation of FCM provides the segmentation to candidate and non-landslide regions in the very high-resolution (VHR) remote sensing images. Then, the segmentation result is further refined using analysis of the difference image. This method has been found to be more superior than level-set and Markov chain-based methods in terms of increased accuracy, reduced parameter tuning, and faster execution. FCM can also be used in conjunction with other data science methods, such as Markov random field (MRF) for improved results of segmentation.

Future Scope

Given the popularity of the FCM clustering algorithm, there has been a constant effort in improving the accuracy and the efficiency of its implementation since the 1980s. One of the earliest attempts at efficient implementation involves the use of approximate values, which are accessed efficiently using a lookup table (Cannon et al. 1986). The lookup table helps in efficient computation of Euclidean distances. This has considerably reduced the execution time on the CPU (central processing unit). In the state of the art, the implementation has been further improved through parallel implementation on multiple processing units, such as the CPUs and GPUs.

The newer challenges in the implementation of FCM are pertaining to the large size of the datasets (Havens et al. 2012). Various strategies prevalent for handling large-scale datasets can be applied in the case of FCM clustering, such as sampling, incremental techniques, and kernelized variants of FCM. The kernel algorithms have been found to have higher time complexity in comparison to the vector-based counterparts. In general, all of these algorithms are motivated to reduce space complexity, followed by reduction in time complexity.

In this article, several aspects of FCM clustering and its geoscientific applications have been discussed. FCM clustering has been proven to be an effective data processing technique for image segmentation, typhoon tracking, change detection in SAR images, and similar other applications.

Cross-References

- [Fuzzy Set Theory in Geosciences](#)
- [K-Means Clustering](#)
- [K-nearest Neighbors](#)

References

- Bezdek JC (1973) Fuzzy-mathematics in pattern classification. PhD thesis
- Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms: advanced applications in pattern recognition. Plenum Press, New York
- Cannon RL, Dave JV, Bezdek JC (1986) Efficient implementation of the fuzzy c-means clustering algorithms. *IEEE Trans Pattern Anal Mach Intell* 2:248–255
- Chuang KS, Tzeng HL, Chen S, Wu J, Chen TJ (2006) Fuzzy c-means clustering with spatial information for image segmentation. *Comput Med Imaging Graph* 30(1):9–15
- Dunn JC (1973) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Cybern Syst* 3:32–57. <https://doi.org/10.1080/01969727308546046>
- Hathaway RJ, Bezdek JC (2001) Fuzzy c-means clustering of incomplete data. *IEEE Trans Syst Man Cybern B Cybern* 31(5):735–744
- Havens TC, Bezdek JC, Leckie C, Hall LO, Palaniswami M (2012) Fuzzy c-means algorithms for very large data. *IEEE Trans Fuzzy Syst* 20(6):1130–1146. <https://doi.org/10.1109/TFUZZ.2012.2201485>
- Lei T, Xue D, Lv Z, Li S, Zhang Y, K Nandi A (2018) Unsupervised change detection using fast fuzzy clustering for landslide mapping from very high-resolution images. *Remote Sens* 10(9):1381
- Pakhira MK, Bandyopadhyay S, Maulik U (2004) Validity index for crisp and fuzzy clusters. *Pattern Recogn* 37(3):487–501
- Tran TN, Wehrens R, Buydens LM (2005) Clustering multispectral images: a tutorial. *Chemom Intell Lab Syst* 77(1–2):3–17
- Wang X, Wang Y, Wang L (2004) Improving fuzzy c-means clustering based on feature-weight learning. *Pattern Recogn Lett* 25(10):1123–1132. <https://doi.org/10.1016/j.patrec.2004.03.008>

Fuzzy Inference Systems for Mineral Exploration

Bijal Chudasama¹, Sanchari Thakur² and Alok Porwal^{3,4}

¹Geological Survey of Finland, Espoo, Finland

²University of Trento, Trento, Italy

³Center of Studies in Resources Engineering, IIT Bombay, Mumbai, India

⁴Centre for Exploration Targeting, University of Western Australia, Crawley, WA, Australia

Definition

A Fuzzy Inference System (FIS) is a system containing a set of *if-then* rules expressed in natural language to simulate inductive reasoning of an expert (Porwal et al. 2015). It is a knowledge-driven inference engine built upon the theory of fuzzy sets. A *fuzzy set* is the extension of a classical set and does not have clearly defined limits. The degree of membership to fuzzy sets grades from 0 to 1 (unlike classical sets, where it is either 0 or 1) (Zadeh 1973). A fuzzy set, hence, allows for a simplified representation of real-world phenomena including geological processes (see “Fuzzy Set Theory in Geosciences”). When more than one fuzzy set are identified in a dataset and combined using logical operators, it forms a fuzzy logic overlay. Numerous fuzzy logic overlays expressed as *if-then* rules and integrated in an inference engine encompass a FIS. A FIS has the capabilities to capture the imprecision and vagueness of natural phenomena within a single system. Genesis of mineral deposits is a result of interplay of stochastic geological processes. This entry describes the application of FISs for modeling these processes and thereby guiding mineral exploration activities.

FIS for Mineral Exploration

The concept of FIS was formulated in the early works on fuzzy logic by Zadeh (1973). Consequent development of FISs for industrial applications is described in Mamdani and Assilian (1975). The FISs are also known as fuzzy-rule-based systems, fuzzy models, fuzzy associative memories (FAM), or fuzzy controllers when used as controllers (Jang 1993).

In the field of mineral exploration, the initial studies involved the use of fuzzy logic overlays (see “Fuzzy Set Theory in Geosciences”). However, the earliest application of FIS to mineral exploration was by Porwal et al. (2004) for mineral prospectivity modeling. Quantifying the possibility

of finding a mineral deposit in a geographic region by mathematical integration of geological variables (representing mineralization processes) is known as mineral prospectivity modeling. Mineral prospectivity modeling can be implemented using different methods ranging from data-driven regression to knowledge-driven expert systems. FISs are intuitive and advanced knowledge-driven systems that can be used for guiding mineral exploration activities at varying scales of prospectivity modeling (Porwal et al. 2015).

Architecture of a FIS

The generic architecture of a rule-based FIS is shown in Fig. 1. The components of a FIS are:

1. Fuzzy set: Let X represent the universal set of values x of a geological variable, such as elemental concentration or distance to a geological feature. A Fuzzy set $\tilde{A}$ in X is a mathematical representation of the expert knowledge of the range of values of X , which is otherwise expressed in natural language. Fuzzy sets are assigned linguistic labels such as “proximal” (e.g., close to a geological feature), “high” (e.g., anomalously large elemental concentrations), and so on. Mathematically, a fuzzy set is a set of ordered pairs, such that:

$$\tilde{A} = \{(x, \mu_{\tilde{A}}(x)) | x \in X\} \quad (1)$$

where $\mu_{\tilde{A}} \in [0,1]$ is the degree of membership of x in $\tilde{A}$.

2. Fuzzy membership function (FMF): An FMF is a mathematical function, such as linear, Gaussian, or sigmoid, that transforms the numeric values of an input variable to the degree of membership in the corresponding fuzzy set. Selection of FMF is based on its ability to represent the influence of a geological feature on modeled geological phenomenon. For instance, Fig. 2 shows a fuzzy set $\tilde{A}$, with linguistic label *Proximal*, defined on set X : *distance to a geological feature* having a degree of membership, $\mu_{\tilde{A}}$ computed using a Gaussian FMF.
3. Fuzzy operators: Fuzzy operators are logical operators that integrate the fuzzy sets in a FIS. These are the building blocks of the *if-then* rules. The three basic fuzzy operators are OR (disjunction), AND (conjunction), and NOT (complement) that perform the union, intersection, and complement of the fuzzy sets, respectively (Fig. 3). Binary t-norm and t-conorm operators can also be used. These enable the generalization of the union and intersection operators, respectively.

4. Fuzzy rules: Fuzzy *if-then* rules are conditional statements formulated to impart logical reasoning to the inference engine. They contain fuzzy sets connected by the operators. A single fuzzy *if-then* rule is in the form of:

$$\text{IF } x \text{ is } A \text{ AND } y \text{ is } B, \text{ THEN } z \text{ is } C \quad (2)$$

where A and B are fuzzy sets in the input variables x and y , respectively, and C is a fuzzy set in the output variable z . The *if* part of the statement (a.k.a. antecedent) defines the condition or the premise of the input variables and the *then* part (a.k.a. consequent) models the output conclusion or the consequence for the given input premises.

5. Defuzzifier: The defuzzifier in a FIS converts the consequent values aggregated across all rules to a crisp numeric output value. The common defuzzification methods are centroid of area, bisector of area, mean of maximum, smallest of maximum, and largest of maximum.

Types of FISs

There are two types of FISs: (1) a Mamdani-type FIS (Mamdani and Assilian 1975), and (2) a Takagi-Sugeno-type (T-S) FIS (Takagi and Sugeno 1985). In a Mamdani-type FIS (Figs. 1 and 4) the output variable in the consequent part comprises fuzzy sets defined by different output membership functions (Eq. 2). For instance, in Fig. 4 the output variable *chemical traps potential* comprises three fuzzy sets *High*, *Moderate*, and *Low*. The degree of membership to these fuzzy sets is defined by Gaussian FMFs of varying parameters. In contrast, a T-S FIS is of the format:

$$\text{IF } x \text{ is } A \text{ AND } y \text{ is } B, \text{ THEN } Z = f(x, y), \quad (3)$$

where $f(x, y)$ is a zero- or a first-degree polynomial function of the input variables.

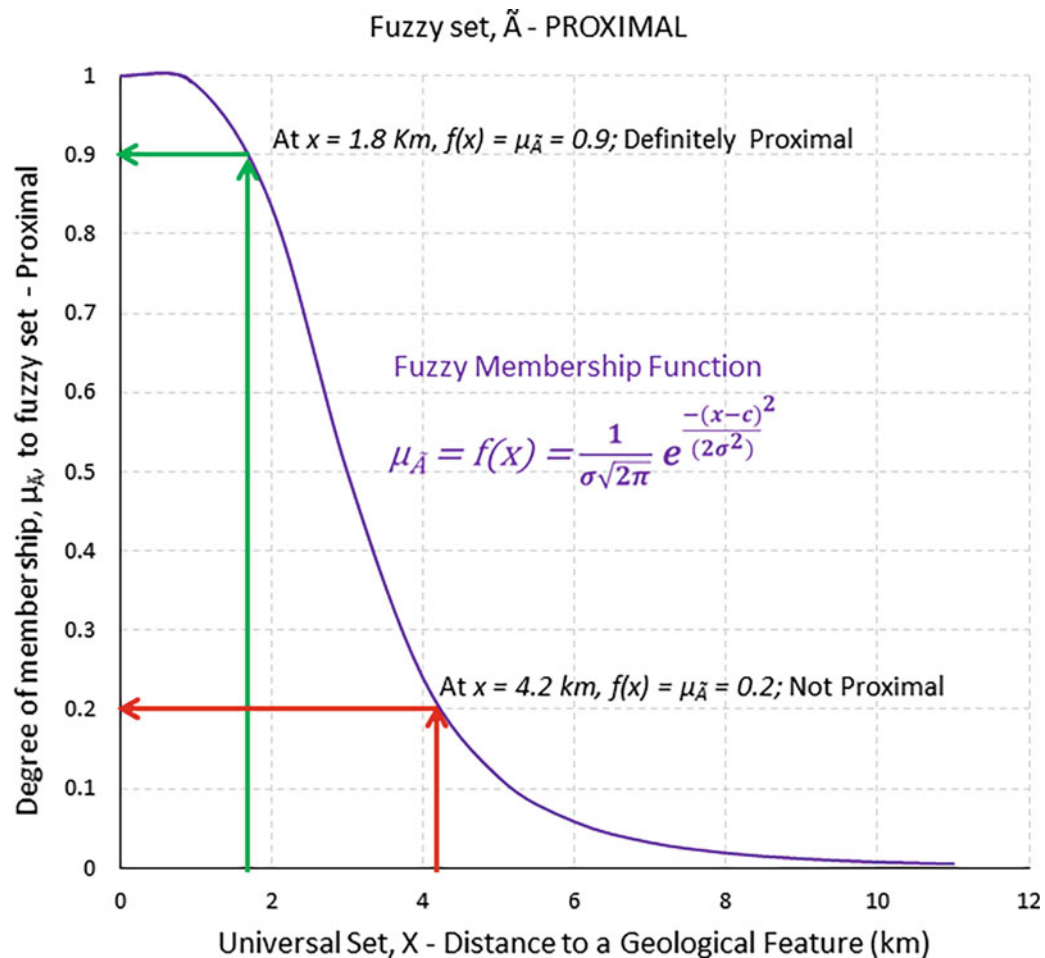
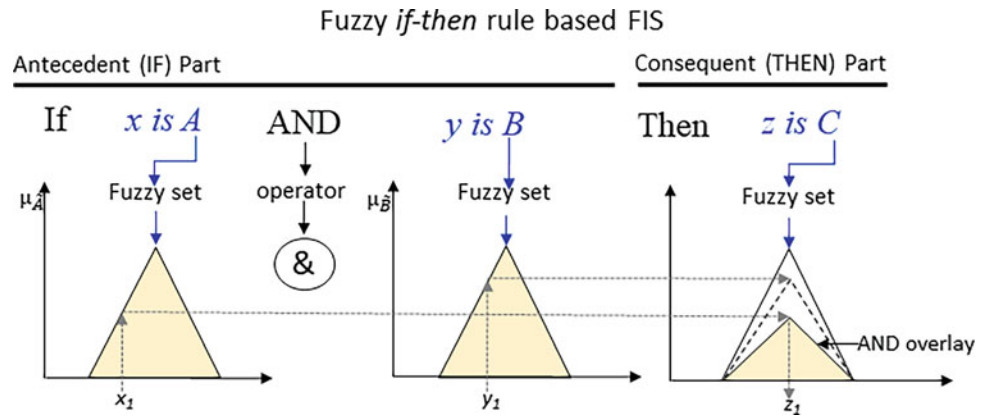
The Mamdani- and T-S-type FISs are knowledge-driven. Jang (1993) demonstrated an adaptive neuro fuzzy inference system (ANFIS), which is a hybrid type of T-S FIS. In an ANFIS the hybrid learning procedure combines the least square estimators and the gradient descent algorithm for parameterization of the FIS.

Example of a FIS Designed for Gold Mineralization

Figure 4 shows an example of a hypothetical Mamdani-type FIS (Mamdani and Assilian 1975) designed for modeling

Fuzzy Inference Systems for Mineral Exploration,

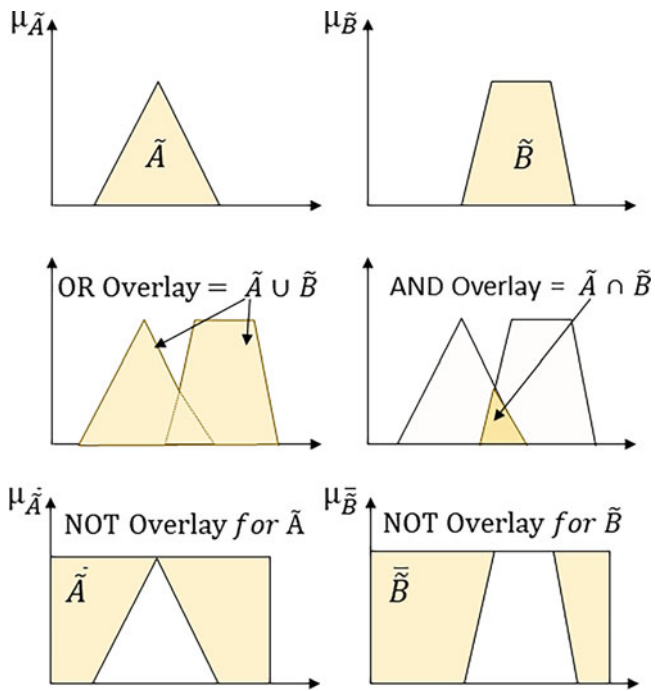
Fig. 1 Generic architecture of a rule-based FIS with one if-then rule



Fuzzy Inference Systems for Mineral Exploration, Fig. 2 Continuous fuzzy membership function for a fuzzy set 'Proximal' in a universal set "Distance to a Geological Feature"

gold mineralization based on the potential of chemical traps. It contains three input evidence layers representing X: distance to Fe-rich rocks, Y: relative conductivity of rocks, and Z: density of lithological contacts weighted by reactivity

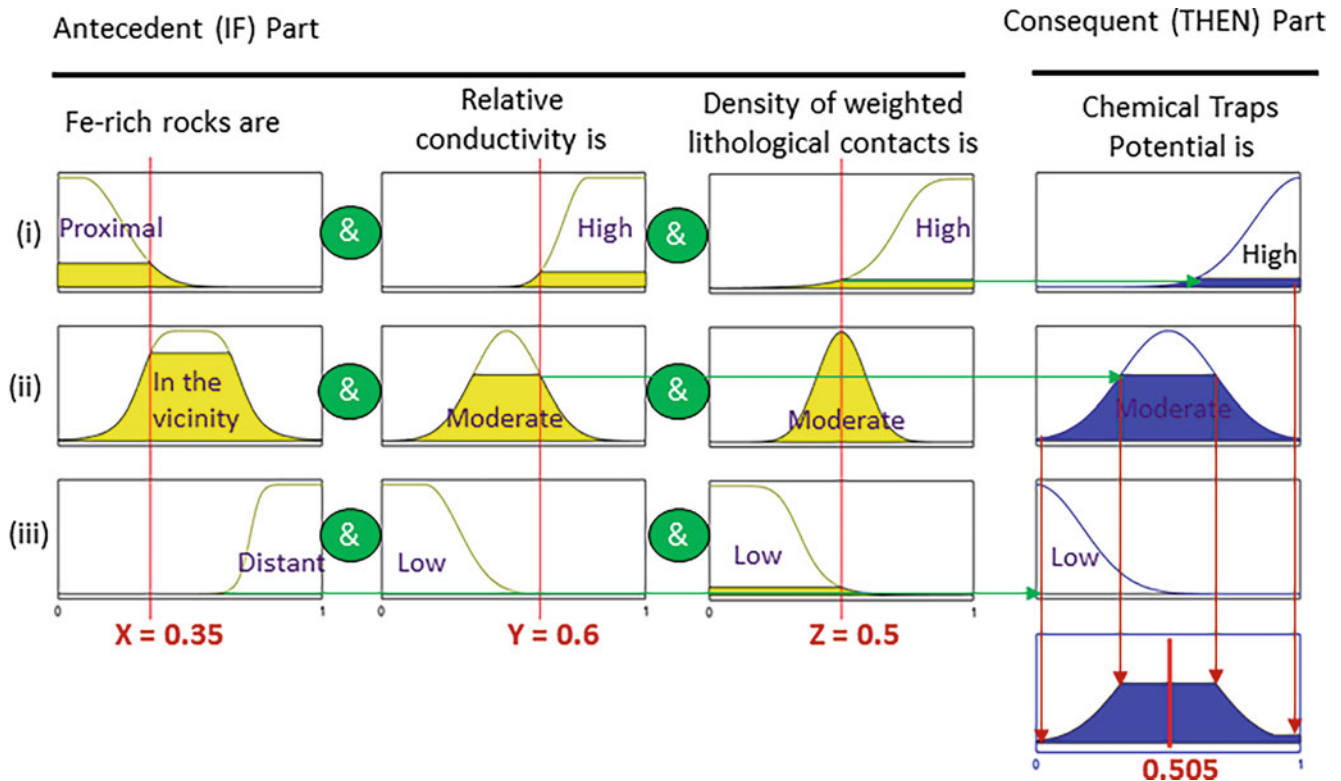
contrasts across the contacts. These input variables are combined in logical statements using the following fuzzy *if-then* rules:



Fuzzy Inference Systems for Mineral Exploration, Fig. 3 Fuzzy-overlay operators – OR, AND, and NOT

- (i) IF Fe-rich rocks are *Proximal* AND Relative conductivity is *High* AND Density of lithological contacts is *High* THEN Chemical Traps Potential is *High*.
- (ii) IF Fe-rich rocks are *In the vicinity* AND Relative conductivity is *Moderate* and Density of lithological contacts is *Moderate* THEN Chemical Traps Potential is *Moderate*.
- (iii) IF Fe-rich rocks are *Distant* AND Relative conductivity is *Low* and Density of lithological contacts is *Low* THEN Chemical Traps Potential is *Low*.

In the above rules, the linguistic values – *proximal*, *in the vicinity*, *distant*, *high*, *moderate*, and *low*, are the fuzzy sets combined using the conjunction operator, AND, for fuzzy overlay in the inference engine. For input values $X = 0.35$, $Y = 0.6$ and $Z = 0.5$, rules i and ii are activated and the chemical traps potential is mapped to the output fuzzy sets *High* and *Moderate*, respectively. Defuzzification using the centroid method translates the fuzzy linguistic values to a crisp numerical value of 0.505 for the *chemical traps potential*. Depending upon the unique feature vectors present in a given dataset, different rule(s) of the FIS can get activated. The corresponding output potential is defuzzified from the



Fuzzy Inference Systems for Mineral Exploration, Fig. 4 Three-if-then rule-based Mamdani-type FIS modeling potential of favorable chemical traps for gold mineralization

aggregate of the *consequents* of all the activated rules using a defuzzifier.

Selected Case Studies

In contrast to purely data-driven techniques, the possibility of incorporating knowledge based on geological expertise has motivated the choice of FIS for efficient modeling of mineralization processes, as demonstrated in recent studies. Porwal et al. (2015) developed and implemented a Mamdani-type FIS with 34 *if-then* rules for prospectivity modeling of surficial uranium deposits in the Yeelirrie area in Western Australia (WA). This study identified several palaeochannels in their study area as highly prospective for mineralization. The palaeochannels in the Deserts and Xeric Shrublands biome in WA are geologically and climatically important areas for formation of surficial uranium deposits. Building from the proto-FIS presented by Porwal et al. (2015), Chudasama et al. (2018) implemented a multistage FIS-based regional scale prospectivity modeling of surficial uranium mineralization in the palaeochannels of WA. The sources and traps components of the surficial uranium mineral system were modeled individually using Mamdani-type FISs. The study identified new greenfields areas for surficial uranium deposits in the less-explored Paterson orogen and Musgraves block.

Chudasama et al. (2016) presented a two-stage Mamdani-type FIS-based prospectivity modeling of orogenic gold mineralization in the Palaeoproterozoic Kumasi Basin in Ghana. Their results identified high prospectivity zones along the structural trends of known mineral deposits capturing the strong spatial association between the known deposits and the metallogenic trends in the area. Of the 32 in situ gold deposits, the model mapped 78% of these within areas of high prospectivity values. New areas identified from the results were taken up for ground exploration by the lease holding company.

Porwal et al. (2004) applied an ANFIS to model base metal mineralization potential in the Aravalli metallogenic province in western India. The Aravalli province contains significant reserves of base metal sulfide deposits, particularly lead and zinc, including the world class Rampura-Agucha deposit. This was a first study where a T-S ANFIS trained using a four-layered feed-forward adaptive neural network was applied to mineral prospectivity modeling. The results captured 96% of the known base metal deposits within 9.75% of the study area and greatly reduced the search areas for guiding further ground exploration activities.

The collective results of the above studies demonstrate that FIS models can be effectively applied (1) for modeling mineralization processes at regional-scales, (2) to aid identification of prospect-scale targets, and (3) to locate drilling areas for expansion of known resources.

Summary and Conclusions

Geological processes leading to formation of mineral deposits are complex and mathematically nondeterministic. Quantification of these processes by a single or hierarchical fuzzy logic overlay leads to an oversimplified representation of the intricate mineral systems. Hence when applied to mineral prospectivity modeling, FISs are powerful at simulating mineralization-related complex geological processes. A FIS is enhanced by the deductive reasoning of the geoscientist for modeling these processes. Being knowledge-driven they are not restricted to learning deposit-specific geological features, but instead are capable of comprehensively modeling mineralization processes. Additionally, in data-rich areas, neuro-fuzzy framework can impart learning capabilities to a T-S FIS for efficient modeling of the targeted mineral system. FISs are hence intuitive and capable of incorporating human cognition and the complex geoscientific reasoning required for mineral exploration activities.

Cross-References

- [Fuzzy Set Theory in Geosciences](#)
- [Machine Learning](#)
- [Mineral Prospectivity Analysis](#)
- [Predictive Geologic Mapping and Mineral Exploration](#)

Bibliography

- Chudasama B, Porwal A, Kreuzer OP, Butera K (2016) Geology, geodynamics and orogenic gold prospectivity modeling of the Paleoproterozoic Kumasi Basin, Ghana, West Africa. *Ore Geol Rev* 78:692–711
- Chudasama B, Kreuzer O, Thakur S, Porwal A, Buckingham A (2018) Surficial uranium mineral systems in Western Australia: geologically-permissive tracts and undiscovered endowment. *IAEA TECDOC SERIES* 446
- Jang JS (1993) ANFIS: adaptive-network-based fuzzy inference system. *IEEE Trans Syst Man Cybern* 23(3):665–685
- Mamdani EH, Assilian S (1975) An experiment in linguistic synthesis with a fuzzy logic controller: *Int. J Man-Machine Stud* 7(1):1–13
- Porwal A, Carranza EJM, Hale MA (2004) Hybrid neuro-fuzzy model for mineral potential mapping. *Math Geol* 36:803–826
- Porwal A, Das RD, Chaudhary B, Gonzalez-Alvarez I, Kreuzer O (2015) Fuzzy inference systems for prospectivity modeling of mineral systems and a case-study for prospectivity mapping of surficial uranium in Yeelirrie area, Western Australia. *Ore Geol Rev* 71:839–852
- Takagi T, Sugeno M (1985) Fuzzy identification of systems and its applications to modeling and control. *IEEE Trans Syst Man Cybern* 15(1):116–132
- Zadeh LA (1973) Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Trans Syst Man Cybern* SMC-3:28–44

Fuzzy Set Theory in Geosciences

Behnam Sadeghi^{1,2}, Alok Porwal^{3,4}, Amin Beiranvand Pour⁵ and Majid Rahimzadegan⁶

¹EarthByte Group, School of Geosciences, University of Sydney, Sydney, Australia

²Earth and Sustainability Science Research Centre, University of New South Wales, Sydney, NSW, Australia

³Centre for Studies in Resources Engineering, Indian Institute of Technology Bombay, Mumbai, India

⁴Centre for Exploration Targeting, University of Western Australia, Crawley, WA, Australia

⁵Institute of Oceanography and Environment (INOS), Universiti Malaysia Terengganu (UMT), Terengganu, Malaysia

⁶Water Resources Engineering and Management Department, Faculty of Civil Engineering, K. N. Toosi University of Technology, Tehran, Iran

Definition

Fuzzy set was introduced by Zadeh (1965- see the entry “Zadeh, Lotfi A.”) as a set whose membership ranges from 0 to 1. The label of a fuzzy set is a linguistic value. For numerical data, the conversion of each data value to fuzzy membership is through a fuzzy membership function. Formally, let X be a set whose elements are represented generically as x . A fuzzy set $\tilde{A}$ in X is a set of ordered pairs:

$$\tilde{A} = \{(x, \mu_{\tilde{A}}(x)) | x \in X\}.$$

where $\mu_{\tilde{A}}$ is the membership function (degree of compatibility or degree of truth) of x in $\tilde{A}$. The fuzzy set $\tilde{A}$ becomes a classical set if its membership value is restricted to either 0 or 1.

Introduction of Fuzzy Set Theory

The membership of a traditional binary set is confined to 1 or 0, implying that an object can either be a member of the set or not a member of the set. The binary set theory therefore disallows partial memberships, which is a major limitation particularly when dealing with poorly constrained and ill-defined phenomenon. The theory of fuzzy sets extends the membership to a value between 0 and 1, thus allowing partial membership of objects in a set. The fuzzy membership value represents the degree of truth that an object is a membership of a given set. Consider, for example, a set defined by high copper anomalies. If this set is considered a binary set, then a given sample can be a member of this set or not a member of

this set. However, it is often difficult to decide the threshold at which a given sample value becomes a member. A fuzzy set requires no such hard thresholding of values, and all sample values are considered members. The membership value defines the degree of truth that a given sample value is indeed a high anomaly.

Geospatial phenomena and processes are often vaguely defined and inherently fuzzy. The variables of geospatial processes are poorly constrained and often described in terms of linguistic values. Therefore, the fuzzy set theory has wide-ranging applications in modeling and predicting geospatial events. The most important process for this sake is assigning the proper membership values to the available data layers considering their importance in the final inference for a more efficient interpretation based on the combined data layers and evidential maps. Such maps are generated using either qualitative analysis (called “knowledge-driven approach”) or quantitative analysis (called “data-driven approach”) (Carranza 2009). In the former one, the fuzzy membership values are assigned by the expert considering the relevant knowledge and experience, but in the latter one, the fuzzy membership values are assigned based on the available exploration data. Such evidential maps are combined considering all the assigned membership values to come up with a single representative map. Then the defuzzification process is applied to the map to define the targets (e.g., mineral exploration targets). Fuzzy sets are more suitable for describing imprecise parameters.

To combine fuzzy sets, a set of operators have been proposed by Zadeh (1965) and Zimmermann (1991). For example, fuzzy AND is a minimum operator, fuzzy OR is a maximum operator, fuzzy algebraic sum and fuzzy algebraic product are types of product operators, the NOT is an exclusion operator, and fuzzy gamma (γ) combines algebraic sum and algebraic product operators (Table 1). These rules and operators allocate a complete fuzzy set to the output over the insinuation process by combination processes for decision-making.

The fuzzy AND and PRODUCT operators should be used where both datasets are necessary for ascertaining the truth of the proposition, while the OR and SUM operators should be used if one dataset is sufficient for ascertaining the truth of the proposition. Obviously, the output of the fuzzy AND operator carries higher confidence.

The fuzzy gamma operator balances the increase and decrease tendencies of the fuzzy algebraic summation and product, respectively, as below:

$$F_{\text{Combination}} = (\text{Fuzzy Algebraic Sum}^{\gamma}) \times (\text{Fuzzy Algebraic Product}^{1-\gamma})$$

For $\gamma = 0$, the output is the same as that of the fuzzy product. As the value of γ increases, the output tends towards

Fuzzy Set Theory in Geosciences, Table 1 The descriptions and equations of the fuzzy AND, fuzzy OR, fuzzy algebraic product, fuzzy algebraic sum, and fuzzy gamma operators (Carranza and Hale 2001; Nykänen et al. 2008)

Operator	Equation*	Description
Fuzzy AND	$\mu_{\text{combination}} = \text{Min}(\mu_A, \mu_B, \mu_C, \dots)$	Minimum operator, equivalent to Boolean AND (logical intersection)
Fuzzy OR	$\mu_{\text{combination}} = \text{Max}(\mu_A, \mu_B, \mu_C, \dots)$	Maximum operator, equivalent to Boolean OR (logical union)
Fuzzy algebraic product	$\mu_{\text{combination}} = \prod_{i=1}^n \mu_i$	The output results always smaller or equal to the smallest contributing fuzzy membership value
Fuzzy algebraic sum	$\mu_{\text{combination}} = 1 - \prod_{i=1}^n (1 - \mu_i)$	Complementary to the fuzzy algebraic product. The output results are always larger or equal to the largest contributing fuzzy membership value
Fuzzy gamma (γ)	$\mu_{\text{combination}} = \left(\prod_{i=1}^n \mu_i \right)^{1-\gamma} - \left(1 - \prod_{i=1}^n (1 - \mu_i) \right)^{\gamma}$	Fuzzy gamma operator is an amalgamation of the fuzzy algebraic product and the fuzzy algebraic sum; it is a parameter chosen in the range of [0,1] (Zimmermann and Zysno 1980). Cautious choice of gamma produces output values that confirm a flexible compromise between the increasive tendencies of the fuzzy algebraic sum and the decreaseive effects of the fuzzy algebraic product.

* μ_n defines a membership value for map n at a particular location ($n = A, B, \dots$). Also, μ_i is the fuzzy membership values for the i th ($i = 1, 2, \dots, n$) map to be combined



Fuzzy Set Theory in Geosciences, Fig. 1 Fuzzy logic modeling steps. (Modified from Porwal et al. 2003)

the fuzzy SUM, and for $\gamma = 1$, the output is the equal to that of the fuzzy SUM.

Stages of Fuzzy Logic Modeling

Fuzzy Logic modeling has three main steps in general that should be applied to the evidential data (Porwal et al. 2003) (Fig. 1): (I) fuzzification (encoding); (II) logical integration using inference engine and fuzzy set operations; and (III) defuzzification (decoding).

The input datasets could be categorical or numerical. All the available datasets are individually imported, then a fuzzifier converts them into fuzzy values of membership, using a membership function $\mu_A(x)$, in a continuous range between 0 (also called “false,” “non-membership,” and “incomplete”) and 1 (also called “true,” “full membership,” and “complete”), respectively. In general, $\mu_A(x)$ is defined based on a priori *knowledge* of the geological system or using training data (Porwal et al. 2003). In general, the fuzzy membership values represent the importance of an input geospatial variable in the geospatial process being modelled. The individual fuzzy sets are then combined

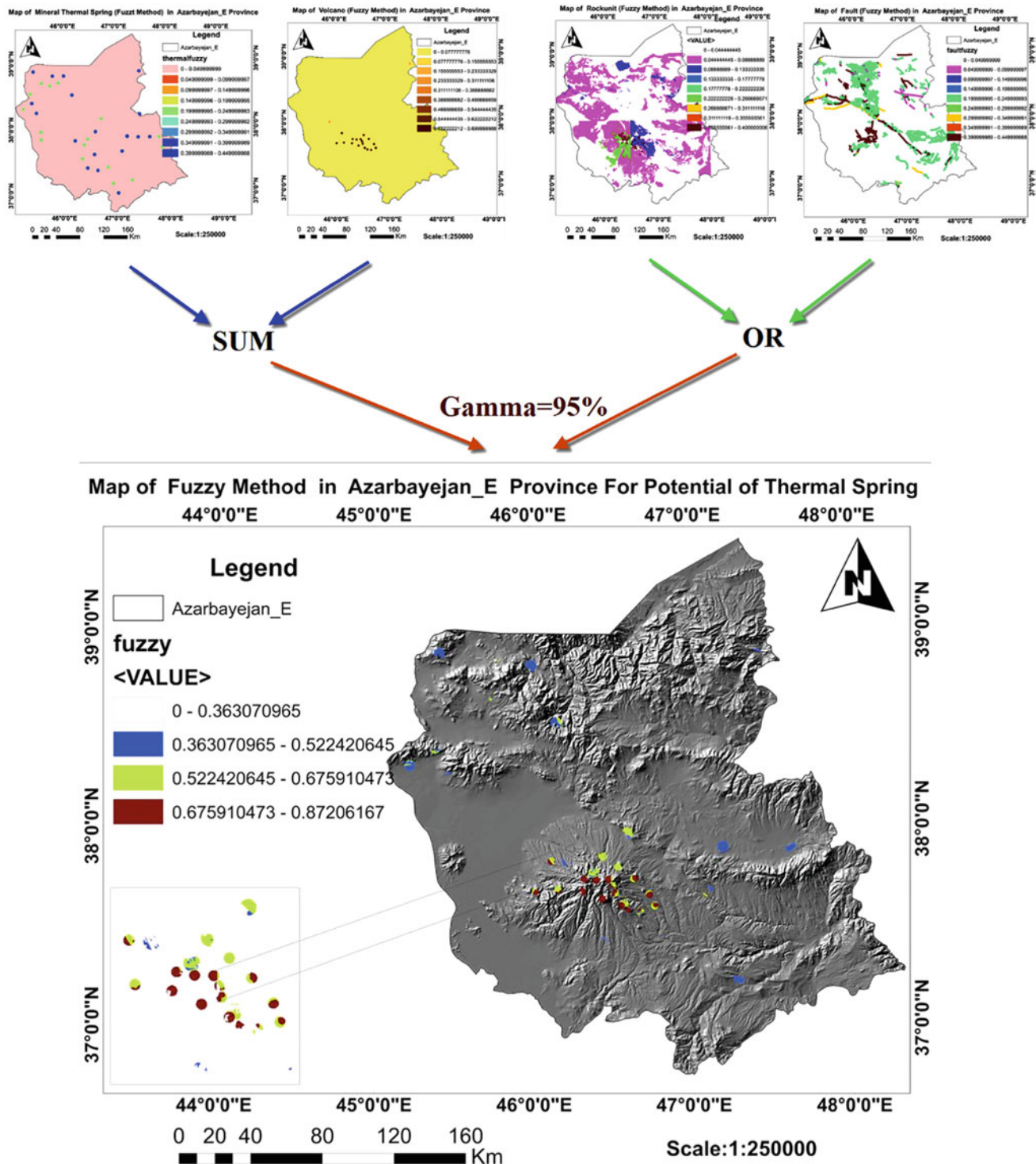
using a number of parallel and/or serial networks that sequentially combine fuzzy sets through fuzzy operators. The design of the inference engine should simulate the geospatial process being modelled (Porwal et al. 2003). The output of the inference engine is a synthesized fuzzy set of data, which is transmitted then to the defuzzifier. Here, a back-transformation process is taken place by a defuzzification mathematical function or any specific threshold fuzzy grade already defined (see Hellendoorn and Thomas 1993). In other words, defuzzification is the process of altering the combination result into a crisp output, e.g., centroid, bisector, middle of maximum (the average of the maximum value of the output set), largest of maximum, and smallest of maximum.

Further Definition Using Various Case Studies

Zimmerman (2001) is a good source to understand fuzzy set theory and its various applications. In geosciences, still Demicco and Klir (2004) with an interesting foreword by Zadeh is one of the best sourcebooks for geoscientists to gain a proper overview regarding the application of fuzzy

logic modeling in different fields of geosciences such as geology, stratigraphic modeling, hydrology and water resources, earthquake research, and sea level estimation. In the field of mineral potential mapping (MPM) using geochemical data (see the entries “Exploration Geochemistry”

and “Mineral Prospectivity Analysis”), Carranza (2009) as the other proper sourcebook has provided interesting demonstrations and details. In geostatistics and modeling spatial uncertainty Chilès and Delfiner (2012) and Caers (2011)



Fuzzy Set Theory in Geosciences, Fig. 2 Target areas detected by the fuzzy logic method (from Sadeghi and Khalajmasoumi 2015). Data layers on the top from left to right: hot springs, volcanoes, volcanic and intrusive rocks, and faults and fractures

provide clear descriptions. In the following section, several case studies are briefly reviewed.

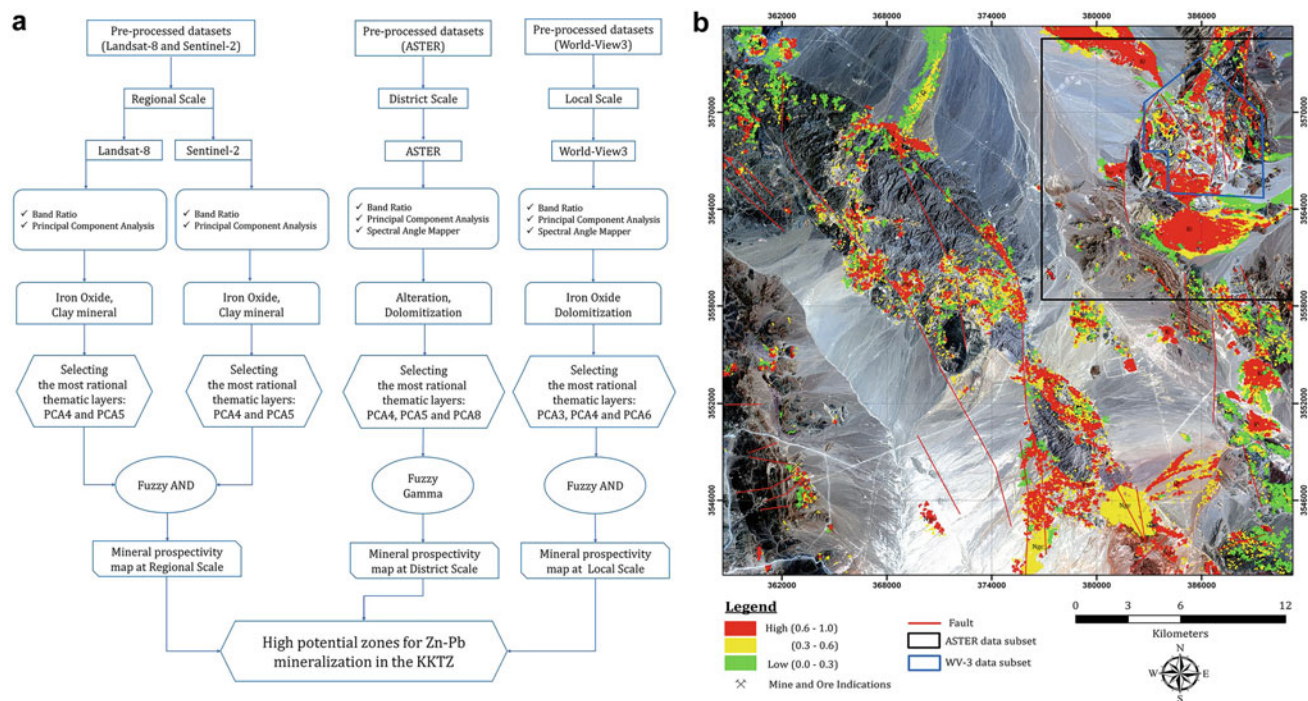
Geothermal Resource Exploration

Due to the climate change and global warming critical consequences, new sources of renewable and sustainable energy and their applications have been taken into serious consideration. Among such sources, the one highly related to geosciences is geothermal (see the entry “Geothermal Energy”). For example, Iceland, Japan, and New Zealand are among the most advanced countries around the world

that have made a great progress in using geothermal energy for a variety of applications such as electricity generation.

From geological point of view, geothermal resources could be discovered in a variety of geological systems such as volcanic rocks, limestones, and shales (Kiavarz Moghaddam et al. 2014). Acidic volcanic and intrusive rocks are considered as the main practical sources of geothermal energy for the industry (Manzella 2007; Huenges 2010).

Geothermal energy is found in either liquid-dominated (over 200°C) or vapor-dominated (with 240°C to 300°C) forms (Sadeghi and Khalajmasoumi 2015), mostly near young volcanoes around the Pacific Ocean, rifts and hot spots, faults, and geothermal alterations (Dickson and Fanelli 2004). Discharge systems such as mud volcanoes and geysers



Fuzzy Set Theory in Geosciences, Fig. 3 (a) The methodology outline applied in this study; (b) mineral potential map of the KKTZ region in Central Iran generated from Landsat-8, Sentinel-2, ASTER, and WV-3 data. (Modified from Sekandari et al. 2020)

Fuzzy Set Theory in Geosciences, Table 2 Fuzzification parameters for the input layers

Data origin	Input layer	Detection	Membership type	Fuzzy operator
Landsat-8 dataset	PCA4	OH-minerals and carbonates	Linear	AND
	PCA5	Iron oxide		
Sentinel-2 dataset	PCA4	OH-minerals and carbonates	Linear	Gamma ($\gamma = 0.6$)
	PCA5	Iron oxide		
ASTER dataset	PCA4	Iron oxide/hydroxides minerals	Linear	AND
	PCA5	OH/S-O/CO ₃ -bearing minerals		
	PCA8	Dolomite		
World-View3 dataset	PCA3	Iron-stained alteration	Linear	AND
	PCA4	Dolomite/Fe ²⁺ oxides		
	PCA6	Fe ³⁺ oxides		

are the other superficial indicators of geothermal systems. They mostly come with hydrothermal alterations.

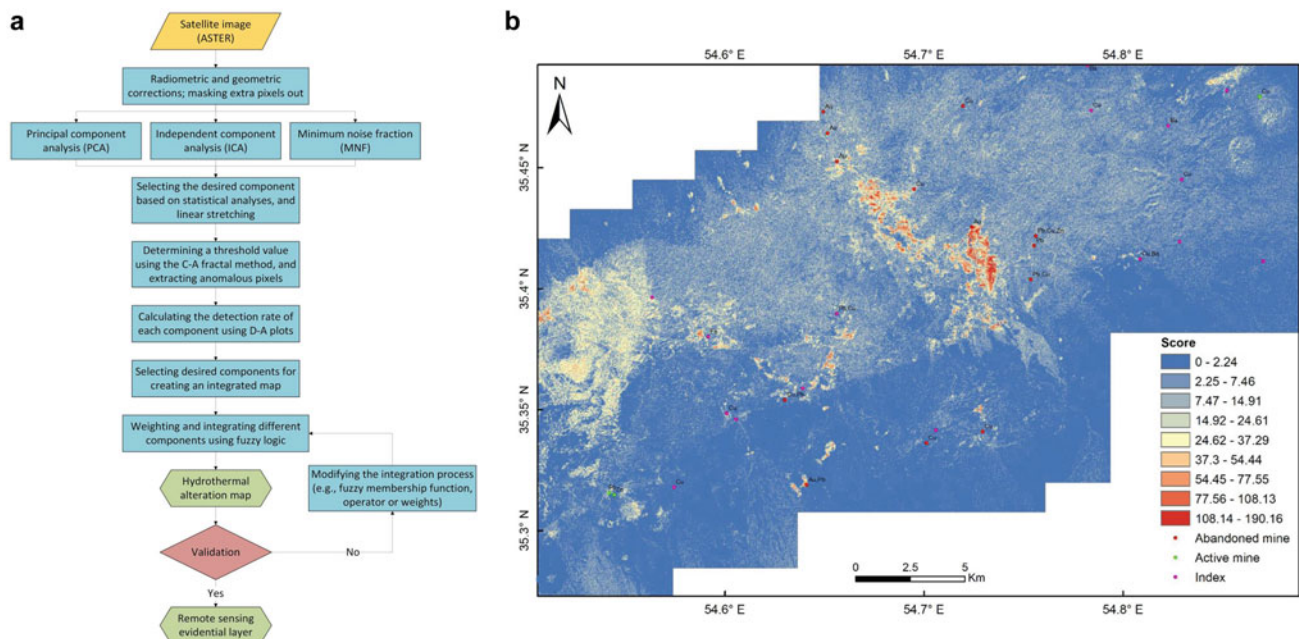
In geosciences there are two main types of uncertainties including (i) stochastic uncertainty and (ii) systematic uncertainty (Porwal et al. 2003; Sadeghi 2020 – see the entry “Uncertainty Quantification”). Simply, fuzziness is the reason of such vagueness and uncertainties. Geothermal system exploration always comes with a high relevant uncertainties and financial risk. Therefore, making the best decision considering all the available data layers could be highly helpful in reducing such uncertainties and the consequent risks.

As a case study, we refer to Sadeghi and Khalajmasoumi (2015), who focused their study on Eastern Azarbayegan Province in NW Iran. In their study, the main data layers taken into account were (1) volcanoes, (2) faults and fractures, (3) hot springs, and (4) volcanic and intrusive rocks. In this case study, based on the expert’s knowledge, SUM was applied to the layers of hot springs and volcanoes, and OR was applied to the layers of faults and volcanic-intrusive rocks. Also, 95% was assigned to γ . In the final processed map (Fig. 2), it turned out that the central part of the study area, including the blue and purple representative areas, is the best area to focus the follow-up exploration and drilling projects on.

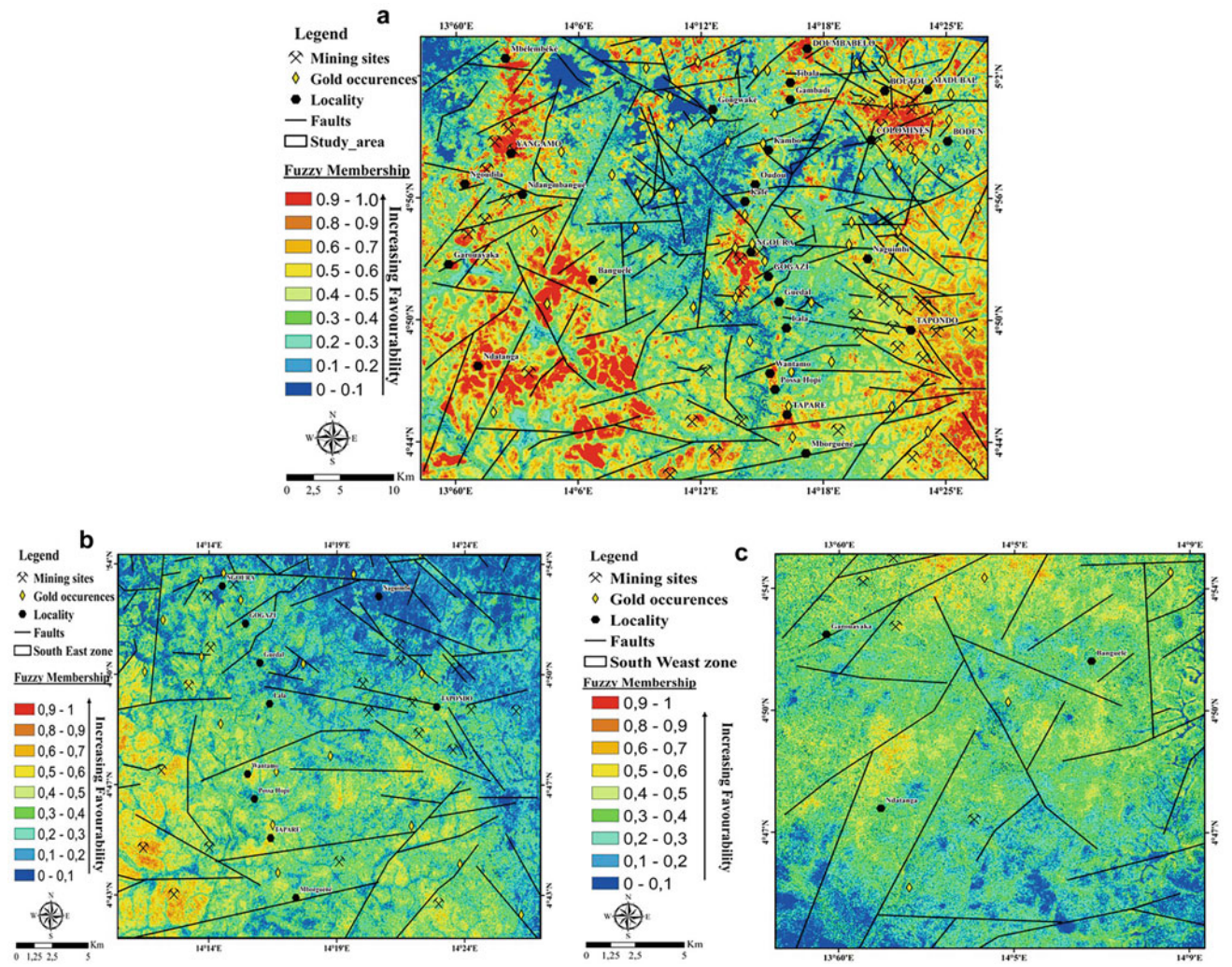
Exploration Geology Using Remote Sensing Data

Remote sensing data are efficaciously used for alteration and lithological and structural mapping (El-Wahed et al. 2021;

Pour et al. 2021). Fuzzy logic modeling (FLM) is effectively implemented to remote sensing data to integrate thematic layers extracted from the data for MPM in a variety of metallogenic provinces around the world (cf. Porwal et al. 2003; Tangestani and Moore 2003; Ranjbar and Honarmand 2004; Rahimzadegan et al. 2015; Sekandari et al. 2020; Wambo et al. 2020; Shirmard et al. 2020; Moradpour et al. 2021). For example, Sekandari et al. (2020) used FLM to integrate the thematic layers derived from Landsat-8, Sentinel-2, Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), and WorldView-3(WV-3) satellite remote sensing data for generating mineral potential maps of Carbonate-Hosted Pb-Zn Deposits in the Kashmar-Kerman Tectonic Zone (KKTZ), Central Iran. Band ratios and principal component analysis (PCA) image processing techniques were applied for mapping alteration minerals and lithological units. Then, the thematic layers of the alteration zones were integrated using FLM to produce mineral potential maps. An outline of the methodological flowchart applied in this study is demonstrated in Fig. 3a. In this study, the multiclass evidential layers were reclassified to 10 classes of equal intervals and then were fuzzified using the linear membership function. The fuzzy AND operator was applied to map the prospective areas by Landsat-8 and Sentinel-2 and WV-3 alteration thematic layers. The fuzzy gamma operator was implemented to ASTER alteration thematic layers. Table 2 shows fuzzification parameters for the input layers. Several high potential zones were identified in KKTZ. Figure 3b shows the KKTZ mineral potential map produced using Landsat-8, Sentinel-2, ASTER, and WV-3 data.



Fuzzy Set Theory in Geosciences, Fig. 4 (a) The methodology outline applied in this study. (b) Mineral potential map of the Toroud-Chahshir range, Central Iran, generated using ASTER data. (Modified from Shirmard et al. 2020)



Fuzzy Set Theory in Geosciences, Fig. 5 (a) Regional scale potential map of Ngoura-Colomines goldfield, Eastern Cameroon, generated based on Landsat-8 data; (b) mineral potential map of the southwestern zone of Ngoura-Colomines goldfield generated using ASTER data; and (c) mineral potential map of the southeastern zone of Ngoura-Colomines goldfield generated based on ASTER data. (Modified from Wambo et al. 2020)

Fuzzy Set Theory in Geosciences, Table 3 Fuzzification parameters for the input layers

Data origin		Data	Selected input layer	Membership type	Fuzzy operator		
	Field laboratory	Geochemical data	Cu, Zn, and Au	Linear	AND		Gamma ($\gamma = 0.9$)
	ASTER dataset	Structure map	Intersection of NE–SW and NW–SE lineaments		OR		
			high lines density zone				
		MF fraction images of n-D class #1 and n-D class #5	Al-OH and Fe, Mg-O-H minerals and jarosite		OR	AND	
		IC4 Image	Sericitically argillically altered minerals in association with jarosite and chlorite/epidote		OR		
	Geological map	Weighted lithology	Lithologies associated with gold mineralization (weighted 3– 5)	–			

Additionally, Shirmard et al. (2020) utilized the FLM to integrate the thematic layers derived from PCA, independent component analysis (ICA) and minimum noise fraction (MNF) techniques applied to ASTER data for generating evidential layers for MPM in the Toroud-Chahshirin range, Central Iran. Numerous prospective zones were discovered using the fused hydrothermal alteration map. The flowchart of methodology used in this study is shown in Fig. 4a. Figure 4b shows the potential map of Toroud-Chahshirin region.

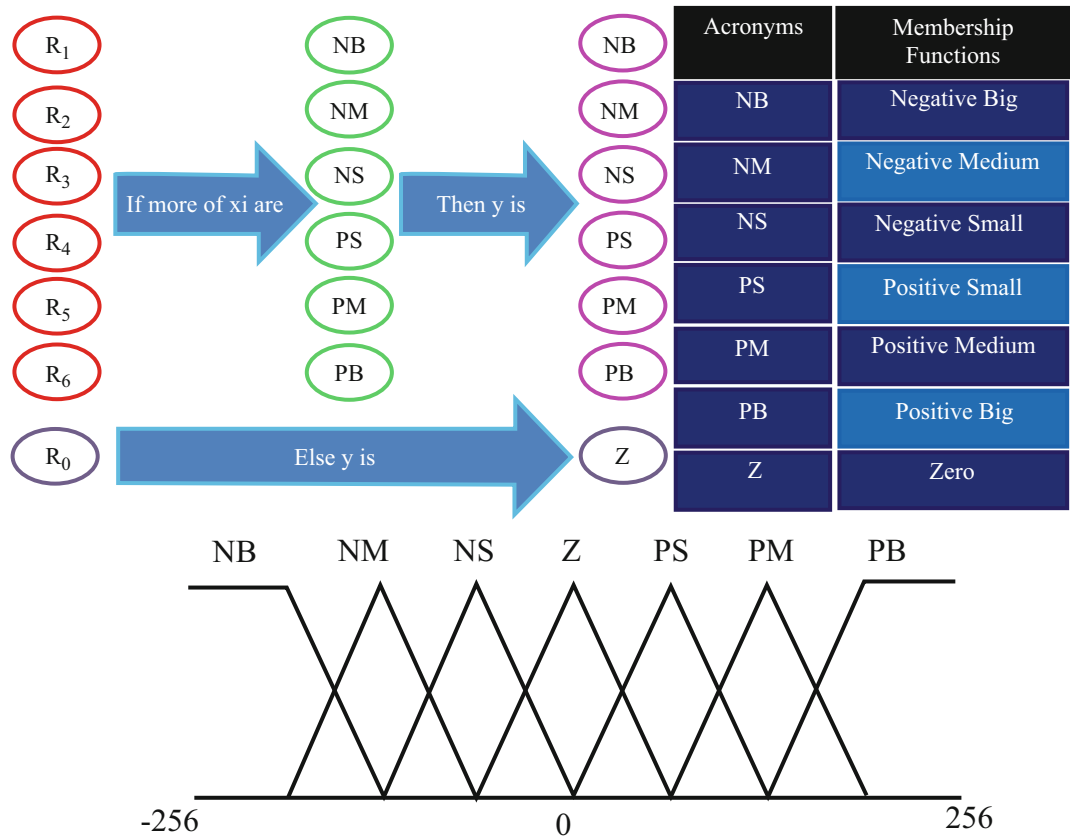
The potential mineral maps produced from thematic layers of alteration minerals using FLM can be fused with other geological maps such as structural and lithological maps and geochemical and geophysical anomaly maps to generate metallogenic provinces' mineral potential models. For example, Wambo et al. (2020) implemented the FLM to alteration thematic layers derived from Landsat-8 and ASTER remote sensing data to identify high potential zones of gold mineralization in the Ngoura-Colomines goldfield, Eastern Cameroon. The resultant mineral potential maps were fused with structural maps of the study area (Fig. 5). The high favorability index zones for gold mineralization were found in the alteration zones corresponded to the fault intersections.

In another case study for gold prospecting in the Astaneh granite intrusive, West Central Iran, MPM was produced by

fusion of ASTER alteration maps, geochemical anomaly maps, and structural and lithology maps using the FLM (Moradpour et al. 2021). Table 3 shows the fuzzification parameters for the input layers applied in this analysis. The MPM indicated several high prospective zones in the study area. Most of the potential zones are situated in the central and southeastern parts of the study area. Such zones are associated with granitic intrusive and contact metamorphic rocks.

Application of Iterative Fuzzy Edge Detection Algorithm to Blurred Satellite Images

Several natural phenomena, including atmospheric effects, aerosols, fog, and clouds and their shadows, may reduce the contrast of satellite imagery. Therefore, edge detection algorithms are applied extract the edge of such degradation effects undesirably. Application of such algorithms is complicated, and using methods based on fuzzy logic can be helpful. In this regard, Rahimzadegan and Sadeghi (2016) applied an iterative fuzzy edge detection (IFED) algorithm to blurred remote sensing images. The IFED method was firstly introduced by Farbiz et al. (1998), which is implemented in two steps, including image enhancement and edge detection. The



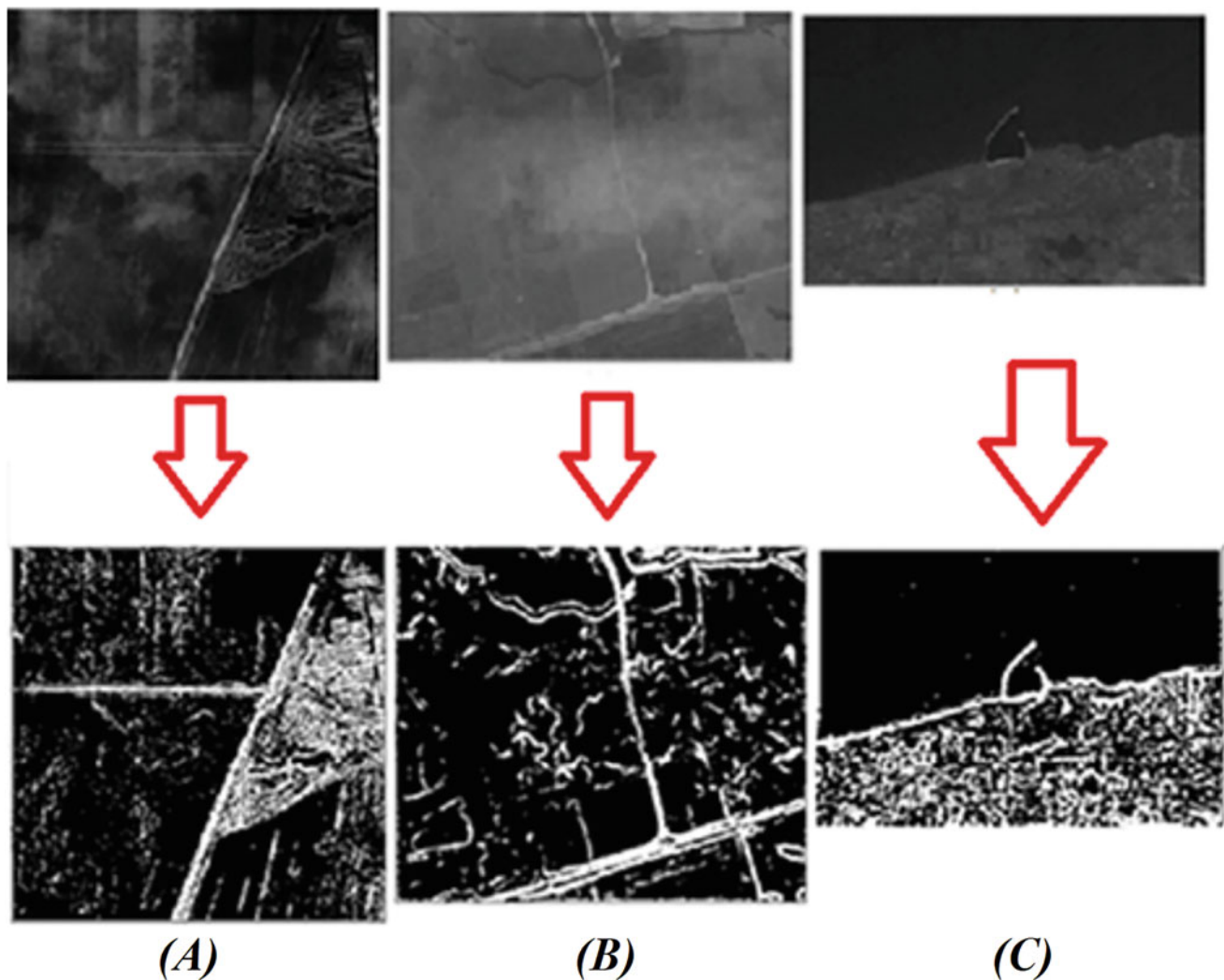
Fuzzy Set Theory in Geosciences, Fig. 6 The used (a) fuzzy rules and membership functions and (b) membership functions based on the fuzzy logic. (Modified from Farbiz et al. 1998 and Rahimzadegan and Sadeghi 2016)

IFED algorithm applies an iterative process based on the fuzzy if-then-else method, initially introduced by Russo and Ramponi (1994) (Fig. 6). The steps of the proposed IFED algorithm are introduced as follows.

Initially the gray values (x_i) are mapped to the range of -256 to 256 . A $N \times N$ window is utilized to calculate the difference between x_i and its neighboring pixels. At the next step, membership functions are calculated based on Fig. 6, and the activity value of the membership functions are acquired. Then, y , the value to add to the pixel at the center of the window to acquire the enhanced gray level is calculated using the correlation-product inference mechanism. The enhanced gray values are computed by adding y to the gray value of the pixel at the center of the selected window. An iteration number with the least edge gray-value rate (EGR) is selected as the best number of the iterations to enhance the selected image. The EGR is calculated using the ratio of the

sum of the gray values in the 10% of higher and lower gray levels of the image. Afterwards, the edges of the regarding image are detected using “min” operator (Russo and Ramponi 1996). In the end, the produced gray value fuzzy edge image is binarized using “mean,” and then a crisp edge image is produced. To evaluate the IFED algorithm, the peak signal-to-noise ratio (PSNR) is utilized.

Three satellite sub-images of the Ikonos, Landsat 7, and SPOT 5 blurred by lowering atmospheric effects were studied in this research. A selected Ikonos sub-image (1 m spatial resolution) covered by a cloud was selected for the first test (Fig. 7a). The edges of the existing roads as the desired edges were extracted using IFED algorithm, while the edges of the clouds were ignored (Fig. 7a). Hence, due to the application of the fuzzy logic, IFED is able to detect the edges of the regarding features of land and ignore atmospheric contamination.



Fuzzy Set Theory in Geosciences, Fig. 7 Original images and results of the IFED algorithm for the sub-images of (a) Ikonos, (b) SPOT 5, and (c) Landsat 7. (Modified from Rahimzadegan and Sadeghi 2016)

A SPOT 5 sub-image (spatial resolution of 2.5 m) was selected as the second case study (Fig. 7b). The clarity of the selected sub-image was lowered by cloud and its shadow. The IFED algorithm can refuse detecting the edges of the cloud and the regrading shadow and extract the edges of the roads, even those are covered by the thin cloud (Fig. 7b).

A sub-image of Landsat (approximately 15 m spatial resolution) acquired over the sea, and its beach was used for the third investigation (Fig. 7c). As sea currents produce pixels with gray values different from the pixels of the sea, which may be detected by edge detection algorithms undesirably, the performance of the IFED algorithm can be evaluated in these areas. As demonstrated in Fig. 7C, the IFED algorithm ignores the undesired edges in the image, including the edges of the currents of the sea, and the desired edges such as the land features are detected clearly. Hence, IFED represented a good performance for the regions of the beaches and ignored undesired edges.

In general, the IFED method represents a significant performance to detect the desired edges of the blurred satellite images, including contaminations of natural phenomena such as sea currents, clouds and their shadows, land features in coastal waters, and fogs and smokes. Moreover, the IFED method can extract the edges of the desired land features, including farms, shorelines, roads, and buildings, efficiently. Such methods developed based on the fuzzy logic are appropriate to enhance blurred satellite images and extract the considered edges, which may include numerous uncertainty sources.

Summary or Conclusions

Given that most geospatial processes are vague and ill-defined, and the variables of the processes are poorly constrained, the fuzzy set theory provides a robust mathematical framework for modeling and predicting these processes. The functionality of all the three components, namely, the fuzzifier, the inference engine, and the defuzzifier, is equally important for the successful implementation of a fuzzy model. The fuzzifier, that is, the fuzzy membership function should simulate the human cognition for the efficient fuzzification of the input geospatial variables. It should be tuned to account for stochastic and systemic uncertainties in the geospatial datasets. The design of the inference engine should capture the subtle relationships between the variables of the geospatial process being modeled. The defuzzifier should return a crisp output that can be used for geospatial decision-making.

Cross-References

- [Exploration Geochemistry](#)
- [Geothermal Energy](#)
- [Mineral Prospectivity Analysis](#)
- [Uncertainty Quantification](#)
- [Zadeh, Lotfi A.](#)

Bibliography

- Caers JK (2011) Modeling uncertainty in earth sciences. Wiley, Hoboken
- Carranza EJM (2009) Geochemical anomaly and mineral prospectivity mapping in GIS. Handbook of exploration and environmental geochemistry, vol 11. Elsevier, Amsterdam
- Carranza EJM, Hale M (2001) Geologically-constrained fuzzy mapping of gold mineralization potential, Baguio district, Philippines. *Nat Res Res* 10(2):125–136
- Chilès JP, Delfiner P (2012) Geostatistics: modeling spatial uncertainty, 218, 2nd edn. Wiley, Hoboken
- Demiccio R, Klir G (2004) Fuzzy logic in geology. Elsevier, Amsterdam
- Dickson MH, Fanelli M (2004) What is geothermal energy? Istituto di Geoscienze e Georisorse, CNR, Pisa
- El-Wahed MA, Zoheir B, Pour AB, Kamh S (2021) Shear-related gold ores in the Wadi Hodein Shear Belt, South Eastern Desert of Egypt: analysis of remote sensing, field and structural data. *Fortschr Mineral* 11(474):10.3390/min11050474
- Farbiz F, Motamedi SA, Menhaj M (1998) An iterative method for image enhancement based on fuzzy logic. Proceedings of the 1998 IEEE international conference on acoustics, speech and signal processing
- Huenges E (2010) Geothermal energy systems. Wiley-VCH, Federal Republic of Germany
- Hellendoorn H, Thomas C (1993) Defuzzification in fuzzy controllers. *Jour Intell Fuzzy Syst* 1:109–123
- Kiavarz Moghaddam M, Samadzadegan F, Noorollahi Y, Ali Sharifi M, Itoi R (2014) Spatial analysis and multi-criteria decision making for regional-scale geothermal favorability map. *Geothermics* 50: 189–201
- Manzella A (2007) Geophysical methods in geothermal exploration. Italian National Research Council, International Institute for Geothermal Research, Pisa
- Moradpour H, Paydar GR, Feizizadeh B, Blaschke T, Pour AB, Khalil Kamran KV, Muslim AM, Hossain MS (2021) Fusion of ASTER satellite imagery, geochemical and geology data for gold prospecting in the Astaneh granite intrusive, west Central Iran. *Int J Image Data Fusion*. <https://doi.org/10.1080/19479832.1915395>
- Porwal A, Carranza EJM, Hale M (2003) Knowledge-driven and data-driven fuzzy models for predictive mineral potential mapping. *Nat Resour Res* 12(1):1–24
- Pour AB, Zoheir B, Pradhan B, Hashim M (2021) Editorial for the special issue: multispectral and hyperspectral remote sensing data for mineral exploration and environmental monitoring of mined areas. *Remote Sens* 13(519). <https://doi.org/10.3390/rs13030519>
- Rahimzadegan M, Sadeghi B (2016) Development of the iterative edge detection method applied on blurred satellite images: state of the art. *J Appl Remote Sens* 10(3):035018. <https://doi.org/10.1117/1.JRS.10.035018>
- Rahimzadegan M, Sadeghi B, Masoumi M, Taghizadeh Ghalehjoghi S (2015) Application of target detection algorithms to identification of iron oxides using ASTER images: a case study in the north of Semnan province, Iran Arabian. *J Geosci* 8(9):7321–7331
- Ranjbar H, Honarmand M (2004) Integration and analysis of airborne geophysical and ETM+ data for exploration of porphyry type

- deposits in the central Iranian Volcanic Belt using fuzzy classification. *Int J Remote Sens* 25(21):4729–4741
- Russo F, Ramponi G (1994) Edge extraction by FIRE operators. *IEEE world congress on computational intelligence. Proceedings of the third IEEE conference on fuzzy systems*, IEEE
- Russo F, Ramponi G (1996) A fuzzy filter for images corrupted by impulse noise. *Signal Proc Lett IEEE* 3(6):168–170
- Sadeghi B (2020) Quantification of uncertainty in geochemical anomalies in mineral exploration. PhD thesis, University of New South Wales
- Sadeghi B, Khalajmasoumi M (2015) A futuristic review for evaluation of geothermal potentials using fuzzy logic and binary index overlay in GIS environment. *Renew Sust Energ Rev* 43C:818–831
- Sekandari M, Masoumi I, Pour AB, Muslim AM, Rahmani O, Hashim M, Zoheir B, Pradhan B, Misra A, Aminpour SM (2020) Application of Landsat-8, Sentinel-2, ASTER and WorldView-3 spectral imagery for exploration of carbonate-hosted Pb-Zn deposits in the central Iranian terrane (CIT). *Remote Sens* 12(1239):10.3390/rs12081239
- Shirmard H, Farahbakhsh E, Pour AB, Muslim AM, Müller RD, Chandra R (2020) Integration of selective dimensionality reduction techniques for mineral exploration using ASTER satellite data. *Remote Sens* 12(1261):10.3390/rs12081261
- Tangestani MH, Moore F (2003) Mapping porphyry copper potential with a fuzzy model, northern Shahr-e-Babak. *Iran Aust J Earth Sci* 50(3):311–317
- Wambo TJD, Pour AB, Ganno S, Asimow PD, Zoheir B, Reis Salles RD, Nzenti JP, Pradhan B, Muslim AM (2020) Identifying high potential zones of gold mineralization in a sub-tropical region using Landsat-8 and ASTER remote sensing data: a case study of the Ngoura-Colomines goldfield, eastern Cameroon. *Ore Geol Rev* 122(103530):10.1016/j.oregeorev.2020.103530
- Zadeh LA (1965) Fuzzy sets. *Inf Control* 8:338–353
- Zimmermann HJ (1991) *Fuzzy set theory – and its applications*, 2nd edn. Kluwer Acad. Publ, Dordrecht
- Zimmermann HJ, Zysno P (1980) Latent connectives in human decision making. *Fuzzy Sets Syst* 4(1):37–51

Genetic Algorithms

D. Arun Kumar

Department of Electronics and Communication Engineering and Center for Research and Innovation, KSRM College of Engineering, Kadapa, Andhra Pradesh, India

Definition

Genetic algorithms (GAs) are techniques with randomized solution search approach that obtain optimum solution using principles and parallelism of evolution and natural genetics. GAs obtain the optimal solution for complex problems using the fitness function. The implementation of GAs involves important steps, namely, a selection of the population, evaluation of fitness, defining crossover, mutation, and convergence to the optimal solution.

Introduction

The advancement of sensor technology has generated a large amount of data. There exist various types of data analysis methods such as data classification, data clustering, and regression (Duda et al. 2000). In data classification, the input data sample is classified into a specific class based on its feature values. The input sample is assigned to a specific class (Theodoridis and Koutroumbas 1999). In data clustering, the input data samples are grouped based on their feature values. In regression, the continuous values of output variable are predicted using linear or non-linear models (Duda et al. 2000).

There exist various mathematical models such as random forests, decision trees, and support vector machines (SVMs) that are used to analyze the data. In this direction, GAs are widely used in data analysis methods. GAs are used in optimization problems such as data classification, clustering, and

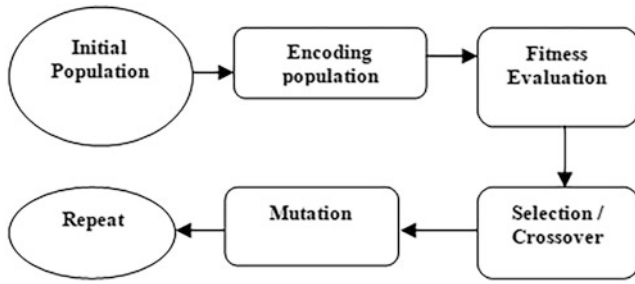
regression analysis. The advantage of GAs, such as the ability to find optimal solutions in complex, large, and multimodal search space, is the reason for their utilization in data analysis.

The implementation of GAs for data analysis is inspired by the natural genetics and evolution processes. The implementation steps of GAs are given in Fig. 1.

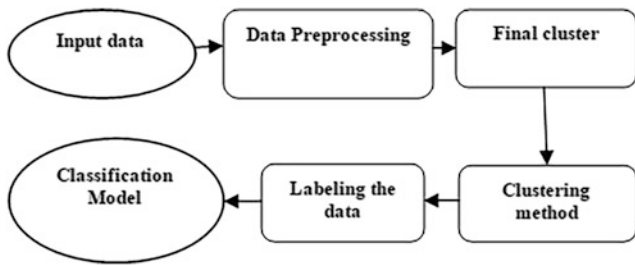
In GAs, the initial population is selected on a random basis through which the final population (representing a solution to the problem) is obtained. The individual element of population is named as chromosome. In the later stage, the chromosomes are selected based on the fitness value and roulette wheel selection-based approach (Maulik and Bandyopadhyay 2000). The selected chromosomes are encoded using corresponding encoding techniques such as binary encoding (Xue et al. 2010). The binary string with p bits has p crossover points. The crossover points for the bit strings are selected randomly and the bit strings are mutated at the crossover points such that the overall fitness value is maximum (Yang 2007). The procedure is repeated for I iterations such that the maximum fitness value is obtained by convergence to an optimal solution in N -dimensional feature space (Forrest 1996). The overall fitness computation is implemented through the process of selection, crossover, and mutation for a maximum number of iterations. The best chromosome that lasts till the end is the solution to the clustering problem. The generations with best strings are preserved to produce maximum fitness function value. In this study, the application of genetic algorithm for data clustering is presented with an example.

GAs for Clustering

Clustering of samples in N -dimensional feature space is implemented as given in Fig. 2. In Fig. 2, the input data is pre-processed to overcome the problems like missing values and redundant features. In the second stage, clustering algorithm is selected to obtain the clusters in the data. The



Genetic Algorithms, Fig. 1 Implementation steps of GAs. GAs start with a selection of the initial population, encoding each chromosome of population into binary format, evaluating the fitness value of each chromosome, selection of crossover, mutating the chromosomes to obtain maximum fitness value, and repeating the steps until the best solution is obtained



Genetic Algorithms, Fig. 2 Methodology of implementing clustering technique. Collection of data, preprocessing of data, selecting clustering method, obtaining the initial clusters, evaluating the method, labeling the patterns, and training and testing the classification model

resultant samples in each cluster are labeled by the domain-specific expert knowledge. Let N be the size of dataset with each sample P having n features. The pattern P_i with n features is given as

$$P_i = [f_1 \ f_2 \ \cdots \ f_n].$$

The clustering problem is defined as grouping N patterns in the dataset to k clusters such that the intercluster distances are minimum. The intercluster distance of S patterns of cluster (P) from the cluster center (C_p) is

$$D_p = \frac{1}{s} \sum_{i=1}^s (C_p - P_i)^2 \quad (1)$$

The summation of k intercluster distances is

$$D = \sum_{i=1}^k D_i \quad (2)$$

The clustering algorithm with minimum D value is considered the best clustering algorithm to group the pixels in the dataset.

GAs provide better cluster centers, unlike the K-means, where the initial clusters are selected on a random basis. This is because GAs avoid converging to local minima unlike the K-means algorithm. GAs provide the optimal cluster centers when the number of training iterations reaches almost nearby infinity. The steps involved in implementing GAs for clustering are given in Fig. 1. In the first stage, the initial population is selected randomly such that each element of the population forms one cluster center. The population in GAs is equal to the number of clusters, i.e., if the number of clusters is equal to k , then the population is k (with k chromosomes in the population). The number of elements in each chromosome is equal to the number of features in the pattern (Nguyen 2015). Each chromosome in the population is encoded using various encoding techniques. In the next stage, the fitness value of each chromosome is computed based on the fitness function. The overall fitness function for clustering is.

$$F = 1/D. \quad (3)$$

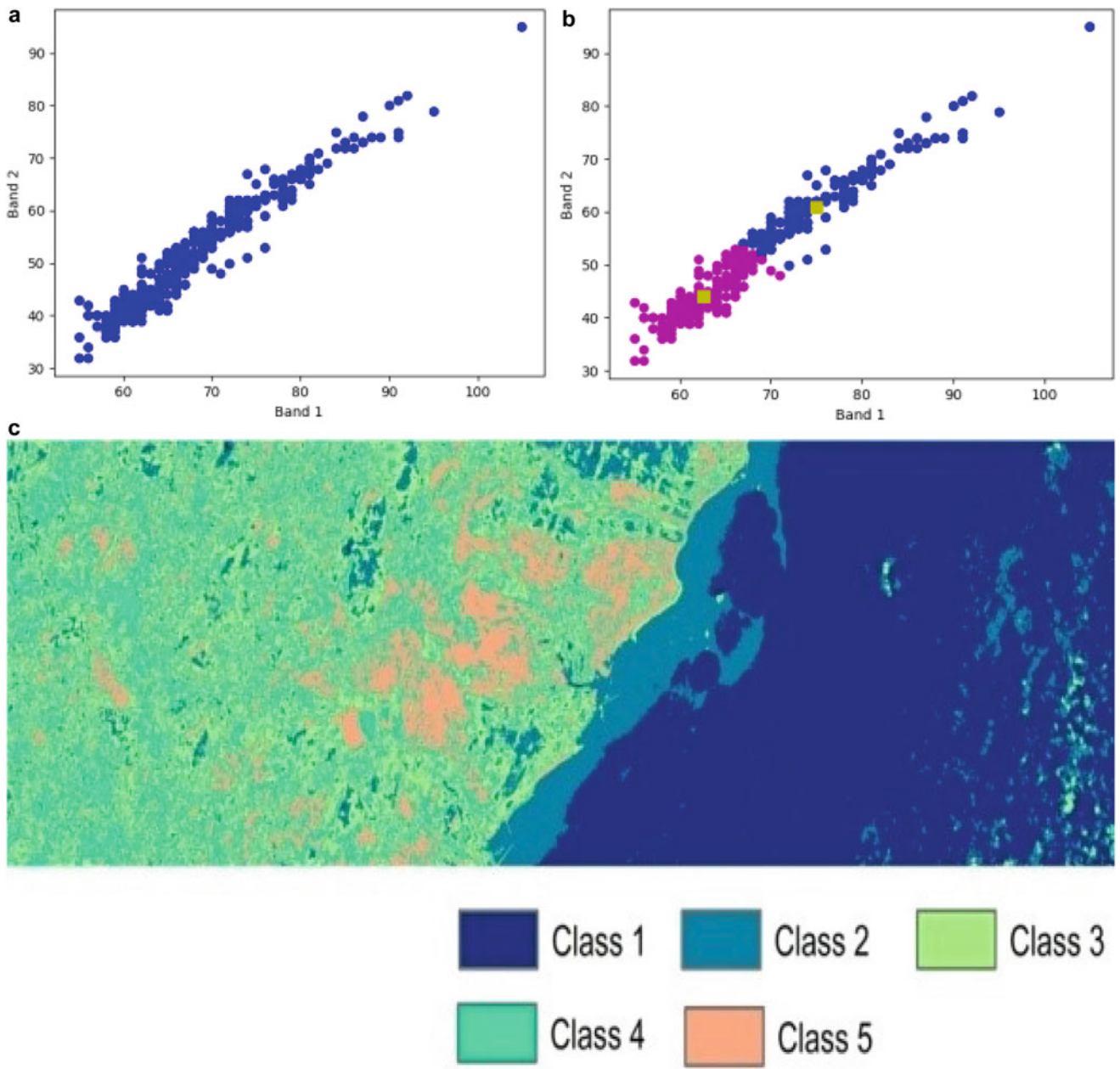
The chromosomes are selected based on fitness value. Furthermore, the selected chromosomes are mutated at the respective crossover points. The procedure is implemented for a number of iterations and converges after reaching the efficient cluster centers (satisfying Eq. (3)).

Example

In this study, the performance of GAs in clustering the data has been evaluated with the IRS LISS III dataset of the Visakhapatnam region, India. The dataset consists of 400 samples without class label. Each pattern in the dataset is characterized by four features (four spectral bands such as bands-4), and 400 pixels are projected into 2D feature space. The distribution of 400 pixels in the bands 1 and 2 feature space is given in Fig. 3a. Two cluster centers generated using GAs are shown in Fig. 3b.

The initial population is considered with two chromosomes such that each chromosome represents a randomly selected cluster center. These chromosomes are encoded into binary sequences to form binary coded chromosomes. The initial population for the dataset is the randomly generated cluster centers in the N -dimensional features. The details of a randomly generated initial population for the present dataset are given in Table 1.

In Table 1, the initial population is generated with two chromosomes representing two cluster centers. Each chromosome consists of four elements representing four features in the dataset. The fitness value (D) of each chromosome is computed, and the chromosome with maximum D value is selected for crossover and mutation. The process is repeated for multiple number of iterations. The final population



Genetic Algorithms, Fig. 3 (a) Distribution of 400 samples in band 1 and 2 feature space. (b) Two cluster centers were generated using GA in the dataset. (c) IRS LISS III image with five clusters (namely, classes 1–5) (Visakhapatnam, India)

Genetic Algorithms, Table 1 Initial population (randomly selected) and final population (obtained after training for 50 iterations)

Population type	Chromosomes
Initial population	60 40 32 28 80 70 48 74
Final population	65 45 25 32 75 61 42 70

Genetic Algorithms, Table 2 Fitness value (D) of the K-means and GAs for a population of 5 is between 3.284 and 5.826; fitness value for GAs is between 3.544 and 3.845 for randomly selected initial population

Population	K-means	GAs
1	3.284	3.845
2	4.524	3.756
3	5.826	3.645
4	3.457	3.673
5	4.321	3.544

contributing to the maximum D value provides efficient cluster centers. In this study, the final population obtained after 50 iterations is given in Table 1. The final population representing the two cluster centers is given in Table 1. The D value obtained with GAs for five different randomly selected initial populations is given in Table 2. In all the five cases with various randomly selected initial populations, GAs produced maximum fitness value and converged to same D value. The convergence of GAs to maximum D value is better than the K-means clustering algorithm as the K-means produced different D values for different instances of initial population. The D values obtained by using GAs and K-means with the present dataset are given in Table 2. The fitness value for GAs is a maximum value of 3.943 in all the five cases while the fitness value of the K-means algorithm is varying between 3.284 and 5.826. The D value with GAs is almost the equal for all the five cases, indicating that in spite of randomly selected cluster centers, GAs produced optimum solution in all the five cases after 50 iterations. The two cluster centers obtained for the present dataset are given in Fig. 3b. The GAs were applied on a large-scale remote sensing image of Visakhapatnam region, India, and the resultant image is given in Fig. 3c. In Fig. 3c, the objects in the image are grouped into five clusters. Furthermore, the number of clusters can be increased to obtain the finite-level information in the images. The image obtained after unsupervised classification is labeled by the field expert. The labeled dataset is used to train classification model, and later, the trained model is used to classify multiple images. Recently, the classification models such as neural networks and deep neural networks have been used to classify large-size images.

Conclusion

In this study, the application of GAs for unsupervised classification is presented with example dataset. GAs were used to obtain the cluster centers for the unlabeled dataset. GAs provided an optimum solution with a randomized search approach. The optimized solution search is implemented through the principles and parallelism of evolution and natural genetics. The implementation steps of GAs such as selection of the population, evaluation of fitness, defining crossover, mutation, and converging to the optimal solution were discussed in this study. The performance of GAs was compared with the K-means algorithm in obtaining appropriate clusters in the unlabeled dataset. GAs with maximum fitness value outperformed the K-means algorithm in obtaining better clusters with remote sensing dataset. Furthermore, GAs provide better clusters in higher-dimensional remote sensing images.

Cross-References

- [Cluster Analysis and Classification](#)
- [K-means Clustering](#)
- [Optimization in Geosciences](#)
- [Particle Swarm Optimization in Geosciences](#)
- [Regression](#)
- [Remote Sensing](#)
- [Spatial Data](#)

Bibliography

- Duda RO, Hart PE, Stork DG (2000) Pattern classification, 2nd edn. Wiley Interscience Publications, USA
- Forrest S (1996) Genetic algorithms. *ACM Comput Surv (CSUR)* 28(1): 77–80
- Maulik U, Bandyopadhyay S (2000) Genetic algorithm-based clustering technique. *Pattern Recogn* 33(9):1455–1465
- Nguyen T (2015) Optimal ground control points for geometric correction using genetic algorithm with global accuracy. *Eur J Remote Sensing* 48(1):101–120
- Theodoridis S, Koutroumbas K (1999) Pattern recognition and neural networks. In: *Advanced course on artificial intelligence*. Springer, pp 169–195
- Xue D, He Z, Chen X, Zhang X (2010) Application of genetic algorithms in remote sensing image processing. In: *2010 international conference on computational intelligence and software engineering*
- Yang MD (2007) A genetic algorithm (GA) based automated classifier for remote sensing imagery. *Can J Remote Sens* 33(3):203–213

Geocomputing

Alice-Agnes Gabriel, Marcus Mohr and Bernhard S. A. Schubert
Department of Earth and Environmental Sciences, LMU Munich, Munich, Germany

Synonyms

Computational geosciences; Computational geophysics

Definition

Geocomputing denotes the use of computer simulation in the field of geosciences, especially in geophysics, to test hypotheses, perform parameter studies, or generate models of planetary interiors. Classical examples include the simulation of seismic wave propagation, porous media flow in reservoir studies, convection in the Earth's mantle, and the geodynamo. Such computational experiments require mathematical-physical equations to describe the system under consideration, numerical

algorithms to determine approximate solutions of these equations, and software implementing the latter such that the simulation can be executed on a computer. In order for the computational experiments to not only return qualitative but also quantitative results, measurements from the rich collection of geophysical datasets (e.g., plate velocities, seismic travel times, or gravity data) are incorporated into the simulations. The large and varied temporal and spatial scales involved in many geophysical questions make simulations often computationally very expensive: For example, capturing a rising plume or subducting slab in a global model of convection in the Earth's mantle might, at least locally, require a resolution of only a few kilometers. Many scenarios, thus, necessitate the use of mid- to extreme-scale parallel supercomputers.

The Why and How

The study of problems in solid earth is complicated by two facts. Firstly, the object under consideration is mostly inaccessible to direct measurements. While for near-surface problems, such as, e.g., determining underground porosity in reservoir engineering, drilling allows to perform sampling on a coarse scale; this is not feasible for most of the interior of our planet. Secondly, performing adequate laboratory experiments is difficult. While mineralogical experimentalists strive to at least partially recreate the enormous pressures and temperatures acting inside Earth, specimen sizes and time scales are vastly different. The same holds for the enormous forces involved in earthquake fracture processes.

Consequently, scientists in the field of geophysics often need to work with indirect observations, i.e., quantities which can be observed on or above the planet's surface and are proxies for internal dynamic processes or the composition of its deeper regions. Examples are measurements of travel times of seismic waves excited by earthquakes, influenced by the material the wave traversed in its path, changes in the Geoid and dynamic topography resulting from convection in the mantle, and associated mass movement or GPS measurements of plate motion.

Observational data alone, however, are insufficient to gain insight. Mathematical models are crucial to connect observations to underlying physics in order to interpret them and extract information. An example would be the comparison of Earth's normal modes to the theoretical model of an ideal ball. The first step is to define the correct equations to use for the model and to decide which effects are important, and which ones can be neglected. Most models in the geosciences involve partial differential equations (PDEs), such as the seismic wave equation in tomography, or the magnetohydrodynamic equations for the geodynamo. Thus, one needs to complement the equations by associated initial and boundary conditions.

After modeling one would like to evaluate the equations for the given data. However, only in the simplest of cases can one find a closed-form analytical solution of the problem. Alternatively, one can resort to numerical techniques from applied mathematics to determine an approximation of the unknown solution. This starts with a process denoted as discretization. It involves replacing the continuous domain, e.g., the Earth's interior, by a discrete computational mesh composed of simple geometric elements, such as tetrahedra or hexahedra. The details of the next step depend on the selected technique, commonly either the method of finite differences, finite volumes, or finite elements. In the former the differentials in the equations are replaced by difference quotients based on the mesh size, while in the latter the unknown analytical solution is approximated by a function from a finite dimensional space, such as piecewise polynomials defined for each element. In all cases the discretization leads to a linear or nonlinear system of equations which needs to be solved to determine the unknowns in the approximate so-called discrete solution. Additionally, as most problems involve time, one also needs to perform discretization of this extra dimension. Usually this involves some form of time-stepping, i.e., one starts from the initial conditions and successively computes approximate solutions at discrete points in time, using previous ones already available.

In order for the discrete solution to be reasonably close to the exact solution, spatial meshes need to be quite fine and time-step sizes need to be small. This results in a large number of degrees of freedom in the system of equations which must be solved for numerously in the time-stepping. Consequently, this cannot be done manually but requires the use of a computer. For this, the numerical algorithms must be implemented in a software program. Once such a simulation was run, its results can be post-processed, e.g., visualized, to allow their interpretation.

Many problems of interest in the solid earth, such as seismic wave propagation, global seismic tomography, earthquake fracture dynamics, convection of the Earth's mantle, reservoir simulations, or the geodynamo, involve spatial and/or temporal scales that require such high resolution that the resulting models fall into the realm of high-performance computing (HPC), (Hager and Weller 2011), and can only be simulated on large-scale parallel supercomputers. In fact such simulations have traditionally been key consumers of CPU hours at academic and industrial supercomputing centers, next to other applications from meteorology, computational chemistry, or astronomy. From 2002–2018, Japan even operated a series of supercomputers termed “Earth Simulator”, dedicated primarily to climate and solid earth physics simulations.

The approach described above is generally referred to as scientific computing or computational science. It can be seen as performing virtual experiments on a computer. It has

significantly gained in importance in the last decades and joined theory and experiment as a third pillar of scientific discovery, see (Gustafsson 2018).

Selected Applications and Challenges

Studying mantle convection is essential to our understanding of how the entire planet works, as the heat transported from deep within our planet to the surface is responsible for most of the geologic activity of Earth. Numerical simulations are an important tool to obtain a quantitative understanding of buoyancy forces in the mantle. Knowledge of these forces is a prerequisite for correctly modeling lithosphere dynamics and tectonic processes. Mantle flow also affects the geodynamo, the process generating Earth's magnetic field. The time scales of geologic processes, on the order of millions to hundreds of millions of years, are sufficiently long that the mantle can be treated as a fluid, although being capable of transmitting seismic shear waves on time scales of seismic waves (seconds to hours). Mantle convection is therefore governed by the equations of motion of a viscous fluid (i.e., the hydrodynamic field equations) expressing the fundamental principles of mass, momentum, and energy conservation. A forward simulation requires physical input parameters for thermal conductivity and expansion, viscosity, specific heat capacity, etc. An equation of state relating density to pressure and temperature is furthermore required to be able to solve this system of coupled PDEs.

The material properties and the energy input (radioactive heating and heat flux from the core) determine the vigor of convection. Owing to the high viscosity of mantle rock, inertial forces are small (Reynolds number $\mathcal{O}(10^{-21})$) and can be neglected in the momentum equation together with Coriolis forces resulting in an instantaneous force balance between viscous resistance and buoyancy, which is referred to as Stokes flow. Once a set of parameters is chosen, a unique solution to the mantle convection equations can be computed given that an initial condition for temperature and boundary conditions for both temperature and flow velocities are provided. Specific nonzero Dirichlet velocity boundary conditions taken from plate tectonic reconstructions can be applied for Earth's surface. Through this sequential data assimilation, a geologically informed prediction of the present-day flow velocities and temperature field in the mantle is obtained.

Such models with earth-like physical parameters require numerous time-steps and extremely fine meshes (~ 1 km resolution at least locally), and models with over 10^9 degrees of freedom can be simulated on large-scale supercomputers, (Bauer et al. 2020; Johann et al. 2015). The past three decades have seen great progress in the adoption of modern numerical techniques and their application in this field. Today a number of powerful software packages exists that allow to simulate

mantle flow in a 3D spherical shell geometry on a global scale on a routinely basis on HPC systems, (Heister et al. 2017).

In these codes the mantle is approximated as a thick spherical shell, which is typically discretized by using either an icosahedral grid, a (modified) cubed sphere, or a yin-yang grid, (Bauer et al. 2020; Ismail-Zadeh and Tackley 2010), and adaptive mesh refinement might be employed. Due to their flexibility, e.g., in incorporating strong parameter variations, and their relaxed smoothness requirements as opposed to finite differences, finite elements are currently the preferred choice for discretizing the hydrodynamic equations. In the resulting discrete equations, it is the solution of the generalized Stokes problem for velocity and pressure one or multiple times per time-step that is the dominating cost factor. For this one relies on iterative solvers, mostly combinations of Krylov subspace methods with multigrid techniques, (Bauer et al. 2020; Ismail-Zadeh and Tackley 2010). For more details on the mathematical and numerical treatment of mantle convection see, e.g., (Ismail-Zadeh and Tackley 2010).

As mantle convection evolves over time scales of millions of years, any testing of predictions for future mantle states is infeasible. While the present-day thermodynamic state of the mantle can be estimated from seismic tomographic models, initial conditions in the past needed for a corresponding forward simulation are unknown. This, however, can be formulated as an inverse problem. Here, one iteratively optimizes the initial condition in the geologic past based on the current, terminal, state of the mantle. Technically this relies on adjoint formulations to compute the gradient of the target functional. Comparing the results of such retrodictions against geologic evidence provides crucial insight into mantle behavior.

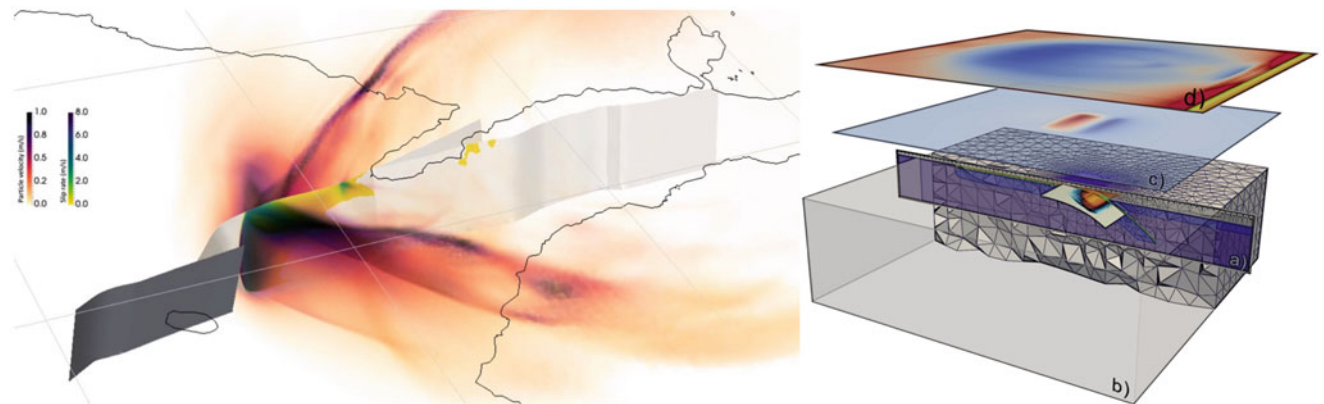
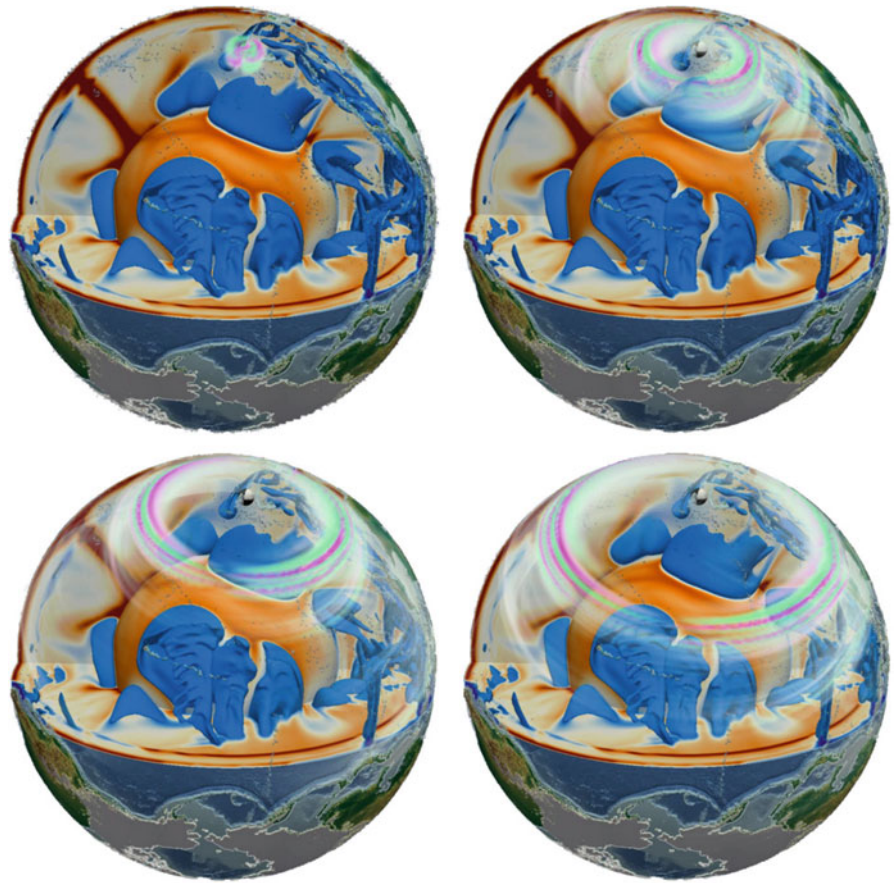
Current challenges in these geodynamic simulations are the huge computational costs for the retrodictions, as they involve multiple forward simulations, and the trend to include complex physical phenomena such as rock compressibility or nonlinear deformation laws into models.

In computational seismology, (Igel 2016), the physics-based forward modeling of the PDEs describing the short-time behavior of elastic solids advanced over the last decades to routinely simulate wave propagation in strongly heterogeneous and complex media (Fig. 1). Taking into account the full 3D complexity of ground motion improves forecasting of strong earthquake-related ground shaking and provides continuously sharper images of Earth's interior via tomography. Since the early 1970s a variety of numerical methods including finite difference, boundary integral, boundary element, pseudo-spectral, finite element, finite volume, discontinuous Galerkin, and spectral element methods were used to demonstrate complex source, path, and site effects on ground motions.

Among these, the spectral element method (SEM) established itself as one of the favored methods for global

Geocomputing,

Fig. 1 Snapshots of the seismic wavefield from a multi-physics simulation using the present-day 3D structure derived from a mantle convection model. 3D global seismic wave propagation was simulated for an earthquake in the Fiji Islands region, (Schuberth et al. 2012). Green and magenta colors depict the wavefield, while shear wave velocity variations in the model are visualized by vertical cross-sections and isosurfaces on a blue to brownish color scale. (Material from: Johnson et al., PRACE DECI (Distributed European Computing Initiative) Minisymposium, published 2013 Springer Berlin Heidelberg, reproduced with permission of Springer Nature)



Geocomputing, Fig. 2 Multi-physics dynamic earthquake rupture and seismic wave propagation simulations which can be linked to tsunami modeling. Left: Physics-based model of the 2018 Palu, Sulawesi, earthquake. The complex dynamic earthquake displacements propagate very fast (at “supershear” speed) and likely played a critical role generating

the Palu tsunami. Right: Illustration of fault slip across a curved mega-thrust (a) emitting seismic waves in a 3D domain (b), and causing seafloor displacement (c) used to source a 2D tsunami model (d). (Figures courtesy of A. Gabriel, adapted from Ulrich et al., *Pure Appl. Geophys.*, 2019 and Madden et al., *Geophys. J. Int.*, 2020)

and regional computational seismology applications (Komatitsch and Vilotte 1998). It is based on a weak formulation of the PDE. This allows to naturally handle both, interfaces and free boundary conditions, yielding very accurate resolution of evanescent interface and surface waves –

phenomena of major importance in seismology. Unknowns are approximated by continuous functions composed of high-order piecewise polynomials. The SEM attains its efficiency from a tensor product representation of these polynomials on hexahedral meshes and a combination of interpolation and

integration points that results in a diagonal mass matrix, which can trivially be inverted in the context of explicit time integrators.

Recently, Discontinuous Galerkin (DG) methods have become a widespread choice for large-scale simulation of hyperbolic problems. DG allows to use meshes composed of triangles and tetrahedra which simplifies meshing of complex geological structures or complex topography, such as volcanoes, sedimentary basins, fault zones, geological interfaces, and sharp impedance contrasts. Further advantages include low numerical dispersion as well as the easy and efficient implementation of local time-stepping. Their use of numerical fluxes allows to naturally include nonlinear interface conditions, arising, e.g., in dynamic earthquake rupture problems, while combination with an arbitrary high-order derivative (ADER) time integrator allows high-order accuracy in time within a single-step scheme.

Understanding the dynamics of earthquakes and faults poses a multiscale, multi-physics problem with profound societal implications, such as secondary effects with unforeseen hazard complexity (Fig. 2). Earthquakes can be represented as frictional shear fracture of brittle solids under compression along internal boundaries of preexisting weak fault interfaces. During the highly nonlinear interaction of frictional failure and seismic wave propagation computational models may, in addition, include multi-physics processes like off-fault deformation or thermal pressurization of rock pore fluids. DG methods have been particularly impactful in this context having the potential to overcome the limitations of short-term (incomplete) earthquake data by incorporating field data and laboratory measured rock behavior.

High-resolution earthquake shaking and rupture modeling based on elastodynamics has been advanced significantly by, and in turn contributed to, large-scale HPC. Recent computational advances now allow to capture nonlinear rupture dynamics on complex faults or fault systems (Fig. 2) up to the scale of megathrust events and on extreme-scale supercomputers (Uphoff et al. 2017).

HPC-empowered computational seismology and earthquake modeling is a key tool to better understand the structure of Earth's interior and sources of seismic energy. It is becoming an important part of the rapid earthquake response toolset by delivering physics-driven interpretations that can be integrated synergistically with data-driven efforts. Ongoing challenges include computational efficiency (resolving smaller-scale heterogeneities affecting the high frequencies of the wave field), addressing model sensitivities and the need for community solutions. Combination with emerging machine learning approaches may close the gap of data quality in real-time conditions, boosting our understanding of earthquake source processes and the associated ground shaking.

Conclusions

We demonstrated that the application of scientific computing to geoscientific problems, denoted as geocomputing, is of major importance in this field. Two challenges are the constant strive to get simulations closer to reality, leading to more sophisticated but also more costly computational models and that many problems of interest can only be simulated on supercomputers.

Cross-References

- [Cloud Computing and Cloud Service](#)
- [Computational Geoscience](#)
- [Deep Learning in Geoscience](#)
- [Differential Equations](#)
- [Forward and Inverse Stratigraphic Models](#)
- [Geoinformatics](#)
- [Geostatistics](#)
- [Machine Learning](#)
- [Mathematical Geosciences](#)
- [Simulation](#)
- [Statistical Computing](#)

Bibliography

- Bauer S, Bunge H-P, Drzisga D, Ghelichkhan S, Huber M, Kohl N, Mohr M, Rüde U, Thönnies D, Wohlmuth B (2020) TerraNeo – mantle convection beyond a Trillion Degrees of Freedom. In: Bungartz H-J, Reiz S, Uekermann B, Neumann P, Nagel W (eds) Software for exascale computing – SPPEXA 2016–2019, vol136 of Lecture notes in computational science and engineering. Springer, pp 569–610. https://doi.org/10.1007/978-3-030-47956-5_19
- Gustafsson B (2018) Scientific computing – a historical perspective, volume 17 of Texts in Computational Science and Engineering. Springer. <https://doi.org/10.1007/978-3-319-69847-2>
- Hager G, Wellein G (2011) Introduction to high performance computing for scientists and engineers. CRC computational science series. CRC Press, Taylor & Francis Group, Boca Raton, London, New York. ISBN 9781439811924
- Heister T, Dannberg J, Gassmöller R, Bangerth W (2017) High accuracy mantle convection simulation through modern numerical methods – II: realistic models and problems. *Geophys J Int* 210(2):833–851. <https://doi.org/10.1093/gji/ggx195>
- Igel H (2016) Computational seismology: a practical introduction, 1st edn. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198717409.001.0001>
- Ismail-Zadeh A, Tackley P (2010) Computational methods for geodynamics. Cambridge University Press. <https://doi.org/10.1017/CBO9780511780820>
- Johann R, Malossi ACI, Isaac T, Stadler G, Gurnis M, Staar PWJ, Ineichen Y, Bekas C, Curioni A, Ghattas O (2015) An extreme-scale implicit solver for complex PDEs: highly heterogeneous flow in Earth's mantle. In: Proceedings of the international conference for high performance computing, networking, storage and analysis, SC '15. ACM, New York, pp 5:1–5:12. <https://doi.org/10.1145/2807591.2807675>. ACM Gordon Bell Prize 2015

- Komatitsch D, Vilotte J-P (1998) The spectral element method: an efficient tool to simulate the seismic response of 2d and 3d geological structures. *Bull Seismol Soc Am* 88(2):368–392
- Schuberth BSA, Zaroli C, Nolet G (2012) Synthetic seismograms for a synthetic Earth: long-period P- and S-wave traveltime variations can be explained by temperature alone. *Geophys J Int* 188(3):1393–1412. <https://doi.org/10.1111/j.1365-246X.2011.05333.x>
- Uphoff C, Rettenberger S, Bader M, Madden EH, Ulrich T, Wollherr S, Gabriel A-A (2017) Extreme scale multi-physics simulations of the tsunamigenic 2004 Sumatra megathrust earthquake. In: *Proceedings of the international conference for high performance computing, networking, storage and analysis, SC 2017*. <https://doi.org/10.1145/3126908.3126948>

Geographical Information Science

Parvatham Venkatachalam

Centre of Studies in Resources Engineering, Indian Institute of Technology, Mumbai, India

Definition

Geographical Information Science (GISc) is a scientific research discipline that provides the technology to handle large volumes of spatial data derived from a variety of sources, study the characteristics and relationships in the data, model, analyze, and provide the development strategies as per the user defined requirements. GISc has evolved over the past six decades generating massive interest worldwide. GISc is the science behind the development of Geographical Information System (GIS) technology and GIS software tools. It provides a spatial framework to assist optimal use of natural resources. Developments in spatial data standards, interoperability, World Wide Web, and the availability of spatial data infrastructure are making GISc a core scientific area in resources management.

Introduction

Data are the raw figures collected from varied sources for specific purposes. When the data are structured, organized, processed, and presented to make them useful and relevant to a context, data gets converted into information. Information science deals with the study of data collection, classification, organization, storage, retrieval, analysis, and dissemination of information. Geographical Information Science (GISc) deals with the research study of geographical/spatial data. GISc is a multidisciplinary field and there has been a rapid development in GISc due to the contributions of experts from several disciplines such as geography, cartography, geodesy,

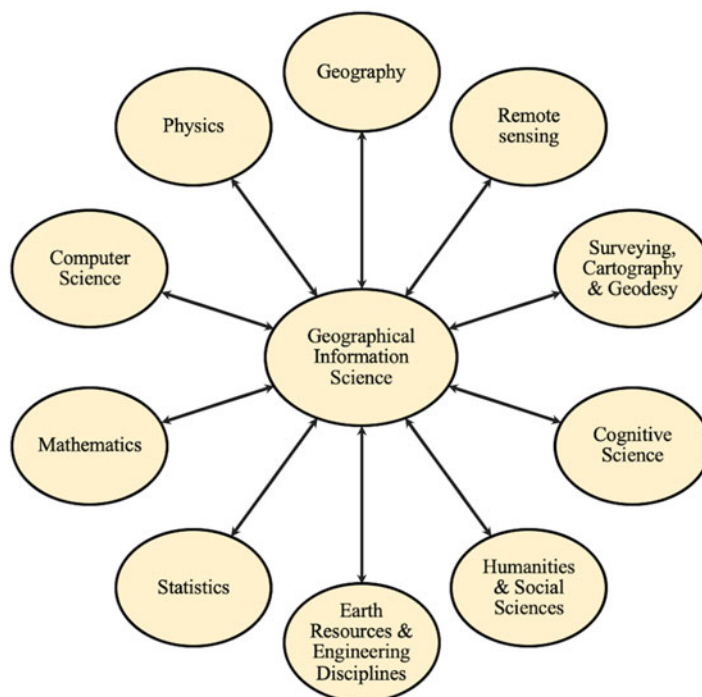
photogrammetry, mathematics, statistics, computer science, remote sensing, earth sciences, engineering fields, humanities, and social sciences (Fig. 1). Geographical Information Science operates through the techniques, methods, and approaches associated with GIS and seeks to redefine the geographical concepts and their use in the context of GIS (Viana et al. 2019). It focuses on spatial data collection, accuracy assessment, data models and structures for efficient storage, spatial relations inherent in the data, interpretation, modeling, analysis, and finally the visualization and presentation of the results for end user benefits. The components of GISc are shown in Fig. 2.

Geographical information science and the geospatial technology using digital platforms started in mid-1960s. But the manual methods of using qualitative maps for decision making existed centuries back. They used general purpose maps and thematic maps for various applications. During 1940s, use of statistical methods for spatial analysis started. With the availability of computer technology and the advent of remote sensing satellites and global positioning systems (GPS), GISc started evolving using digital spatial data in decision making. There is no strictly a logical progress or evolution toward the development of geospatial technology (Venkatachalam 2009). Initiatives were taken in different parts of the world by academia, government agencies, and industries to use computer technology for spatial data handling and analysis. With the introduction of internet, Internet of Things (IOT), drones, and high resolution satellites, spatial data are getting generated faster than they can be analyzed. Today millions of professionals use geospatial technology and it is playing a central role in e-government initiatives. Geospatial technology has become a part of information technology and provides enterprise solutions. This entry discusses briefly various components of GISc in the present scenario and the open research challenges in the field of geographical information science.

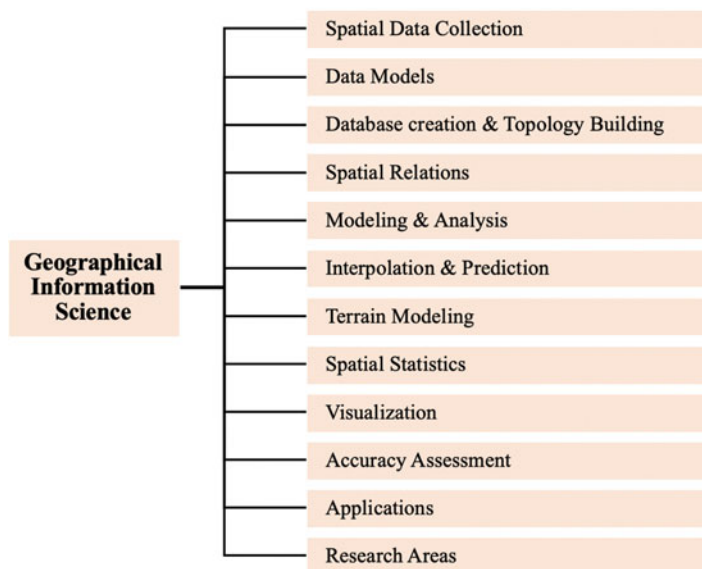
Spatial Data Concepts

Geography of the Earth varies continuously and is very complex to quantify digitally and bring it within distinct boundaries. To describe an area, key features existing in that area are identified simplifying the inherent complexity. The features get described as water body, built-up area, hilly terrain, or individual locations such as building, electrical pole, and so on. These geographical features are related to each other in terms of in the left, in front of, etc. These conceptual data get converted into digital form using spatial data models. We view the real world through this digital data base. To describe a school spatially, we give its location (Lat/Long) and its properties (name, address, number of students, etc.). Location

Geographical Information Science, Fig.1 A Multidisciplinary Field



Geographical Information Science, Fig. 2 Components of GISc



gives the spatial information and associated properties give its characteristics.

Spatial database contains digital information of real objects such as water bodies, roads, buildings represented as point features (buildings, line features (roads), and area features (lakes). Some features vary continuously on the Earth such as temperature, elevation, vegetation, etc. By dividing the study area into zones with the assumption that the feature is constant within each zone, measurements are taken at sample points within each zone. Also contours are prepared connecting points of equal value with lines.

A spatial entity (spatial object) is identified by a single identifier and it cannot be subdivided into like units. The spatial object can be a point/line/area feature. The indivisibility is based on the characteristics used to define the object and one cannot expect complete homogeneity within the spatial extent of the object. A spatial object is defined by an identifier (name), position on the Earth's surface (location), character, and function of the object and its spatial properties (Laurini and Thompson 1994). In general, spatial features are represented digitally in three ways. A point object has a position in space and it has no length and has zero dimension

(e.g., a building, a tower). A point can be a node or vertex. A line is a one-dimensional object and has the property length. A line has two end points and points in between to mark the shape of the spatial feature (e.g., road, contour). An area is a two-dimensional object and has the properties area and perimeter. An area can be alone or can share the boundaries with other areas. Existence of holes within an area (pond inside a village) means that the area has external and internal boundaries. An area is generally called as polygon.

Spatial features are represented as point or line or area depending on the scale of the original map source, the purpose of the study, and other factors. A city may be a point on a small scale map (1:250,000) and can be an area in a large scale map (1:25000). The map scale of the source document helps to decide the level of details to be stored in the spatial database.

A spatial database has location, spatial relationship with other data, attributes or the characteristics, and time wherever essential. The attributes are stored in a table where each spatial object corresponds to a row in the table and the attributes correspond to columns in the table. Temporal information can be stored by giving interval of time over which the object exists or giving values at certain points of time. Relationships between spatial objects can be defined in terms of proximity, distance, intersect, containment, touch, disjoint, etc. For example, a point object can be in north of another point object or can be on a line object or can be inside a polygon object. During spatial analysis, new objects can be generated from original objects. Area object can be transformed to point objects using centroid calculations. Thiessen polygons generate area objects from point objects.

The spatial database contains both spatial and attribute data. In a geospatial database, the spatial data is stored using vector or raster data structure in a file while attribute data is stored in a relational database. Unique ID given to each spatial object helps to link the spatial and attribute data. Most well-known GIS packages use the geo-relational database model. Some of the recent GIS tools try to store both spatial objects and attributes in a single relational table using object-relational model.

Spatial Data Sources

Spatial database is created collating data from variety of sources and is one of the most important parts in geospatial technology. It is an expensive task. Data comes from topographic sheets, surveyed maps, government records, and remote sensing data including satellite images, aerial photos, and drones. Plenty of data may be available in different agencies but accessibility, accuracy, and quality are important issues. Available and accessible data may not be in digital form. Presently, government departments globally are

working on creating spatial data warehouses with necessary metadata information. Open Geospatial Consortium (OGC) regularly brings out spatial data standards and data exchange formats so that interoperability becomes feasible at the user end. Based on the time, cost, and scale of problem, one can generate spatial data from primary sources or collate from secondary sources.

Remote sensing data is one of the main sources for primary data generation. With the availability of high resolution images from satellites, air platforms, and drones, very detailed spatial information can be generated. They also help in updating old thematic information. Digital orthophotos help in extraction of elevation data with improved vertical resolution. In addition, field survey helps to collect primary data (e.g., soil, geology, etc). Global Positioning System (GPS) is a common affordable tool for getting geographical location and elevation information. Secondary data sources are the national survey departments and they provide topographic sheets, thematic maps, and census data on demography, socioeconomic information, agriculture, etc.

When data are collated from different sources, we must know the scales of measurement used for numeric values. That determines the type of mathematical operations that can be performed on the data. The numerical values are defined with respect to nominal, ordinal, interval, and ratio scales of measurement. On nominal scale, numbers give identity while in ordinal scale, the values give the order (1st, 2nd, etc.). In interval scale, the differences between values are given and the scale does not start at zero. Hence, subtraction makes sense and not division. In ratio scale, measurement has an absolute zero and division makes sense.

Geographical Coordinate System

Even though the Earth appears flat at a closer look, it is relatively spherical. The geographical coordinate system (Latitude and Longitude) helps to specify quantitatively the location of any spatial feature on the Earth's surface. Latitudes are measured in degrees toward north or south from equatorial plane starting from 0° to 90° . Longitudes are measured in degrees toward east or west from Greenwich starting from 0° to 180° . Using these coordinates, distances and areas are calculated using spherical geometry and radius of the Earth. Lines connecting points of same longitude are called meridians. The prime meridian passes through Greenwich, England, and referred as 0° meridian. Lines connecting points of same latitude are called parallels. Any two parallels are always the same distance apart. Meridians are farthest apart at the equator and converge to a single point at the pole. Parallels and meridians cross one another at right angles.

The Earth is not a perfect sphere. It is wider along the equator than the poles. Hence it is approximated to an

ellipsoid and the difference between the major axis (along the equator) and the minor axis connecting the poles is approximately 21 kms. As the Earth has surface irregularities, an ellipsoid that fits to one region or country is chosen as the ellipsoid for that region. When an ellipsoid approximates the shape of the Earth, a datum defines the position of the ellipsoid relative to the center of the Earth. Many countries have their own datums. An earth-centered (geocentric) datum using the Earth's center of mass as the origin is the most widely used datum known as WGS84 (World Geodetic System 1984). While doing spatial data analysis, all data layers must be on the same reference datum.

The vertical datum is used while measuring elevation. GPS gives elevation from the ellipsoidal surface. A geoid is a closer approximation of the Earth, which is considered as the mean sea level in that region. In order to get elevation from mean sea level (geodetic) of a land feature, the difference between the geoid surface and the ellipsoid surface must be measured locally.

A scale of a map is the ratio of the map distance to the ground distance. A 1:2500 scale map means a map distance of 1 cm on the map represents 2500 cms on the ground. Generally a 1:2500 scale map is called larger scale map and 1:250,000 scale map is called small scale map as a 1:2500 scale map gives more details corresponding to a smaller area compared to 1:250,000 scale map.

Map Projections

A set of mathematical techniques are developed to project 3-dimensional spherical surface of the Earth on a 2-dimensional planar surface, which are called map projections. The projected map gives a systematic arrangement of parallels and meridians on a plane. However, the map projection involves distortion in terms of shape, area, and distance and no map projection is perfect. Every map projection preserves certain spatial properties while sacrificing other properties. Choice of a map projection depends on the purpose of the map, type of area covered, extent, location, properties to be preserved, needed accuracy, etc. Map projections are of three types: conformal, equivalent, and equidistant. A conformal projection preserves local angles and shapes while an equivalent projection preserves areas. An equidistant projection preserves distance in a direction or at a place.

The principle of map projection is by keeping a light source inside a globe and projecting the spatial features of the globe onto a 2-dimensional surface surrounding the globe. A cylinder, a cone, or a plane paper can be used as a projected surface. The process of projecting a spherical surface onto a projected medium is done using mathematical principles of geometry and trigonometry. All projections are grouped as cylindrical projections, conical projections, and azimuthal

projections. The orientation of a projection surface can be changed as needed. A cylinder is normally oriented tangent along the equator and a cone is normally oriented tangent along a parallel. A plane is normally oriented tangent at the pole. If the projection surface is changed to 90° from normal, the result is a transverse projection. If the projection surface is at an angle between the normal and transverse position, the result is an oblique projection. The line of tangency between a projection surface and the globe surface is called standard line and there is no distortion on that line. When the projection surface intersects the globe, the result is a secant projection with two standard parallels. In the case of azimuthal projection, the light source can be at the center of the globe or at the opposite pole or from infinite position.

To get the coordinates on a projected map, the midpoint of the central parallel and central meridian becomes the origin of the coordinate system and the coordinates are calculated as per four quadrants. As the map is generally oriented toward north, x-value in the east direction is called easting and y-value in the north direction is called northing. The most well-known projection globally is the Universal Transverse Mercator (UTM) projection. It uses varied cylinders with different central meridians for different zones to minimize the distortions. There are 60 zones, each zone covering 6° longitude and numbered sequentially starting with 1 at 180° W and moving toward east. Each zone is further divided into north and south hemispheres.

Spatial Data Models

The continuous geographical variations of the real world have to be approximated and abstracted for converting them into quantitative or digital form in finite dimensional space. A spatial data model converts these variations into discrete objects. A spatial data model gives a set of rules for representing spatial objects and the spatial relations between them. Two major spatial data models are raster data model and vector data model. A geographical space can be considered as a container of a series of thematic map layers such as land use, drainage, topography, soils, administrative boundaries, etc. The spatial objects can be categorized into points, lines, and areas. Points represent features such as schools, buildings, wells, electric poles, etc. while lines represent roads, topographic contours (lines connecting points of equal heights), drainage, etc. Area features can be water bodies, land use boundaries, administrative boundaries, etc.

In a raster data model, grid is a unit and the value in each grid gives the characteristic of the spatial feature in that grid. A raster data layer has rows, columns, and grid values. Grids are also known as pixels. Well known examples are remote sensing images and scanned maps. A raster layer is a matrix with rows and columns and grid values are stored in

2-dimensional array. Each grid in the raster model is intrinsically related to other grids in terms of row and column position. Hence spatial relations are explicit and inbuilt in raster data model. Point features are shown by single grids, line features are shown by a sequence of neighboring grids, and area features by a cluster of contiguous grids. The grid value can be binary, byte, or integer or a real value based on the theme it represents. The grid size gives the resolution of the raster data. If the grid resolution is smaller, data accuracy will be better. However, the data volume and processing time will increase. Remote sensing images and digital elevation models (DEMs) use raster data structure and grid values continuously change to bring out high spatial variability. When spatial variability is not very high (e.g., village boundaries, land use), data compression methods are used to reduce the size of the data file. Well-known data compression methods are run-length encoding, value point coding, and quadtree. In a raster data layer, a header file gives metadata information, geographical coordinates of corners of the map layer, rows, columns, resolution, etc.

The vector data model helps in precise positioning of spatial features. The map layer is assumed to be in a continuous coordinate space and features are represented in terms of x,y coordinates in Cartesian coordinate system. A point feature is given by a single x,y coordinate pair, a line by a sequence of x,y coordinates, and area as a closed boundary of x,y coordinates. Along with spatial objects, corresponding attributes are stored in a separate table.

Initially, spaghetti model had been in use as vector data structure. A point is encoded as a single x,y coordinate and a line feature as a series of x,y coordinates with a start node and an end node. An area feature is treated as a polygon and is stored as a closed loop of x,y coordinates that defines its boundary. In spaghetti data structure, the common boundary between neighboring polygons is stored twice once for each polygon. The limitation in this data structure is that the spatial relationships among the spatial objects such as polygon inside a polygon or a polygon adjacent to another polygon are not getting encoded.

To overcome these limitations, topological data model has been adopted. Topology is the study of those properties of geometric objects that remain invariant under certain transformations such as enlarging, skewing, bending, stretching, etc. Basic entity in this model is an arc (a series of x,y coordinates) with a start and end node. In the case of network data such as road network, for each arc along with its own arc ID, the connecting arcs' IDs on both of its ends are also stored. Also the minimum bounding rectangle (MBR) for each arc gets stored to assist in query search. In the case of polygon data, along with each arc, the IDs of the polygons which fall on the right side and left side of the arc get stored. This avoids duplication of common boundary between polygons and at the same time, query such as adjacency can be

done easily. It also helps to store polygons inside polygons or nested polygons.

The topological data model helps to store efficiently each polygon, its area, perimeter, minimum bounding rectangles (MBRs), and polygons inside polygons efficiently. In addition to points, lines and polygons, regions and routes can also be stored in vector data structure. A region is a cluster of polygons with uniform characteristics and the polygons can be joint or disjoint. Example is a group of islands like Andaman and Nicobar islands that can be stored as a single spatial entity. Similarly, a route consists of many arcs like a highway or a stream. In vector data structure along with spatial objects, associated attributes are stored in suitable database packages.

Triangulated Irregular Network (TIN) is a topological model to represent terrain elevation data. It stores terrain surface as a set of nonoverlapping triangles with the assumption that each triangle has a constant gradient. Flat surface can be represented by fewer but larger size triangles while the terrain with high variability in elevation has to be represented by denser but smaller size triangles. From the irregularly sampled elevation points, a TIN data model is built using Delaunay Triangulation method. It uses the principle that the triangles have to be as equiangular as possible and the circumcircle of a triangle should not contain another sample point. This model helps to handle even features such as ridge lines, valleys, high peaks, etc. In TIN model, along with vertices, slope and aspect of each facet also get stored.

Spatial Database Creation

In any geospatial technology application project, around 75% of the cost goes in spatial database creation. It is a time taking, labor-intensive and error prone task. Many countries have started national spatial data warehouses, and users can download the data on payment or free of cost. As digital data may be in varied formats, global agencies are working on spatial data standards and interoperability among the datasets.

Generally, the original maps are in paper forms in different scales and projections. These maps have to be scanned, manually digitized (vectorized), cleaned, and edited to convert them to topological structured data. In order to bring all map layers of a study area (e.g., drainage, contours, land use, soils, forestry, administrative boundaries, etc.) to a common reference system, several transformations may have to be carried out such as projection and datum change, geometric transformation, etc.

In the first step, the paper maps are scanned using the best resolution feasible. Then vectorization of spatial features using on-screen digitization method is done manually or using semiautomatic method like line follower. Spatial data cleaning is an important step so that spatial relations can be built using topology. Several errors such as undershoots,

overshoots, duplicate arcs, and sliver polygons (generally obtuse angled thin elongated triangles) have to be removed. Many GIS packages provide automated methods to clean these errors using the tolerance limits given by the user.

Geometric transformation or geo-registration is a step to bring all thematic layers into a common reference system. This step is done in a similar way like in remote sensing and photogrammetry. Using a set of known control points (tic marks or ground control points) between the base map and slave map, a mathematical relation (a second/third order polynomial) is formed. Using that polynomial, the slave map gets transformed to base map coordinate system. In the case of transforming a satellite digital image to a raster map layer coordinate system, in addition to geometric transformation, pixel resampling has to be done using nearest neighbor or bilinear interpolation or cubic convolution method. In geospatial technology, in addition to spatial objects, the associated attributes (characteristics/ properties) have to be handled. Most of the GIS tools use relational database model to store attribute data either in a single table or multiple tables and link the corresponding spatial objects using unique IDs.

Spatial Relations

A spatial relation describes how an object is located in relation to another object in space based on geometric properties. In the case of geospatial science, it is used to describe the spatial relations among point, line, and polygon objects. There are three types of spatial relations, namely, topological, directional (cardinal), and distance (metric) relations. Directional relation specifies the direction where an object is located with respect to a reference object (e.g., left, right, north, south, etc.). Distance relation explains how far an object is located with respect to a reference object in terms of distance and angle. In addition, there can be spatial relations between objects in terms of morphology such as size, shape, and structure of the objects.

Topological relations are spatial relations that are invariant under topological transformations such as translation, rotation, and skewing (Egenhofer and Franzosa 1991). Topology helps to capture the most important spatial properties. Space can be defined as 0-dimensional (point), 1-dimensional (x-axis), 2-dimensional (x-y plane), and 3-dimensional (x-y-z volume) space. Similarly a spatial object can be defined as 0-dimensional object (a point), 1-dimensional (a line), 2-dimensional (a polygon), and 3-dimensional (a volume). The difference between the dimension of the embedding space and the dimension of the spatial object is called co-dimension. When a point is in a plane (a well in a village), the co-dimension is 2. Topological spatial relations are described based on the spatial object and the space where it is embedded. A spatial object A has 3 components: an interior

(A°), a boundary (∂A), and an exterior (A'). An example using 2 polygons (A and B) in a plane (R^2) is explained. The polygons A and B can be topologically related using the fundamental distinction of intersection: empty ($\emptyset$) and nonempty ($-\emptyset$). With 2 possibilities and 9 intersections (3×3), the number of possible topological relations is 2^9 and all the possible 9-intersection (9-i) matrices can be physically written. However, one cannot get a geometrical figure for each of those matrices. It depends on the type of the object and the type and the dimension of the embedding space. In the case of 2 polygons A and B in a plane (R^2), the eight spatial relations and the corresponding 9-i matrices in terms of $\emptyset$ and $-\emptyset$ are given below (Fig. 3).

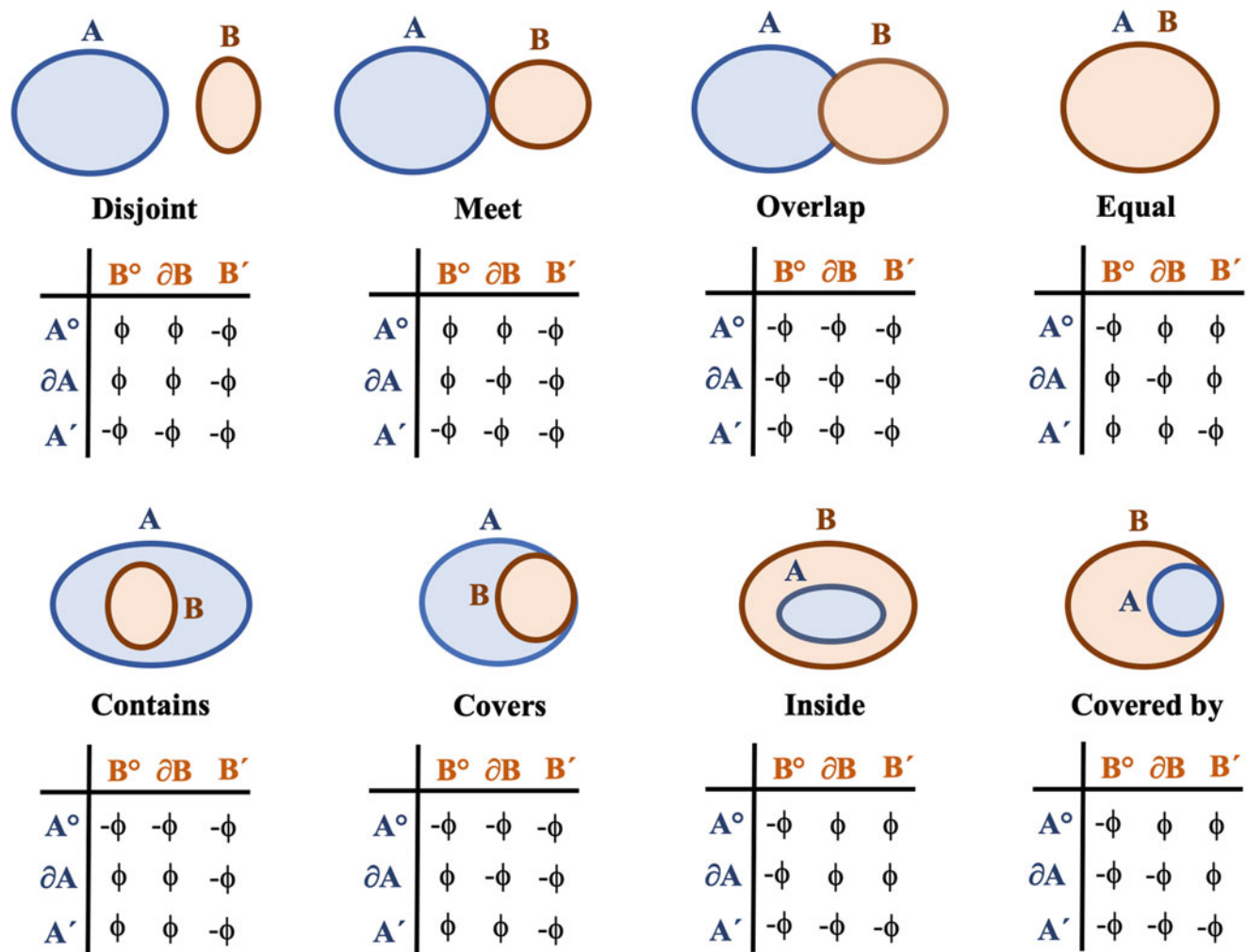
Spatial Data Analysis

The data analysis is the main activity in geospatial technology. Based on the application, it can perform simple data retrieval to a complex modeling analysis. The method of analysis depends on the data model that has been used to represent spatial data.

Vector Data Analysis

In vector data model, the attributes of spatial objects are stored in a database. The simplest analysis function is to run a query in attribute database and show the results both in tabular form and the corresponding spatial objects in a map form on the screen. The query can be a simple query on a single table or a complex query joining multiple tables. The query is given in SQL (structured query language) format. For example, highlight the villages in a district which have a population greater than 10,000 and a bank. Also using any attribute, density map can be generated. In these types of attribute analysis, existing spatial objects are used and no new spatial object is generated.

Under vector data operation, as spatial objects are stored in point, line, and polygon layers, the queries can be executed using two layers. The spatial operations can be categorized under union, intersect, identity, clip, update, erase, and split. All overlay methods are based on the Boolean connectors AND, OR, XOR (Chang 2009). For example union operation combines two layers and generates a new polygon layer with topology and a new attribute table. It is equivalent to the Boolean operation OR and is a spatial join operation. In addition to two layers operations, single layer operations such as buffer, eliminate, and dissolve can be done. Buffer operation helps to generate buffers of specific distances around a spatial object. Any complex spatial analysis can be made into a stepwise flow chart identifying suitable operations and can be executed using these capabilities. The



Geographical Information Science, Fig. 3 9-i Intersection Matrices for 2 Polygons in R^2

important aspect is that all layers must be brought to a common reference system before starting the analysis.

Raster Data Analysis

In raster data structure, regular grid pattern is used to cover the space and the value in each grid gives the characteristic of the spatial phenomena in that grid. In raster data analysis, operations are done on these grid values and the operations can be categorized as local, focal, and zonal operations.

In local operations, a new raster layer gets created from one or more input layers. The value of each new grid is computed using the values of the same grid of the input layers based on a mathematical or trigonometric or logarithmic function. In local operations, the neighboring or distant grid values of a layer have no effect. A well-known local operation is map algebra, which is similar to band algebra in digital image analysis of remote sensing data. In map algebra, the input

raster layers can be combined using various conditions involving algebraic, logical, and relational operators. In addition, under local operations, raster layers can be overlaid (similar to union in vector data model) and a new layer can be created.

In focal operations, a single input layer gives a new output layer and the grid values for the output layer are computed using the neighborhood of the grid values of the input layer. Similar to low pass/high pass filtering operation in image processing, filtering can be done on raster layer with a 3×3 or 5×5 moving window giving suitable weights. Another important utility is the generation of slope and aspect maps from the input layer of elevation values. Buffering, proximity maps (thiessen polygons), and distance mapping from a given facility can be carried out.

Zonal operation uses a zone (a group of grids that have the same value) map and a thematic map (rainfall, elevation, temperature, etc.). This operation computes the minimum, maximum, mean, standard deviation, etc., for each zone

based on the thematic data and generates a table giving the statistical parameters for each zone and also corresponding maps.

The raster data model uses a simple data structure and is very useful when multiple thematic layers have to be combined using various criteria. It is desirable to use raster data model for applications related to natural resources management. Vector data model is more suitable for network analysis such as shortest path finding, location-allocation problems, infrastructure planning, etc.

Spatial Data Interpolation

Very often, the data are available at few sample points and using a suitable algorithm, one has to estimate the values at other points in a study area. One of the examples is elevations data. Spatial interpolation methods use the data values at known points and estimate the values at other points. The principle behind these methods is that the points closer in space are more likely to have similar values than the points far apart. There are several spatial interpolation techniques used in spatial data science.

A global interpolation technique uses every known data point to estimate the values at other points. It results in a smooth surface. A local interpolation technique uses a small portion of known data points to estimate the values in nearby points. It brings out local variations. Exact interpolation methods preserve the values at known points while the approximate interpolation methods may get new values at known points. Some interpolation techniques are deterministic and some are stochastic, which help to assess the predicted errors.

Trend surface modeling is a global technique where a first-order (linear) or a second-order (quadratic) or a third-order (cubic) polynomial equation is formed using the data at known points. The resultant polynomial is used to calculate the data at unknown points.

A well-known method is Inverse distance Weighted (ICW) method where the estimated value of a point depends more on nearby points than far off points.

The general formula is

$$z = \left(\sum_{i=1}^n w_i z_i \right) / \left(\sum_{i=1}^n w_i \right)$$

where w is a function of distance such as $1/d^k$, z is the value to be estimated, z_i are the known values for n number of points, d_i is the distance between the i^{th} point and the point for which the value is estimated and k is a power. A variety of algorithms can be generated by changing the distance function w , varying

the number of sample points and changing the directions from where these sample points are selected.

Kriging is a spatial interpolation technique based on Geostatistics principles. The basis of this technique is the rate at which the variance between points changes over space. This is expressed in terms of variogram, which shows how the average difference between values at points changes with distance between points. Many kriging algorithms are available such as ordinary kriging, universal kriging, simple kriging, block kriging, etc. When the number of input sample points is high, kriging method becomes computationally intensive.

Digital Terrain Mapping

The terrain is a continuous surface with undulations, and digital terrain modeling helps to visualize this undulating terrain in a map form. Generally, digital terrain data is given by the elevation value (z) above the mean sea level at known points. In raster data model, z value gives the elevation of a grid while in vector data model, z value is given as an attribute for a point object or for a line object (e.g., contour). The representation of elevation in raster data model is called as digital elevation model (DEM) and in vector format it is known as Triangulated Irregular Network (TIN) model.

A digital elevation model is created using suitable spatial interpolation technique from the input of known elevation data at regularly or irregularly spaced points. Input data sources are field measurements using GPS, stereo images, digital contours, Radar, LIDAR data, etc. The quality of DEM depends on the spatial and vertical resolution of the input data. The quality of DEM influences the accuracy of the terrain measures such as slope, aspect, automatic delineation of drainage network, etc.

A Triangulate Irregular Network represents the terrain data with a series of nonoverlapping triangles. There are several algorithms to generate TIN from the input data. Input source data can be from points, contours, breaklines such as ridges, valleys, etc. Delaunay triangulation algorithm is a well-known TIN generation algorithm. In this technique, all the nodes are connected to the nearest neighbors to form triangles such that the triangles are as equiangular as possible and the circumcircle of a triangle does not contain another sample point inside it. It is advisable to include sample points beyond the boundary of the study area during TIN generation and then clip out the study area. This helps to avoid elongated and stretched triangles along the boundary.

The most common method of terrain mapping is contouring and it can be generated from a DEM or TIN. Contours connect points of equal elevation and the contour interval is the elevation difference between two adjacent contours. Contours do not intersect one another and do not

stop in the middle of the map. Using the terrain data, vertical profiling of a proposed road or canal alignment can be plotted to calculate the earth work (digging/filling) volume. Hill shading technique simulates the terrain view with the intersection between sunlight and the surface. Three-dimensional view (Perspective view) of the terrain helps to view the terrain in various rotations and navigate through it. Another advantage is the generation of intervisibility maps, which are useful in defense applications, telecommunication, etc. Slope and aspect are the important parameters that are derived from terrain data. Slope gives the maximum rate of change of elevation at a location and aspect gives the direction of the slope. Both these parameters are used in watershed mapping, site-suitability mapping, road/canal alignment, etc.

Network Analysis

A spatial network is defined as a set of geographical locations interconnected by several routes. It is a system of linear features connected at intersections (nodes) and it can be defined as a directed graph. It can be stored as a system of linear features using a topological structure. A network is represented as node-arc incidence matrix or node-node adjacency matrix or forward star representation. Common applications of network analysis are finding the shortest route from a source and destination, locating a closest facility, and solving location-allocation problem.

In the case of road network, the database consists of spatial features such as road links and nodes and attributes, namely, link impedance, turn impedance, one way, flyovers, underpasses, etc. In the shortest path analysis, the system finds the optimal path with the lowest impedance from an origin to destination passing through desired stops. It is very useful to get solution in traveling salesman problem. A well-known algorithm is Dijkstra's algorithm. Network analysis is used to identify the closest facility (e.g., a post office) from a given location. It also helps to allocate the part of the network to a closest facility center in terms of time and distance. Service area mapping is one of the well-known utilities in network analysis. In location-allocation analysis, optimal distribution of resources from warehouses to demand nodes can be mapped minimizing the travel distance. It is a powerful tool in transportation planning and management.

Statistical Analysis

As geospatial science and technology revolves around spatial and attribute data, many statistical measures help in analyzing the data and highlight the measures in visualization. Descriptive statistics uses the data, which are in interval and ratio scales and calculates mean, median, mode, standard

deviation, skewness, and kurtosis. As spatial objects have significant impacts on their neighborhoods, spatial autocorrelation is another important parameter that helps to measure the dependence on nearby values and bring out spatial patterns.

Trend surface analysis helps to understand the spatial trend locally or regionally for various properties such as rainfall, elevation, income distribution, infrastructure availability, etc. The trend surface can be linear or quadratic or third degree surface generated using the input location data.

Spatial interaction occurs due to movement over space and it can be population migration, commodity flows, etc. Gravity models are the most widely used types of spatial interaction models (Lo and Yeung 2002). If two cities have population p_1 and p_2 and they are separated by a distance d , then the spatial interaction can be expressed as the product of population divided by the distance ($p_1 \times p_2/d$). The distance can be geometrical distance or time distance or social distance. With population, exponents can be added in terms of income, etc. To build decision support systems, external statistical models get integrated with geospatial tools to provide solutions to the decision makers.

Visualization of Data Analysis

The results and outcomes of spatial analysis have to be expressed in such a form that the end users can easily understand and get the benefits. Spatial data visualization is an important step and maps are the most effective medium of communication. The map elements are title, body, legend, scale, north arrow, projection name, data sources, etc. Cartography is the subject that gives the methods on preparation of maps.

In vector data visualization, varied symbols can be used for point features, different width of lines for line features (e.g., thin blue line for a tributary and a thick blue line for a river) and varied fill patterns to show area features. The visual variables for a map display are size, shape, texture, fill pattern, hue, saturation, and intensity. Generally, elevation data is shown on 3-D surface and draping a thematic feature over it enhances the visualization. Color maps are preferred over black and white maps. The properties of hue, saturation, and intensity help to improve the color maps presentations. In the case of raster maps as grid is the basic unit, different colors are given to the grids to depict the features.

Maps can be categorized based on the usage. A topographic map is a general purpose map giving information on administrative units, drainage, elevation contours, land use, settlement areas, etc. A thematic map depicts a theme (land use, soils, etc.). A dot map shows quantitatively a property such as population where a dot may indicate one million. Statistical maps use pie and bar diagrams to quantitatively show more than one property (e.g., number of males

and females in a state in a country map). Isarithmic maps show isolines where an isoline connects points of equal value. While preparing the layout, the map body is placed at the center and other map elements have to be placed in such a way that the map looks balanced. A properly designed map can communicate the results well to the end user.

Spatial Decision Support Systems

A spatial decision support system (SDSS) provides a framework for integrating spatial database with geospatial analysis tools, statistical packages, and well-developed economic and environmental models. A SDSS tries to offer modeling, optimization, and simulation features for generating, evaluating, and recommending alternate strategies in decision-making task. The process in SDSS involves iterative, integrative, and participative steps. There cannot be a single SDSS to solve all problems. For each specific application, a SDSS must be designed as per the requirements. To build a SDSS, the initial step is to clearly define the problem, decompose the problem into subunits, and identify suitable models and algorithms to solve the subunits. The system design must have the capability to acquire knowledge from the end user, support multiple representations, and provide better visualization of the results in tables, maps, and through virtual navigation.

Spatial Data Accuracy Assessment

One of the main research issues in geospatial information science is accuracy. It includes several parameters such as original data quality, error, uncertainty, scale, resolution, data input methods, analysis techniques, presentation, interpretation of the results, etc. One cannot make an assumption that the input data are fully perfect. The results must give some indicators on data quality and an estimate of confidence about the results. Generally, the data are collected from different sources. In addition, vast amount of data are downloaded using the internet without knowing the quality at the production level. Even though the input spatial data are inaccurate to some degree, they are represented in high precision digitally in the computers.

The components of data quality are positional accuracy, attribute accuracy, geographical extent of dataset, completeness of data, the data source documentation, classification and sampling methods, boundary definitions, lineage, etc. Positional accuracy is defined as the closeness of location information to the true position on the ground. If a line width is 0.5 mm on a 1:50,000 scale printed map, it is

equivalent to 25 meters on the ground. Knowing the errors on the source document, during geo-registration and digitization process, a root mean square value can estimate the positional accuracy. In most of the cases, attributes are time dependent. Attribute accuracy can be estimated using the time of collection of the source data. The extent of data set must cover the entire study area. Data completeness is controlled by the rules of selection, generalization, and scale. Use of larger scale maps help to improve the quality. Another factor is the source media. A paper map may shrink or swell due to atmospheric conditions, which may introduce distortion during geo-registration.

Spatial data may have classification errors. In raster data, classification is done based on grid and the variations within a grid are not shown. In vector data, the classification of a theme is user defined. In primary data collection, sampling on the ground may be random, systematic, or stratified sampling. Important data may get missed due to the sampling method used. In thematic maps such as soil maps, the boundaries between two categories cannot be clearly defined and there is always fuzziness in the boundary areas. With every data, additional information on the source, method of preparation, symbols used, area of coverage, time of the survey, etc., will help to assess the data quality. Lineage provides the history of the source data. Presently metadata standards are emphasized globally to be adopted along with spatial data sources. The data quality is directly proportional to the data cost. The inherent uncertainties present in the data may result in making flawed decisions. The provision of data quality in proper documentation is very essential for improving the accuracy in the analysis.

In addition to data quality, errors occur during scanning, digitization, cleaning, multilayer registration, error propagation during analysis, improper visualization, and misinterpretation of the outcomes. One of the important and challenging areas in geographical information science is the quantification of error at every step and indicates it in terms of percentage of accuracy in the results.

Applications

Development of geographical information science and the implementation of its outcome through geospatial technology are to solve the real-world problems. It includes simple spatial query to high-end modeling problems. Availability of more and cheaper data and the success of applications for socio-economic benefits have increased the number of users in many folds. Internet technology provides easy to use simple and low cost methods to access, query, and view the spatial data. Major areas of applications cover natural resources

mapping and management, transportation, utilities, health and human services, engineering areas, cartography and map publication, communications, disaster management, education, defense, business, and so on.

Under natural resources sector, GIS is applied in agriculture, hydrology, land use planning, forestry, landscape conservation, archaeology, caves mapping, environmental management, marine and coastal mapping, mining and earth sciences, petroleum, etc. In forestry, geospatial technology helps to monitor forest growth, forest depletion, forest fire, reforestation, harvest scheduling, and so on. In the area of mining and mineral exploration, it helps to integrate data from geology, geophysics, hyperspectral airborne images, field surveys to delineate the mineral zones qualitatively, and estimate the ore content quantitatively. It empowers us to study environmental problems and impact assessment to protect the natural resources.

In transportation, geospatial technology is used for varied purposes like modeling travel demand, planning the road network, monitor the real-time traffic, management of transportation infrastructure, maintaining the roadway assets, etc. Routing and scheduling for on-time delivery and services is another important application. It is a powerful tool in utility management for utilities such as water supply, sewerage, electricity, telephone, etc.

In the government sector, geospatial technology helps in the areas of economic development, elections, security, law enforcement, crime mapping, public safety, disaster mitigation and management, land records and cadastral solutions, urban and regional planning, and sustainable development. It supports all aspects of disaster management, namely, planning, response and recovery, coordination with multiple agencies, and protection of the resources. It plays a key role to establish the principles of sustainable development. In the health sector, it is a vital tool for scientists and public health officials to investigate the cause and spread of diseases around the world. In the business sector, a host of applications are available including selecting the best sites, profiling customers, analyzing market areas, updating and managing assets and location-based services to users. It enables telecommunication professionals to integrate location-based data into analysis and management processes in network planning and operations.

In the area of defense, geospatial technology helps in mapping, base operations and facility management, force protection and security, command and control, intelligence, surveillance and reconnaissance systems, logistics, mission planning, modeling, simulation and training, terrain analysis, visualization, etc.

Research Challenges in Geographical Information Science

Geospatial technology is an outcome of the research in Geographical Information Science by academia and coordinated efforts of government organizations and industries with academia. This technology has witnessed a spectacular growth in the last four decades. As an offshoot from academic research, it has become a core technology for information resources management and decision making in government and businesses. Opportunities are growing in hardware and software development, research areas, and practical implementations. With the introduction of Web, it has entered in the homes of millions and hands of billions.

There are several issues that need our attention. Basically, geographical information is inherently complex in nature. People want to use the technology for wide range of applications and the scope is increasing. In the technological front, there is a rapid advancement. Huge amount of spatial data are getting generated using various technologies. There is a need for well-trained manpower to handle the implementations.

Vast amount of data are available through government sources and commercial suppliers. Public participation generates Volunteered geographical information (VGI). Multi-source data integration is a complex task as the data may be in incompatible formats with varied cartographic specifications such as scale, map projection, accuracy, and symbology. Metadata standards are getting developed globally. Sharing geographical data through formal standards is yet to become reality. Data ownership, copyright, and cost recovery discourage users from using the existing data. Also when the data are updated or modified or augmented, the ownership becomes an issue. Next comes the data security. A balance act is needed between the public's right to know and the individual's right to maintain privacy. The inherent uncertainties in the data may increase the possibility of getting wrong outcomes. Handling geometric incompatibility among multidata sets due to scale, projections, datum, etc., is a challenge. During mosaicing nearby maps, mismatch of features across boundaries may occur. The datasets need updation at regular intervals, which is a tedious and time-consuming task.

In the context of big data, the spatial data has 3Vs, namely, variety, volume, and valuable. New data models and new computational algorithms have to be developed to handle these semi-structures and unstructured data. The big data must be semantically annotated to make it understandable by both humans and machines. The machine learning techniques and process models have to interact with each other. There is a need for parallel/stream/cloud/grid computing engines for real-time analysis. New database structures,

search engines, and visualization techniques have to be developed. Modeling uncertainty and ambiguity in big data is a research challenge.

Research Areas

Study of geographical information both in systems and science perspectives have opened new research opportunities in computer science, computational geometry, spatial statistics, geography, etc. Technologies such as LIDAR, drones, high resolution satellites, and scanners produce volumes of data with high accuracy and speed. Integration of these datasets collected by different technologies in different accuracies and formats is a big challenge. Attempts are being made to develop techniques to extract information from these large set of spatial data using spatial data mining and knowledge discovery.

The focus on GIScience needs to shift from representation and analysis of the form of the Earth's surface to a much stronger concern for the processes that define its dynamics. (Goodchild 2004). Research is needed to develop representation of geographical processes independent of geometry but based on human cognition. It has to be on spatial concepts, categories, processes, and their interactions. Spatial computing algorithms have to be developed with input from Mathematics, Statistics, Computer Science, and Geography. A spatiotemporal process can be very complex to model and the complexity can be simplified by breaking the process into several components (Singh and Venkatachalam 2013). In space-time modeling, several approaches are being developed to model space time data, namely, physical modeling, space-time geostatistical modeling, time series of spatial processes, spatial processing of time series, and hierarchical modeling approach.

In spatial statistics, statistical methods must be developed to compute and analyze the data in spherical coordinates. Algorithms for space time interpolation of large datasets are needed. Use of statistical methods may help to quantify positional uncertainty, fuzzy boundaries, and accuracy of predictions. Present data structures and algorithms have to be modified to speed up computations.

One of the important research challenges is spatial data security. Very little attention has been given to address security concerns such as access control, security, and privacy policies at storage and dissemination level. Issues on research are protection of the ownership of geospatial data at dissemination; task-based and feature-based control at storage level; and secure, efficient, and computationally inexpensive geospatial data outsourcing. One way for data protection is watermarking. The requirements are watermarks that have to be invisible, robust during geometric transformations, cropping, format change, etc. One of the important

characteristics of geospatial vector data is its positional accuracy (Zope et al. 2015). The embedding watermark in vector data should preserve positional accuracy and topological relationships. In raster data, it must be near lossless and maintain the classification. It is better to combine watermarking with cryptography for better protection. During storage, there has to be control-based access with specifications and enforcement of policies. While outsourcing the data, efficient transformations can be followed such that the authorized users can get the results using keys received from the original source.

Open research areas in geographical information science include development of multidimensional GIS, spatiotemporal GIS, modeling of spatiotemporal phenomena, representation of variable resolution data, parallel/distributed/grid/cloud computing in geospatial domain, prediction of moving objects, estimation of uncertainty and modeling of error propagation in analysis, and improved visualization techniques.

Web technology has changed everything and GIS is no exception. WebGIS is the combination of Web and GIS (Fu 2018). WebGIS has grown during the last two decades. WebGIS has unlocked the power of GIS and has reached millions of users. Web technology with read and write has resulted in abundance of user generated content. Web services with cloud help to build new applications. Databases are getting generated with public participation (VGI) and there are several web mapping applications such as Google Maps, Google Earth, Microsoft Bing maps, Yahoo Maps, etc. OGC has developed open standards for geospatial web services. WebGIS is a powerful tool for E-Governance and a huge facilitator of public services. It can provide broad geospatial intelligence to decision makers with its rich source of real-time datasets. It provides a new business model with advertisement-based web mapping, location-based services, etc. It is one of the easiest tools to teach spatial literacy in schools and in Neogeography.

There are several open areas of research in WebGIS. With large amounts of VGI, the research has to focus on quality, integrity, accuracy, and personal privacy. Geotagging can enrich the spatial data and can be used in data mining. Geoparsing can help in disaster management and military intelligence. Geotargeting will be useful in e-commerce and tourism. Sensor web can help in near real-time analysis. Cloud-based GIS can provide quicker development time, mobility, reduced cost, and dynamically scalable databases. Semantic web is trying to add meaning to every information on the web so that the web data can be understood and processed by machines.

Conclusions

This entry discusses briefly the various components of Geographical information science and the open areas of research.

Geospatial technology and Computers can synthesize data and perform analysis and modeling. But, people make decisions. Users are from diverse technical background. There is a huge shortage of skilled specialists who can design the systems and develop applications. Sound GIScience education is needed to provide solid formation of the theoretical aspects including mathematics, statistics, computer science, geography, etc. In the academic front, GISc has widened its scope and research getting evolved on new concepts, model, geographical computing, and application. Challenges and the solutions in Geographical Information Science will enhance the capabilities of geospatial technology to provide massive benefit to the society.

Cross-References

- [Geoinformatics](#)
- [Geostatistics](#)
- [Interpolation](#)
- [Spatial Analysis](#)
- [Spatiotemporal Analysis](#)
- [Topology in Geosciences](#)
- [Uncertainty Quantification](#)

Bibliography

- Chang K (2009) Introduction to geographic information systems. McGraw-Hill Inc., New York
- Egenhofer MJ, Franzosa RD (1991) Point-set topological spatial relations. *Int J Geogr Inf Syst* 5(2):161–174
- Fu P (2018) Getting to know WebGIS, 3rd edn. ESRI Press, Redlands
- Goodchild MF (2004) GIScience: geography, form and process. *Ann Assoc Am Geogr* 94(4):709–714
- Laurini R, Thompson D (1994) Fundamentals of spatial information systems. Academic, London
- Lo CP, Yeung AKW (2002) Concepts and techniques of geographic information systems. Printice-Hall Inc., Upper Saddle River
- Singh MK, Venkatachalam P (2013) Efficient implementation of hierarchical model to study space time variability of latent heat flux. In: Computer science and its applications- ICCSA Part IV. Lecture notes in computer science LNCS, vol 7974. Springer, Berlin, Heidelberg, pp 33–44
- Venkatachalam P (2009) Geographical information system (GIS) as a tool for development. Information technology and communications resources for sustainable development, Encyclopadia of life support systems (EOLSS). <http://www.eolss.net/Sample-Chapters/C15/E1-25-01-12.pdf>
- Viana CM, Abrantes P, Rocha J (2019) Introductory chapter: geographic information systems and science. INTECH. <https://doi.org/10.5772/intechopen.86121>
- Zope SC, Venkatachalam P, Buddhiraju KM (2015) Assessment of distortion in watermarked geospatial vector data using different wavelets. *Geo-spatial Inf Sci* 18(2–3):124–133

Geographically Weighted Regression

Xiang Que^{1,2} and Shaoqiang Su^{1,3}

¹Computer and Information College, Fujian Agriculture and Forestry University, Fuzhou, Fujian, China

²Department of Computer Science, University of Idaho, Moscow, ID, USA

³Key Laboratory of Fujian Universities for Ecology and Resource Statistics, Fuzhou, China

Definition

Geographically weighted regression (GWR) (Brunsdon et al. 1996; Fotheringham et al. 2015, 2003) is a technique that extends the traditional regression framework by allowing different linear relationships to exist at various locations in space. It is a local method for analyzing *spatial heterogeneities* of processes and is widely applied in geography and other related disciplines. The underlying idea of GWR is that each observed point in the study areas borrows a certain number of its nearest neighbor points to build pointwise regression models.

Introduction

The general global linear regression model commonly assumes that the same relationships hold throughout the entire study area (i.e., constant across space). It can build the relationship between a dependent variable and one or more independent variables, whose general form is as follows:

$$y_i = \beta_0 + \sum_{k=1}^n \beta_k x_{ik} + \varepsilon_i \quad \text{for } i = 1, 2, \dots, n \quad (1)$$

In Eq. 1, x_{ik} is the k th independent variable measured at a location i , and y_i is the corresponding dependent variable. ε_i denotes the random error term, which is assumed to be independent and drawn identically from the normal distribution $N(0, \sigma^2)$, i.e., β_0, β_k are coefficients which are usually to be estimated by minimizing the value of residual sum of squares (RSS), i.e., $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ ($\hat{y}_i$ denotes the predicted value for the i th observed location). Such a global model is usually fitted by the method of **ordinary least squares (OLS)**. The coefficients can be estimated by OLS estimator take the matrix form:

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (2)$$

In Eq. 2, $\hat{\beta}$ denotes the vector of estimated coefficients, X is the matrix which contains the independent variables and a

constant column of 1. X^T is the transpose of X , and $(X^T X)^{-1}$ is the inverse of the **variance-covariance matrix**.

In the geospatial context, it is not always the case with some assumptions underlying the general global linear regression model. As Tobler's first law of geography goes "Everything is related to everything else, but near things are more related than distant things" (Tobler 1970), both the errors and the variables in the model might exhibit spatial correlations. Observed spatial variables will usually have spatial effects, which are caused by their spatial correlation structures. Two broad classes of spatial effects may be distinguished, referred to as spatial dependence and **spatial heterogeneity** (Anselin 1988). Spatial error or spatial lag models are appropriate when there appears to be spatial dependence structures in the residual term or the variables. It is assumed that relationships being modelled are the same everywhere in the study area in general global linear regression. However, there is often appropriate that the spatial processes might vary across space, which is referred to as spatial heterogeneity (Charlton et al. 2009).

Model Formulation

Geographically weighted regression (GWR) relaxes the assumptions of constant coefficients across space, that is, allows the relationships between the dependent variable and one or more independent variables to vary with their spatial locations. It is a useful tool for modelling spatially heterogeneous processes whose general formulation can be defined as:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^n \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \quad (3)$$

for $i = 1, 2, \dots, n$

In Eq. 3, the **coefficients** β_0, β_k vary according to their location i in the study area, which are spatially varying coefficients (SVC) that quantifies variations in the processes and the relationships between spatial independent and dependent variables. These coefficients may be estimated anywhere given a dependent variable and a set of independent variables observed at places whose locations are known. Its estimator, which is based on the **weighted least squares (WLS)**, conditioned on the coordinates (u_i, v_i) of location i , which takes the form:

$$\hat{\beta}_k(u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i) y \quad (4)$$

In Eq. 4, $W(u_i, v_i)$ denotes the weighted matrix whose off-diagonal elements are 0. The weights in the matrix are generally generated by a distance decay function that is known as a kernel function.

$$w(i) = \begin{bmatrix} w_1(u_i, v_i) & 0 & \cdot & \cdot & 0 \\ 0 & w_2(u_i, v_i) & \cdot & \cdot & 0 \\ \cdot & \cdot & w_3(u_i, v_i) & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & w_n(u_i, v_i) \end{bmatrix} \quad (5)$$

Gaussian and **Bi-square** kernel are two widely used kernel functions in the GWR-based model, which are given in Eqs. 6 and 7, respectively.

$$\text{Gaussian : } w_{ijS} = \exp \left[-\frac{1}{2} \left(\frac{d_{sij}}{b_S} \right)^2 \right] \quad (6)$$

$$\text{Bi-square : } w_{ijS} = \begin{cases} \left[1 - \left(\frac{d_{sij}}{b_S} \right)^2 \right]^2, & \text{if } d_{sij} < b_S \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

In Eqs. 6 and 7, d_{sij} is the spatial (Euclidean) distance between the regression point i and an observed data point j . b_S is a quantity known as the spatial bandwidth, which is in the same units as the coordinates used in the dataset. There two different weighting strategies of using the bandwidth for calculating the weighted matrix in GWR: a fixed spatial bandwidth strategy and an adaptive spatial bandwidth strategy. The former uses a specified spatial bandwidth for calculating weights, which is usually based on sufficient knowledge of the problem. In contrast, the latter uses a searching strategy to obtain the optimized spatial bandwidth through model diagnostic values.

Model Calibration

The GWR, which combine the WLS estimator, kernel and bandwidth, is regarded as a local model for exploring spatial non-stationarity. Several diagnostic values are used for the measurement of goodness of fit, such as the corrected **Akaike Information Criterion (AICc)** (Hurvich et al. 1998) or **cross-validation (CV)** score (Cleveland 1979). Their calculating formulations are given in Eqs. 8 and 9, respectively.

$$AIC_c = 2n \ln(\hat{\sigma}) + n \ln(2\pi) + n \left\{ \frac{n + \text{tr}(S)}{n - 2 - \text{tr}(S)} \right\} \quad (8)$$

In Eq. 8, n is the sample size, $\hat{\sigma}$ is the estimated standard deviation of the error term, and $\text{tr}(S)$ denotes the trace (the sum of the values in the leading diagonal) of the hat matrix S (Hoaglin and Welsch 1978). The hat matrix S is an interesting matrix in regression modelling. We will obtain the vector of predicted dependent variable $\hat{y}$ when the observed values of y are pre-multiplied by S , i.e., $\hat{y} = Sy$. In a GWR model, the effective number of parameters, which depends on the

bandwidth and the number of independent variables, can be obtained by the expression $2tr(S) - tr(S^T S)$.

$$CV = \sum_{i=1}^n [y_i - \hat{y}_{\neq i}(b)]^2 \quad (9)$$

In Eq. 9, $\hat{y}_{\neq i}(b)$ is the fitted value of y_i with the observed data for location i omitted from the calibration procedure. This approach calibrate model only on samples near to location i and not at i itself, countering the “wrap-around” effect.

In the calibration of the GWR model, we usually use a searching strategy (e.g., Golden-section search) to obtain the optimized spatial bandwidth and then calculate the weighted matrix. Thus, the coefficients $\hat{\beta}_k$ can be estimated by Eq. 4.

Model Output

GWR model can output a series of location-specific estimated coefficient values that can help users analyze the spatial non-stationarity. Besides, the model can also provide the estimated local standard errors, calculated local goodness of fit, local t statistics, local p -values, and so on. Out-of-sample points can also be predicted by the GWR model, given the values of the independent variable at the corresponding position. GWR model can output a series of location-specific estimated coefficient values that can help users analyze the spatial non-stationarity. Besides, the model can also provide the estimated local standard errors, calculate local goodness of fit, perform inferences based on local t statistics, evaluate the significance of spatial changes, and so on. The GWR model can also be used to predict out-of-sample points, given the values of the independent variable at the corresponding position.

Case Study

We selected hourly weather data and the elevation as independent variables to analysis their relationships with the value of **air quality index** (AQI) of 16 monitoring stations in Beijing, China. We build the **GWR** model as the formation below:

$$\begin{aligned} AQI_i = & \beta_0(u_i, v_i) + \beta_1(u_i, v_i)TEM_i + \beta_2(u_i, v_i)HPA_i \\ & + \beta_3(u_i, v_i)WET_i + \beta_4(u_i, v_i)WS_i + \beta_5(u_i, v_i)WD_i \\ & + \beta_6(u_i, v_i)ELE_i + \varepsilon_i \end{aligned} \quad (10)$$

where AQI is the dependent variable, TEM denotes the air temperature ($^{\circ}\text{C}$). HPA is the air pressure (hPa), WET is the relative humidity (%). WS is the wind speed (m/s), and the WD is degrees of wind direction (start from counterclockwise east). ELE is the elevation.

The hourly AQI data in April 30, 2016, source from the website <http://beijingair.sinaapp.com>, and the corresponding weather data source from the Beijing Municipal Data Resource Network (<http://106.38.57.109/>). Table 1 shows some fitting results of fitted GWR model. The residual sum of squares (RSS) is only 764.78, which is only 6.3% of the OLS. And both the $AICc$ and R-Squared (R^2) are significantly better than the OLS model. The results of OLS shows that it performs very poorly in the study area because it is trying to fit a global model to a nonstationary process. The means of estimated coefficient values by GWR are also different from the OLS estimated ones.

In this case, we use the inverse distance weighted (IDW) interpolation method to generate the surface of each independent variable. We then use the fitted GWR model to predict the spatial surface of AQI (as shown in Fig. 1). Through this method, we can obtain the estimated spatial surface of the dependent variable (as well as the predicted coefficient surfaces), which is useful for analyzing the heterogeneity of the spatial process.

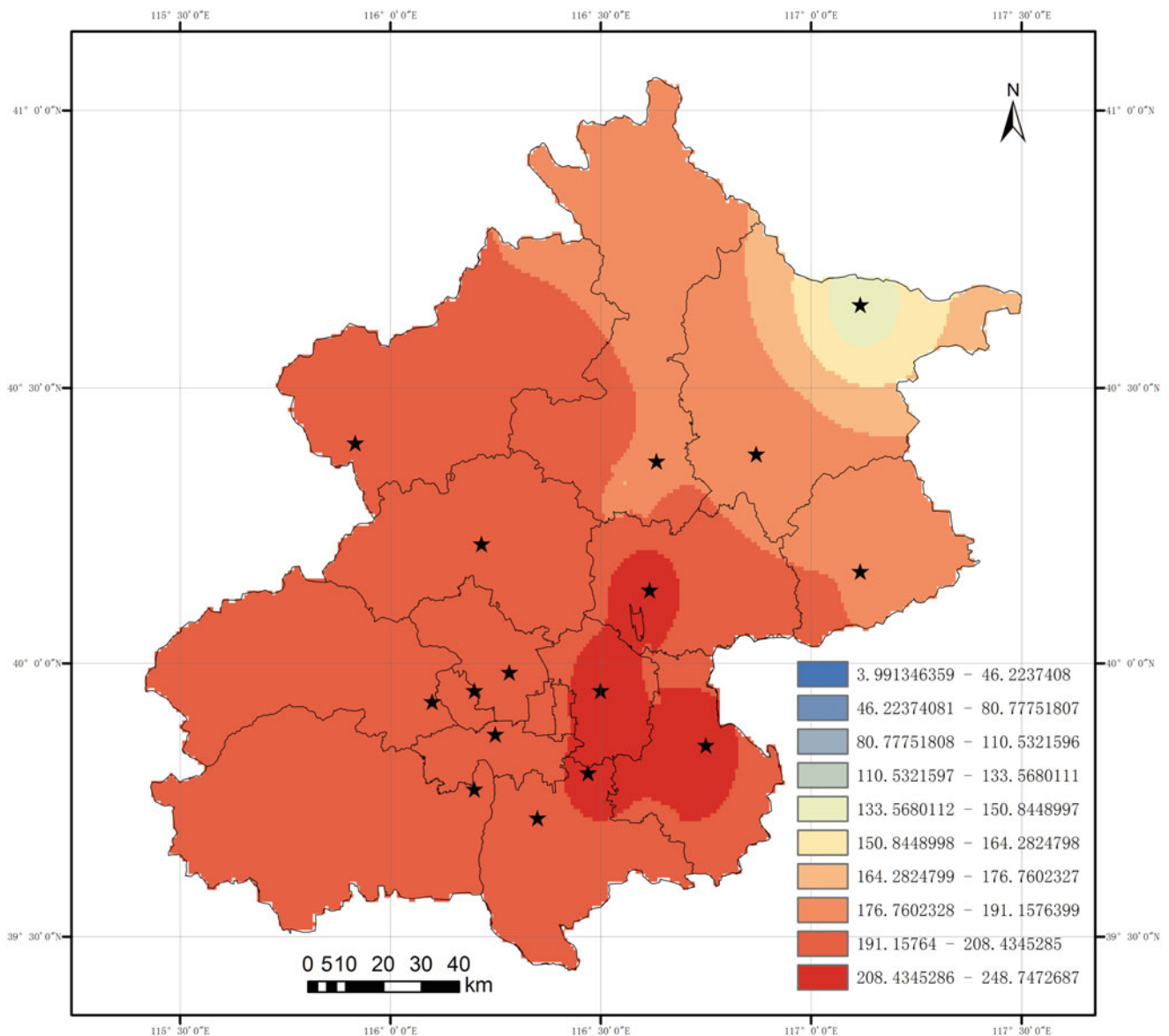
Extensions to GWR

The GWR model is implemented in many open-source packages and software, such as R package “spgwr,” python package “MGWR,” and GWR 4.0. Therefore, GWR is explicitly applied in many spatial-related fields, including mineral potential modelling (Wang et al. 2015), health risks (Li et al. 2017), climate variation (Hu and Xu 2019), and so on. However, this model is not appropriate for all types of spatial data. There are many variants of the GWR model to deal with various spatial heterogeneity issues. For examples, if the dependent variable is count data, a **Poisson** extension of GWR model should be used for analyzing. If the dependent variable is binary, a **logistic** extended GWR model is needed (Atkinson et al. 2003). And there are also temporal extensions of GWR models, such as the **geographical and temporal weighted regression (GTWR)** (Fotheringham et al. 2015) and **spatiotemporal weighted regression (STWR)** (Que et al. 2020, 2021).

Geographically Weighted Regression, Table 1 Modelling results of case study

Models	RSS	$AICc$	R^2	$Est \beta_0$	$Est \beta_1$	$Est \beta_2$	$Est \beta_3$	$Est \beta_4$	$Est \beta_5$	$Est \beta_6$
OLS	12133.09	165.50	0.31	18381.63	23.791	-18.80	3.83	-15.82	-0.05	-1.94
GWR	764.78	137.83	0.96	7497.01	11.81	-7.59	1.63	-16.63	-0.07	-0.25

Note: $Est \beta_0$, $Est \beta_1$, $Est \beta_2$, $Est \beta_3$, $Est \beta_4$, $Est \beta_5$, $Est \beta_6$ of GWR are the mean of the estimated coefficient of const, TEM , HPA , WET , WS , WD , and ELE , respectively



Geographically Weighted Regression, Fig. 1 Predicted AQI surface of GWR model

Conclusions

GWR is a widely used statistical model, which can be applied to explore the spatial heterogeneities. It can build pointwise regression models by borrowing points from spatial neighbors. This model can generate the local regression coefficients, which can help reveal the strength of the spatial effect of each independent variable to the dependent variable at some places.

Cross-References

- [Regression](#)
- [Spatiotemporal Weighted Regression](#)

Bibliography

- Anselin L (1988) Spatial econometrics: methods and models. Kluwer Academic, Dordrecht
- Atkinson PM, German SE, Sear DA, Clark MJ (2003) Exploring the relations between riverbank erosion and geomorphological controls using geographically weighted logistic regression. *Geogr Anal* 35(1): 58–82
- Brunsdon C, Fotheringham AS, Charlton ME (1996) Geographically weighted regression: a method for exploring spatial nonstationarity. *Geogr Anal* 28(4):281–298. <https://doi.org/10.1111/1467-9884.00145>
- Charlton M, Fotheringham S, Brunsdon C (2009) Geographically weighted regression. White paper. National Centre for Geocomputation. National University of Ireland Maynooth
- Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc* 74:829–836. <https://doi.org/10.1080/01621459.1979.10481038>

- Fotheringham AS, Brunson C, Charlton M (2003) Geographically weighted regression: the analysis of spatially varying relationships. Wiley, Chichester
- Fotheringham AS, Crespo R, Yao J (2015) Geographical and temporal weighted regression (GTWR). *Geogr Anal* 47(4):431–452
- Hoaglin DC, Welsch RE (1978) The hat matrix in regression and ANOVA. *Am Stat* 32:17–22. <https://doi.org/10.1080/00031305.1978.10479237>
- Hu X, Xu H (2019) Spatial variability of urban climate in response to quantitative trait of land cover based on GWR model. *Environ Monit Assess* 191(3):194
- Hurvich CM, Simonoff JS, Tsai CL (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J R Stat Soc Ser B Stat Method* 60:271–293. <https://doi.org/10.1111/1467-9868.00125>
- Li Y, Chen Z, Li J (2017) How many people died due to PM_{2.5} and where the mortality risks increased? A case study in Beijing. In: 2017 IEEE international geoscience and remote sensing symposium (IGARSS). IEEE, pp 485–488. <https://doi.org/10.1109/IGARSS.2017.8126997>
- Que X, Ma X, Ma C, Chen Q (2020) A spatiotemporal weighted regression model (STWR v1.0) for analyzing local nonstationarity in space and time. *Geosci Model Dev* 13(12):6149–6164. <https://doi.org/10.5194/gmd-13-6149-2020>
- Que X, Ma C, Ma X, Chen Q (2021) Parallel computing for fast spatiotemporal weighted regression. *Comput Geosci* 150:104723. <https://doi.org/10.1016/j.cageo.2021.104723>
- Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. *Econ Geogr* 46:234–240
- Wang W, Zhao J, Cheng Q, Carranza EJM (2015) GIS-based mineral potential modeling by advanced spatial analytical methods in the southeastern Yunnan mineral district, China. *Ore Geol Rev* 71:735–748. <https://doi.org/10.1016/j.oregeorev.2013.08.005>

Geohydrology

Faramarz Doulati Ardejani^{1,2}, Behshad Jodeiri Shokri³, Soroush Maghsoudy^{1,2}, Majid Shahhosseini^{1,2}, Fojan Shafaei^{1,2}, Farzin Amirkhani Shiraz² and Andisheh Alimoradi⁴

¹School of Mining, College of Engineering, University of Tehran, Tehran, Iran

²Mine Environment and Hydrogeology Research Laboratory (MEHR Lab), University of Tehran, Tehran, Iran

³Department of Mining Engineering, Hamedan University of Technology, Hamedan, Iran

⁴Department of Mining Engineering, Imam Khomeini International University, Qazvin, Iran

Synonyms

Hydrology

Definition

Groundwater in a surface mining operation can create a number of water-related problems affecting the mines' design and

economic viability. Groundwater inflow from the surrounding strata toward the mining excavation requires installing large-scale dewatering facilities to lower the water table and create an extensive and prolonged cone of depression (Doulati Ardejani et al. 2003a). A hydrogeological investigation of the mine site is necessary to design an appropriate dewatering technique. The hydrogeological characterization program is required to be implemented at an early stage of the mine evaluation process, specifically to help quantify the probable pore pressure conditions in the pit slopes (Read et al. 2013). The collection of hydrogeological data such as hydraulic conductivity, storage coefficient, and pre-mining groundwater flow regime are the main factors in designing an appropriate dewatering technique. Sequential investigation levels for a hydrogeological project include an inventory of existing data, remote sensing interpretation, hydrogeological fieldwork, geophysical surveying, and exploratory drilling. The method presented in this chapter provides an appropriate framework to investigate hydrogeological conditions around an open-pit mine to control water-related problems.

Introduction

Hydrogeological characterization of a specific area can be achieved by estimating a set of aquifer parameters, such as aquifer thickness and extent, hydraulic conductivity, hydro-lithological units, and tectonic features (Keller and Pinter 1996). To accurately design a dewatering system in an open-pit mine, it is necessary to determine the exact amount of water that enters the pit. For this purpose, a hydrogeological investigation has to be conducted to properly identify the aquifer surrounding the mine site. The obligations of the mine hydrogeology group usually include (Doubek et al. 2013):

- Detailed characterization of the groundwater system
- Installing a monitoring or piezometer network in order to support continued hydrogeological characterization and to monitor the performance of dewatering and pit slope depressurization measures
- Installing horizontal drains
- Performing aquifer tests
- Locating suitable dewatering well sites by drilling and testing pilot holes
- Designing and supervising the installation of production dewatering wells
- Sizing pumping equipment
- Collecting data and maintaining a robust database of water levels, pumped volumes, flow rates, and water chemistry
- Reporting dewatering progress to mine management

The hydrogeological fieldwork must include the groundwater system's adequate characterization and hydraulic properties within the mine area and the surrounding region.

Open-pit mining operations below the groundwater table can result in extreme inflow to the mine workings and water-related severe problems affecting the mining operations' design and economic viability. In other words, unexpected inflow in large quantities can cause (Singh and Atkins 1984; Morton and van Mekerck 1993):

- Impeding production
- Delaying the project
- Reducing the stability of the strata and resulting in the pit slope stability problem
- Causing many environmental problems
- Forming a pit lake and causing a long-term water quality problem
- Losing access to all or part of the mine working area
- Increasing the use of explosives
- Rising the explosive failures resulting from wet blast holes
- Needing the use of special explosives
- Resulting in inefficient loading and hauling
- Creating unsafe working conditions

This inflow from the surrounding layers toward the excavation requires dewatering facilities to keep the mine workings dry, create an extensive and prolonged cone of depression, and increase the pit wall slope stability. Mine abandonment, groundwater rebound due to cessation of dewatering and associated pollution issues, is considered a severe mining problem throughout the world (Henton 1981). If the effects and magnitude of a water-related problem can be recognized in advance of mining, appropriate water management strategies to minimize the socioeconomic and environmental impact of mine dewatering often can be efficiently incorporated into the mine plan. To design an effective drainage scheme for a surface mine, the prediction of water inflow into a mining excavation is a crucial component of the operation's feasibility stage (Singh and Reed 1987).

Dewatering of open-pit mines during mining operations can place considerable hydrological stress on the regional groundwater flow system by creating an extensive and prolonged cone of depression, regional groundwater table lowering, overlapping cones of depression, land subsidence, and water quality deterioration, which can endanger mine productivity and human life (Keqiang et al. 2006). When mining is completed, and the dewatering operations have ceased, the water table or hydraulic head (in the confined aquifer) rebounds. The rising groundwater forms a pit lake. The groundwater in contact with the pit wall strata containing sulfide materials may be contaminated with pyrite-oxidation products. Hence, prediction of post-mining water table elevation within the pit is essential. The adverse effects can be countered with an appropriate addition of alkalinity to the contaminated water and special rehandling techniques. Many analytical and numerical models have been applied in the past

to estimate water inflow into surface mining excavations (Singh and Atkins 1984; Singh and Reed 1987; Hanna et al. 1994; Doulati Ardejani et al. 2003a; Bahrami et al. 2014). Comprehensive studies including analytical and numerical modeling and field observations into groundwater recovery after cessation of dewatering operation in mining excavations have been reported by many researchers (Reed and Singh 1986; Doulati Ardejani et al. 2003b).

In this chapter, the main steps for predicting groundwater inflow into an open-pit mine and post-mining groundwater rebound using numerical models, proposed dewatering system, and the associated costs of each stage study have been presented.

Open-Pit Mine Dewatering

A mine dewatering program's main aim is to lower the water table below the pit floor and remove residual groundwater seepage and surface water runoff inside the pit. According to Norton (1983), the dewatering process has the following effects:

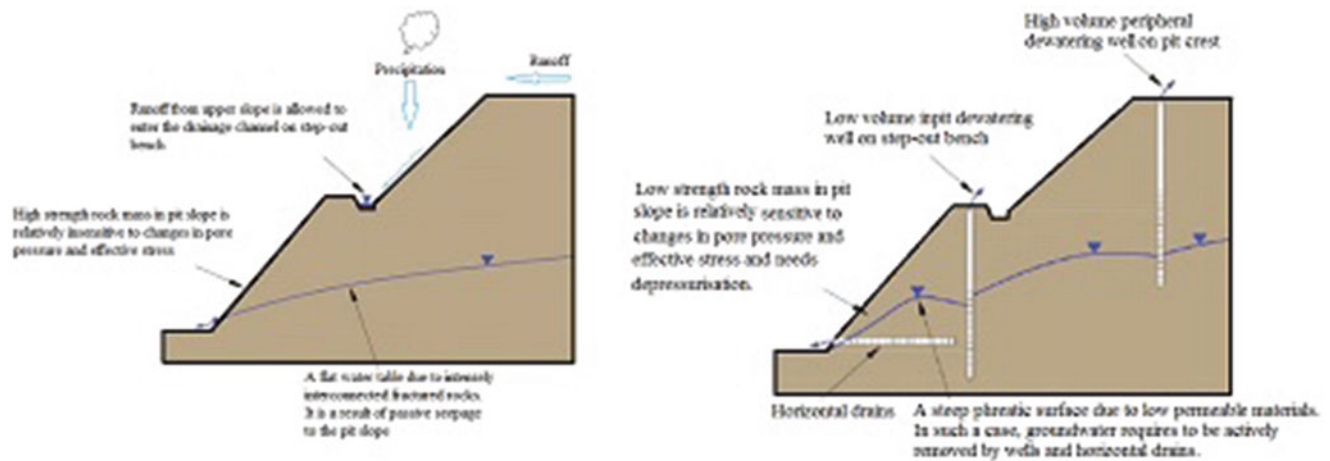
- Creates a cone of depression in the vicinity of the mining excavation
- Reduces the hydraulic gradient in the aquifers
- Increases the slope stability
- Controls mine excavation flooding

A hydrogeological investigation of the mine site is necessary to design an appropriate dewatering technique. The collection of hydrogeological data are the main factors in designing appropriate dewatering techniques. Depending on the pit's geology and hydrogeological condition, several groundwater control methods, including peripheral (active) and in-pit dewatering techniques, are proposed to conduct a successful mine dewatering operation. Active dewatering techniques (Norton 1983) are often used in surface mine sites where a large amount of groundwater inflow into the excavation is expected.

Dewatering Techniques

Passive Dewatering

As soon as the pit is progressed to below the water table, small quantities of natural drainage will usually occur on the pit wall, in response to seepage and potentially in response to the rock mass's mining-induced relaxation. This drainage results in a reduction in the pore pressures in the pit slopes. In passive dewatering, a sump is often excavated at the pit bottom, and the groundwater is allowed to enter the sump, where it is eventually pumped out. This method is recommended for



Geohydrology, Fig. 1 Schematic view of passive (left) and active (right) dewatering techniques (Read et al. 2013)

open pits where a small amount of groundwater is expected and for controlling surface runoff and direct precipitation (Norton 1983).

Active (Advance) Dewatering

In many open-pit mines, the pore pressure dissipation rate obtained by passive dewatering is too low to maintain the water table drawdown in advance of the mine excavation level or achieve the pit walls' target depressurization levels. In these cases, an active dewatering program can be implemented to reduce pore pressure in the wall and keep the pit faces relatively dry. In this method, electrical submersible pumps must be installed in vertical boreholes drilled within the geological layers inside and around the perimeter of the pit. However, horizontal drains can be considered an effective way to depressurize and drain aquifer in the west slope and similar situations that may occur during the next phases of mine development (Norton 1983). Figure 1 shows a schematic representation of dewatering techniques.

Groundwater Flow Model

Successful dewatering necessitates a hydrogeological assessment of the mine site. This may be achieved through numerical modeling. Models are extensively used to quantify the dewatering rates, design dewatering programs, and examine such programs' efficiency. To design an effective dewatering program for an open-pit mine, predicting the water inflow into the pit using an appropriate groundwater flow model is a significant task. A groundwater model is a computational technique representing an approximation of a groundwater system (Anderson and Woessner 1992). There are different types of groundwater models. The mathematical model is the

most common model used to simulate groundwater flow problems. A mathematical model attempts to simulate the flow system's actual behavior through the solution of mathematical equations (Schwartz et al. 1990). Groundwater flow models can simulate the hydraulic head distribution and groundwater flow rates within and across the systems' boundaries under consideration (Ünsal 2013). The use of numerical modeling codes such as SEEP/W (Geo-slope International 2014), MODFLOW (McDonald and Harbaugh 1988), and FEFLOW (Diersch 1992) enables us to address complicated groundwater problems associated with the mining operation.

Study Schedule

A comprehensive hydrogeological study is carried out in an open-pit mine site as the basis for the development of a proper dewatering plan to overcome water inflow. Table 1 summarizes the activities which are conducted concerning the mine hydrogeological investigation.

Conclusions

A variety of groundwater control techniques can be used to conduct a successful mine dewatering, depending on the geology, hydrogeological condition, and mine type. This allows pit dewatering and lowering the groundwater level adjacent to pit walls. According to the hydrogeological study result, the water inflow is a critical issue in an open-pit mine area, affecting the pit slope stability and causing a wet working environment for the mining operation. The passive dewatering from the lowest zone of a pit floor level is just a temporary solution.

Geohydrology, Table 1 A step-by-step investigative procedure of open-pit mine hydrogeology

No.	Task
1	Literature survey Literature survey of the mining operations and associated water inflow problems, inflow sources into open-pit mines, inflow types, and flow regime in aquifers Open-pit mine dewatering and pit slope depressurization Post-mining groundwater rebound and pit lake formation A review of numerical methods, governing flow equation, groundwater flow models Preliminary interpretation of the hydrological and rainfall data
2	Data collection/preliminary design Remote sensing, satellite images, and airborne magnetic data interpretation Hydrological data (rainfall, evaporation, temperature) Surface water (runoff, stream flow records) Hydrogeological features (springs) Data extraction from available maps (geology, ore geology, hydrogeology, topography, and structural features) Preparing a checklist for fieldwork, sampling, and water quality monitoring program Designing the horizontal drains for those segments of the pit wall where there is a risk of instability Supplementary interpretation and modeling the geophysical (IP/RS, magnetic) data Large- and local-scale watershed and sub-watershed analyses using digital elevation model and GIS tool Determining the general water flow direction Designing the new geophysical profiles and points
3	Field operation and surveying the study area Site visiting and primary investigation (double-check of designed geophysical profiles and points) Determining the locations and drilling open observation holes and measuring groundwater levels Groundwater level measurements in springs and the existing exploration boreholes Conducting a geophysical survey – resistivity and magnetic surveys Site detailed investigation for designing a conceptual hydrogeological model (investigating the hydrological, hydrogeological, environmental, and ecological characteristics of the study area) In situ measurements for water quality parameters (pH, temperature, EC, DO, ORP, and salinity) Hydro-geochemical sampling from available boreholes, open observation holes, springs, and seeps Soil sampling for permeability test around the pit and watershed surface soils Double-checking the geological features, lithology, faults, and discontinuities according to the structural map Inspecting the lithological logging during the drilling of exploration well, inspecting the horizontal drains and open observation holes Preliminary determination of groundwater inflow to the pit and inflow type Preliminary determining the rate of pit lake formation Determining and categorizing different seeps around the pit wall Performing a tracer test around the pit
4	Office works and data interpretation Geo-electrical data reduction, processing, preparing the resistivity pseudo-sections and interpretation Modeling the resistivity profiling measurements and producing 2D inverted resistivity sections Modeling the vertical electric sounding (VES) measurements and producing an inverted resistivity model Magnetic data reduction, processing, and enhancement Producing magnetic maps (total intensity, residual field, RTP, analytical signal, vertical derivatives, and upward continuation) and interpretation Producing a modified structural map of the mine area based on the magnetic data surface structural features Determining the hydraulic conductivity of pit wall rocks based on tracer test and hand auger experiment Generating a groundwater table level map from all available groundwater level data and determining groundwater flow direction Generating a groundwater table depth (depth to groundwater level) map Generating bedrock elevation and bedrock depth (depth to bedrock) maps Producing supergene zone map (related to porphyry copper mines) Preparing various cross sections of the aquifer representing geological units, pit layout, groundwater table, and bedrock level
4	Piezometric network and location design Monitoring the piezometers, continuous groundwater measurements, and data evaluation Time-based changes in groundwater levels within the mine site Compiling the hydrological data and deriving a hydrological model of mine catchment
5	Hydrogeochemical data analysis and interpretation (to support hydrogeological study and provide a preliminary estimate of potential environmental concerns) Major, trace, and rare earth element analysis Graphical representation of hydrogeochemical results and preparing water chemistry diagrams Plotting water chemistry maps Statistical analysis of water quality data Speciation and solubility modeling of the hydrogeochemical data Water quality assessment and evaluating potential health effects

(continued)

Geohydrology, Table 1 (continued)

No.	Task
6	Description for operating a drilling program Preparing a job description for drilling operation and pumping out test Installation of piezometers Drilling the testing wells
7	Developing a conceptual model Development of a conceptual model based on geological, topographic, structural, geophysical, hydrogeochemical, hydrological, hydrogeological, drilling and mine geometry, and development phases
8	Developing a mathematical model Inspecting the drilling operation, sampling, data collection, and geological logging Inspecting the pumping out test Investigating the aquifer system, boundaries, and zones containing water and determining the hydrodynamic coefficients Developing a groundwater inflow model during mine advancement Determining the height of the seepage face in excavation highwall as a significant parameter of the slope stability Developing a groundwater rebound model Developing a model for hydrogeological regime and hydrodynamic system of a surface mine area
9	Proposing a dewatering system Proposing a dewatering and drainage scheme for a mine area Preparing a job description for dewatering wells Drilling the dewatering wells and inspection Conducting dewatering operation

Apart from lowering the water table near the pit, pumping out operation causes insignificant change to groundwater levels in the surrounding aquifer. Advance dewatering must be carried out to overcome the water inflow issue to the pit. This dewatering technique will be an essential and permanent solution for the future mining operation. A general active dewatering program's primary focus is to lower the water table below the pit floor and remove residual groundwater seepage and surface water runoff inside the pit. It will also positively affect pit slope stability and control the water accumulation on the bench floors in the mine working faces. Runoff must also be controlled to allow effective mining operation. Runoff from the surrounding areas must be diverted away from the pit using collector drains and diversion bunds. Within the mine, surface water from direct precipitation and runoff must be controlled by the drains and sumps to pump water away from working areas. The water pumped out of the dewatering wells or the accumulated water in the water-collecting sumps must be disposed of using pumps and a proper pipeline system. Polyethylene pipes can be used for water transfer. The water collected from the pit should be transferred to a pond outside the pit. Centrifugal pumps are suitable for water removal. Either submersible pumps or pedestal pumps can be used for this purpose. Submersible pumps are put underwater, while pedestal pumps are positioned with the pump motor out of the water. Disposal of the pumped water can create problems, because the discharge water may have a high sediment load, elevated concentrations of iron, sulfate and toxic elements, and low pH which can cause environmental problems at the disposal point, requiring sediment removal (e.g., by passing through a sediment lagoon) and special water treatment techniques. It is

recommended to use a geomembrane liner with appropriate diameters in the external pond to prevent water infiltration into the ground. However, no need to install geomembrane liner for in-pit sumps. It is recommended to add alkaline materials such as limestone, lime, sodium carbonate, or sodium hydroxide to the in-pit sumps to neutralize the acidity and precipitate the trace elements. This in turn increases the quality of the water. Dewatering and mine water management are essential aspects of many surface mine operations. A planned and integrated approach toward water management, incorporating the overall mine development strategy, operational requirements, and environmental considerations, can be developed. Effective dewatering and mine water management strategies bring cost-saving benefits through managing risks and minimizing uncertainties.

It is necessary to develop innovative, cost-effective, and environmentally acceptable technologies, which will be applied on a field scale to prevent acidic drainage generation from hazardous mining waste dumps and neutralize acidic drainages. Treatment of the effluents will be attempted to prevent the transportation of the toxic elements and minimize their adverse effects on the surface- and groundwater systems and soil. Besides, the development of an integrated environmental management scheme, based on the result of a risk assessment analysis, will produce a novel methodology for the proper disposal of the hazardous wastes associated with the future mining development and the application of preventative and remedial technologies for the protection of soil and water quality.

Bibliography

- Anderson MP, Woessner WW (1992) Applied groundwater modeling: simulation of flow and advective transport. Academic Press, San Diego
- Bahrami S, Doulati Ardejani F, Aslani S, Baafi E (2014) Numerical modelling of the groundwater inflow to an advancing open pit mine: Kolahdarvazeh pit, Central Iran. *Environ Monit Assess* 186: 8573–8585
- Diersch H-JG (1992) Interactive, graphics-based finite element simulation of groundwater contamination processes. *Adv Eng Softw*. http://www.feflow.info/uploads/media/fefflow_reference_manual.pdf
- Doubek G, Creighton A, Dowling J, Price M, Hawley M (2013) Chapter 2: Site characterisation. In: Beale G, Read J (eds) *Guidelines for evaluating water in pit slope stability*. CSIRO Publishing, Collingwood, pp 65–152
- Doulati Ardejani F, Singh RN, Baafi EY, Porter I (2003a) A finite element model to: 1. Predict groundwater inflow to surface mining excavations. *Mine Water Environ* 22(1):31–38
- Doulati Ardejani F, Singh RN, Baafi EY, Porter I (2003b) A finite element model to: 2. Simulate groundwater rebound problems in backfilled open cut mines. *Mine Water Environ* 22(1):39–44
- Geo-slope International, Ltd (2014) SEEP/W for finite element seepage analysis. <http://www.geo-slope.com/products/seepw.asp>
- Hanna TM, Azrag EA, Atkinson LC (1994) Use of an analytical solution for preliminary estimates of groundwater inflow to a pit. *Min Eng* 46(2):149–152
- Henton MP (1981) The problem of water table rebound after mining activity and its effect on ground and surface water quality. *Proc, Quality of groundwater, Noordwijkethout, The Netherlands*. vol 17. Elsevier, Burlington, pp 111–116
- Keller EA, Pinter N (1996) Active tectonics: earthquakes, uplift and landscape. Prentice Hall, Upper Saddle River
- Keqiang H, Dong G, Xianwei W (2006) Mechanism of the water invasion of Gaoyang Iron mine, China and its impacts on the mine groundwater environment. *J Environ Geol* 49:1163–1172
- McDonald MC, Harbaugh AW (1988) A modular three-dimensional finite-difference groundwater flow model: techniques of water resources investigations. US Geological Survey, Reston
- Morton KL, van Mekerck FA (1993) A phased approach to mine dewatering. *Mine Water Environ* 12:27–33
- Norton PJ (1983) A study of groundwater Control in British surface mining, PhD Thesis, Nottingham University, 460 pp
- Read J, Beale G, Ruest M, Robotham M (2013) Introduction. In: Beale G, Read J (eds) *Guidelines for evaluating water in pit slope stability*. CSIRO Publishing, Burlington, pp 19–64
- Reed SM, Singh RN (1986) Groundwater recovery problems associated with opencast mine backfills in the United Kingdom. *Int J Mine Water* 5(3):47–74
- Schwartz FW, Andrews CB, Freyberg DL, Kincaid CT, Konikow LF, McKee CR, McLaughlin DB, Mercer JW, Quinn EJ, Rao PSC, Ritmann BE, Runnells DD, Van der Heijde PKM, Walsh WJ (1990) Groundwater models. National Academy Press, Washington, DC
- Singh RN, Atkins AS (1984) Application of analytical solutions to simulate some mine inflow problems in underground coal mining. *Int J Mine Water* 3(4):1–27
- Singh RN, Reed SM (1987) Estimation of pumping requirements for a surface mining operation. *Proc, Symp on pumps and pumping*, Nottingham University, England, p 1–24
- Ünsal B (2013) Assessment of open pit dewatering requirements and pit lake formation for KIŞLADAĞ Gold mine, UŞAK-TURKEY, Ph.D. Thesis, The Graduate School of Natural and Applied Sciences of Middle East Technical University, 145 p

Geoinformatics

Jeffrey M. Yarus¹, Jordan M. Yarus² and Roger H. French¹

¹Material Sciences and Engineering, Case Western Reserve University, Cleveland, OH, USA

²Geoscience Lead Site Reliability Engineering Halliburton, Houston, TX, USA

Definition

Geoinformatics is the general term used to describe the process of information entry and retrieval to transform geoscience data and information into knowledge. First introduced in the early 1990s (Merriam 2004), geoinformatics implies the use of computing systems and comprises activities such as “data collection, digitization, management, and transmission of digital data” (Merriam 2004; Allaby 2013).

Introduction

The seeds of geoinformatics lie within the field of quantitative geoscience with an excellent historical review in Merriam and Merriam and Harworth (Merriam 2004; Merriam and Harworth 2004). In its early history, the geosciences were rooted in qualitative methods, focused on non-numeric, descriptive data. The early emphasis on qualitative analysis stems in part from the integral relationship between the geoscientist and fieldwork, observing and describing the various features on the Earth’s surface. This descriptive information is responsible for key theoretical underpinnings including the “Principle of Original Horizontality,” published in 1669 by Nicholas Steno (Allaby 2013), and the *Strata Identified by Organized Fossils*, published first by French zoologist Georges Cuvier in 1812, and expanded by William “Strata” Smith in 1815 (Gohau 1990). Both of these principles were based on qualitative analysis and serve as the foundation of modern stratigraphy.

Quantitative Geosciences: Evolving Computational Systems

The emergence of quantitative geoscience had a significant impact on the early evolution of the discipline, what Merriam (Merriam 2004) refers to as the “origins” and “formative” stages. Prior to the nineteenth century, graphs and charts may have been common, but statistical analyses using large amounts of non-numeric data were problematic. This was due in part to the labor necessary to gather and convert descriptive information into numbers for statistical

computations. The emergence of modern statistics did not occur until the end of the nineteenth century with Karl Pearson, credited with initiating the discipline of mathematical statistics (Cubitt and Henley 1978). While the use of statistics was uncommon, there were significant contributions made. Sir Charles Lyell (1797–1875), whose work contributed to the basis of modern geology, was one of the first to use statistics to establish the order in facies sequences. This was influenced by an impressive interdisciplinary group of colleagues including Charles Darwin, Charles Babbage, Michael Faraday, and other key players in the industrial revolution and the birth of modern science (Merriam 2004).

By the late nineteenth century through the mid-twentieth century, the *exploration* and *development* stages (Merriam 2004), early computer systems were introduced, and the use of quantitative analysis became more common. However, most geoscientists lacked access and/or the technical support required to use these computer systems. Further, they lacked sophisticated methods for data handling and management (e.g., converting the non-numeric data into digital form) (Agterberg 1974). A small group of geoscientists recognized the importance of computers, mathematics, and statistics. Among these players were John C. Griffiths (petrography and statistics), W.C. Krumbein (size-frequency distributions, computer-based stratigraphic facies analysis), L.L. Sloss (statistics and sequence stratigraphy), and Georges Matheron (geostatistics) (Merriam 2004). Joint contributions from other disciplines were also highly influential, an excellent example being Andrey Nikolaevich Kolmogorov. His classic work on “The solution of a probability problem related to the question of the formation of strata” influenced Matheron, Krumbein, and others and arguably planted the seeds of what is today referred to as “mathematical geology” (Shiryaev 1989).

Quantitative Geosciences: Early Computer Engagement

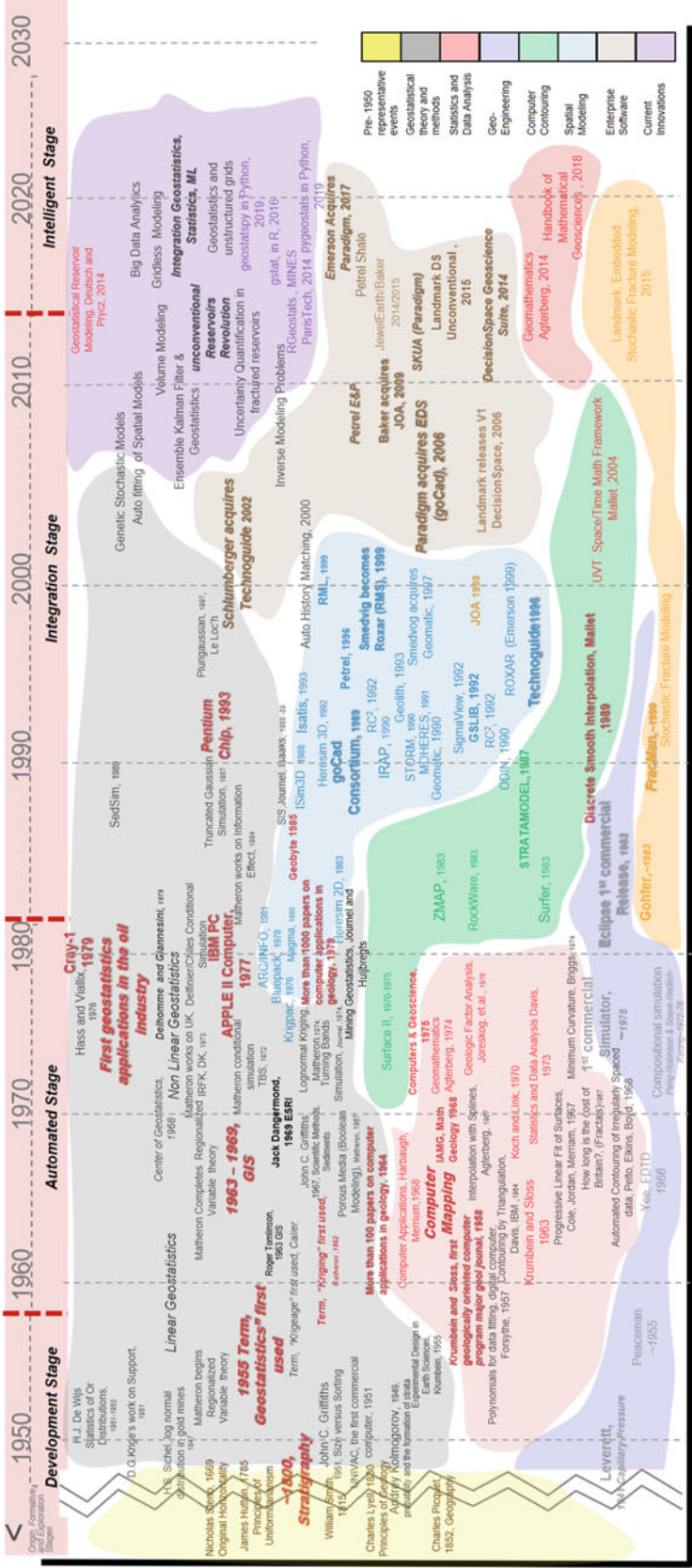
Merriam’s *automated stage* covers the period between 1958 and 1982 with the emergence of personal and desktop computers and the exponential growth in mathematical and statistical applications (Merriam 2004). Figure 1 represents a timeline of representative topics from Earth geoscience history through the present. The topics encompass papers, concepts, events, and technologies covering three geoscience disciplines: geostatistics, classical statistics and data analysis, and engineering. The colored clusters loosely represent topics of similarity and generally show a progression from early theory to commercialization.

The world moved from “mainframe” computers to early supercomputers like the Cray-1 in 1976 and the Cray X-MP in 1982. The Cray X-MP was able to perform 0.8 GFLOPS (Giga Floating Point Operations Per Second) at a cost of

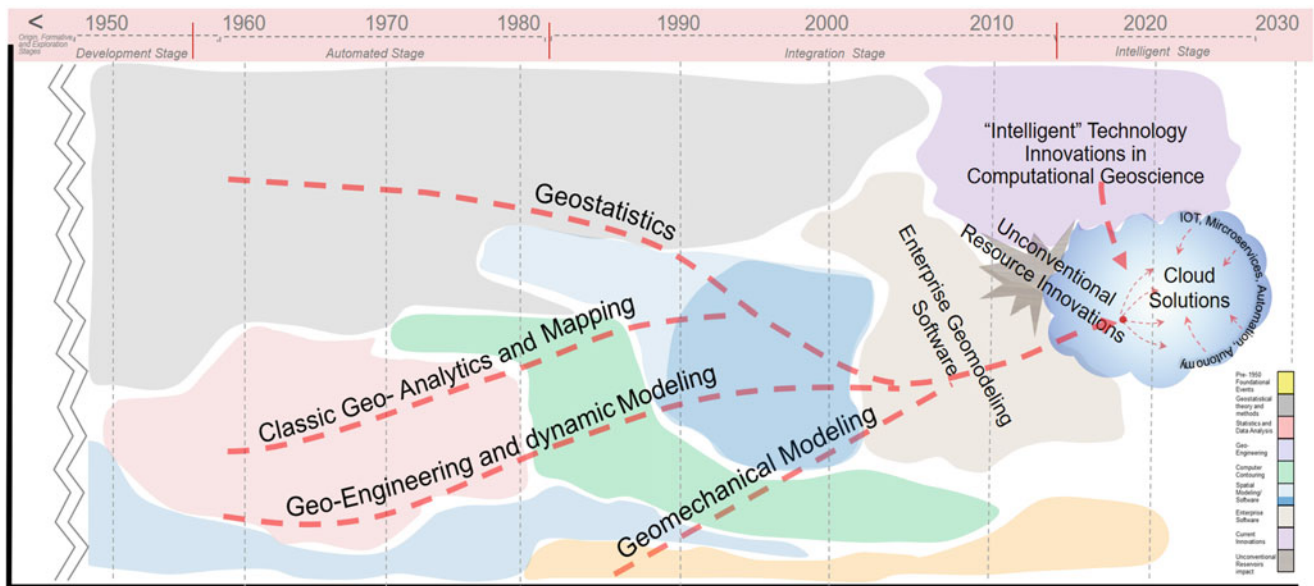
\$15,000,000, which enabled new categories of scientific problems to be addressed (Cray 2020). As of this writing, an AMD Ryzen 3600 CPU and NVIDIA RTX 3080 GPU can produce 30254 GFLOPS at a cost of \$0.04/GFLOP, and a Samsung Galaxy S20 achieves 1,267 GFLOPS for a total cost of \$1000 (FLOPS 2020). At the same time floortop, desktop, and “luggable” computers (Wang 2200-VP, IBM 5100, Osborn 1, Kaypro II, HP9807A) became available (Rhode Island Computer Museum 2020; IBM 2003; History of Computing 2020; National Museum of American History 2020; Hewlett Packard Computer Museum 2020). With the introduction of Silicon Graphics’ high-performance IRIS 1000 and 1400 in 1983 and 1985, respectively, and Windows 1.0 being launched in 1985 and broadly adopted in 1992 as Windows 3.1, a new generation of computational and graphical power emerged (Silicon Graphics – Wikipedia 2020; Earle, et al.; Microsoft Windows 2020). Some universities and large oil companies seized the opportunity to use these new tools and advanced quantitative applications.

This explosion in computational power and its greater accessibility enabled the development of statistical software which paralleled the computer evolution-revolution. A number of large mainframe-based statistical software programs were introduced including the Statistical Package for Social Scientists (SPSS) in 1968, Statistical Analysis System (SAS) in 1976 Biomedical Data Package (BMDP), and the S language for data analysis that was developed in 1976 by John Chambers of Bell Labs in close relationship to the Unix operating system concurrently developed by Dennis Ritchie (S was the precursor to today’s R statistical programming language) (Nie 2020; Aster 2016; Frane 1976; Becker 1994). These packages were in general use in a variety of scientific disciplines including the growing quantitative geoscience community. Some early software development within the geoscience disciplines made a significant impact on Earth sciences. Many of these impacts are identified on the timeline shown in Fig. 1. Of particular note is Surface II, written in Fortran by Robert Sampson of the Kansas Geological Survey (Sampson 1975). This mainframe-based program was perhaps the first integrated geoscience graphics software package that incorporated not only traditional automated contouring and statistical map analysis but also geostatistical kriging and drift analysis (Olea 1975) which had been formalized only 10 years earlier by George Matheron (1965).

As computers have become more computational and graphically more capable, a tension develops between graphics for visualization of data and the main method for user interaction. Donald Knuth points out that the ideal approach is that code should be interwoven with its explanation, and the results of the data analysis should compile to be the final report. This is a useful definition of reproducible science, a challenge to this day. Progress is being made using code-based data science tooling (e.g., RStudio integrated



Geoinformatics Fig. 1 Timeline includes papers, concepts, events, and technologies intended to represent key developments. The chart follows Merriam's quantitative development stages and Ma's suggested current *intelligent stage* (Ma 2019). The color coding is a product of simple clustering by general topic. Events prior to 1950 are lumped together and meant only to remind us of some fundamentals upon which geoscience is based



Geoinformatics, Fig. 2 The *development stage*, *automated stage*, and *integration stage* converge on the Enterprise Geomodeling Software solutions. The path forward in the *intelligent stage* shows movement

into cloud-based high-performance solutions and to the dismantling of enterprise software to microservices across CPUs, GPUs, and TPUs

development environment (IDE) for R coding) (Peng 2011; Broman et al. 2017; Bajuk and Howe 2020).

A successful software package originating from geosciences is Environmental Systems Research Institute's (ESRI) Arc/Info released initially in 1982. Arc/Info was an early geographic information system (GIS), built on the legacy of earlier products (GIS Geography 2020). Today, Arc/Info is a full-featured, menu-driven, enterprise software solution (Fig. 2) and is part of the current ArcGIS Desktop product line. Currently, ESRI is the largest GIS software company. Code-based data science approaches have also been rapidly advancing in geospatial analysis as discussed, for example, in Lovelace's book *Geocomputation in R* (Lovelace et al. 2019).

There is a need for a descriptive term for computational systems and knowledge generation from geoscience information. The vast amount of qualitative data, prior to the current digital age (what Ma calls the *intelligent stage*, Figs. 1 and 2), the influence of early mathematicians and statisticians, and the compelling draw to computer systems by early adopters, would lead today's geoscientists to invent the term "geoinformatics," if it had not been suggested already (Ma 2019). In *Stratigraphy and Sedimentation*, Krumbein and Sloss all but canonized the definition. They stated, "Automatic data processing [in computers] provides the essential means for coping with such large masses of data by providing a mechanism by which data-searching, arranging and analysis can be performed with a minimum of hand labor" (Krumbein and Sloss 1963).

Quantitative Geosciences: The Twenty-First Century

The foundation of modern informatics systems is its integration with information. Prior to the computer revolution, the fundamental informatics system resembled a filing cabinet. The process of composing and adding information (data entry) is considered part of the informatics process. Today, data entry and retrieval has become more sophisticated. Databases can be federated and connected to meaningful analytical processes like data mining, forecasting, imputation, and other multidirectional information feeds, all of which serve to expand and improve data accessibility and richness. The extent of geoinformatic activities is dependent on the infrastructure surrounding the information. Sophisticated federated databases can be designed to find and support multiple data sources and formats, but their effectiveness depends on the volume of accessible rich information and human-computer interfaces or automated machine decisions. Ma suggests extending Merriam's quantitative development stages to include an *intelligent stage* (Ma 2019). He postulates that we have entered a period of information science and computational intelligence, a move that can drive real-time data integration, automated analytics using code-based data science, and deep learning and artificial intelligence.

Modern database technology, machine readable data, and changes in computational systems are examples of geoinformatic computational intelligence and changing the way we work to solve technical challenges.

Databases. Database systems can be broken down into **relational** and **non-relational databases**. Relational databases store data in a variety of tables related to each other by a common data field. **Relational Database Management Systems (RDBMS)** are used to create, update, and administer relational databases. Examples include Oracle, MySQL/MariaDB, and PostgreSQL; each enforces strict schemas for data management and is based on **Structured Query Language (SQL)**. Relational databases are widespread and commonplace largely due to their ease of access and ability to be scaled or extended.

Non-relational databases are classified on the basis of how they organize data. Unlike relational databases, they do not rely on traditional tables (fixed table schema) as storage structures or SQL-type management systems and thus are referred to as **NoSQL** (meaning non-SQL, or not only SQL). This flexibility facilitates Big Data management of various data formats found across the geoscience domain. Such organized data is referred to as a data warehouse or data lake when applied to storing very large amounts of data (**Big Data**). A complete discussion of these classification types can be found in Jatana et al. (2012). NoSQL databases are growing in popularity for use with geospatial data (Zang et al. 2014) and spatiotemporal data in energy applications such as photovoltaic power plants (Hu et al. 2017).

Machine Readable Data. The World Wide Web Consortium (W3C) has implemented metadata models that relate to language, meaning, or logic in an effort to make Internet data machine-readable. This effort exists as an extension of the web called the **Semantic Web** (French et al. 2015) which is a layered architecture comprising elements such as **Extensible Markup Language (XML)**. XML is a metalanguage specifying syntax without semantics. The **Resource Description Framework (RDF)**, a set of World Wide Web metadata specifications, appends ontological tags to Internet data in a pre-defined format (subject, predicate, and object), thereby providing semantic conceptual descriptions. The RDF is one example of how the Semantic Web has leveraged metadata to achieve machine readable information. **Web Ontology Language (OWL)** is another more powerful example. This is done through the use of structured ontologies required for data entry that classify data objects into nouns and relate data objects with verbs. This allows applications to process the information instead of simply presenting it. Software can be made to “read” and manipulate ontological structures. These can be serialized or converted to digital format for storage and transfer. Such Web services are particularly useful for data discovery exemplified by the **US Geological Survey (USGS)** that uses **JavaScript Object Notation (JSON)** to serialize the ontological information used in their National Digital Catalog (Runnels 2016).

And as computing advances to Exascale (10^{18} floating point operations per second = 1 exaFLOPS) (Kothe et al. 2019; Schulthess et al. 2019), there is an effort to enable

artificial intelligence (AI) and deep learning in an explicit Semantic Web implementation, based on the FAIR principles, that datasets, software, and models should all be **Findable, Accessible, Interoperable, and Reusable**, so as to enable machines to implement AI in modeling (Wilkinson et al. 2016, 2019; Fagnan et al. 2019).

Interaction with Computational Systems. The more common use of continuous integration in software engineering (CI) in combination with high-level software frameworks, such as TensorFlow2 and programming languages like R and Python, increased capacity of Central Processing Units (CPU), the expanded neural network capabilities now enabled in **graphic processing units (GPU)**, and **tensor processing units (TPU)**, has initiated a reengineering of many of the research and enterprise software solutions (Software framework – Wikipedia 2020; Jouppi et al. 2017). Cloud computing through platforms like Amazon’s AWS, Google’s GCP, and Microsoft’s AZURE is encouraging a transition to service-based cloud, hybrid cloud, and edge computing. These services provide cloud-based virtual computer systems or infrastructure as a service (**IaaS**), **Software As A Service (SaaS)**, and development **Platforms As A Service (PaaS)**. Such “aaS” paradigms provide sophisticated computing that offer flexibility while minimizing the need for expensive desktop or computer hardware. Figure 2 shows a generalized geotechnology trajectory and current migration of enterprise desktop software to the Cloud, accommodating the trend toward high-performance distributed computing and “aaS” paradigms.

The solution to current canonical geospatial problems is not just technical and scientific but also societal. For example, the world currently exists in a transitional state between fossil and renewable resources, a drive that ultimately pushes toward greener energy and a reduction in carbon footprint to help mitigate climate change. Juggling the science behind finding energy resources, the economics to develop them, and the policy solutions for fair governance, are nontrivial problems. R and Python can serve as interface languages that allow easy incorporation of machine learning, deep learning, and artificial intelligence, with GPUs (like in the AMD/Nvidia compute mentioned above) and deep learning frameworks (such as TensorFlow2 and Torch). This current and rapidly evolving technology addresses such immense challenges. Now this technology is accessible to geoscientists and other scientific communities (LeCun et al. 2015; Abadi et al. 2016; Paszke et al. 2019).

Current Investigation Examples

Geoscientists now have better insights into some of their canonical problems through the use of geoinformatics. The following are some examples of these investigations.

Planet Tectonism. New ways of understanding the cooling and transfer of heat from planet interiors and the impact on volcanic activity through heat pipe cooling are being investigated (Moore et al. 2017). While many proposed methods have been reliable, various efficiencies are required to optimize their use in modern computational systems and allow communication and comparison among different solvers. Working directly with common applications, an open-source parallel solver and discretization library has been proposed as a community tool supplying scalable, portable, reproducible performance toward novel science in regional and planet-scale geodynamics (Sanan et al. 2017). Scalable geospatial data analytics and other tools can be leveraged to inform efforts to recover valuable resources and plan for human needs (Klein et al. 2015).

Predicting Climate Change and the Associated Implications Using Geological Records. Open data access to climate information is now ubiquitous. Climate change is one of the most important hazardous impacts facing the future of human civilization (Schulthess et al. 2019). Recent geospatiotemporal analytics applied to the geography of climate change in California and Nevada using the PRISM and WorldClim interpolated climate databases shows the spatio-temporal relationships and heterogeneity at different scales. The research shows potential impacts of climate change and guides the evaluation and implementation of conservation strategies, which are identified as one of the main Exascale computing challenges (Ackerly et al. 2010). Another example is the Energy Exascale Earth System Model, in which the complex biogeochemistry and the geo-spatiotemporal aspects are used to simulate the historical carbon cycle dynamics, including carbon losses predicted in response to land use and land cover change and the responses of the carbon cycle to changes in climate (Burrows 2020).

Predicting Earthquakes

USGS developed an automated system for quickly estimating the fatality and economic loss from earthquakes and automatically assessing public response needs and informing government. **Prompt Assessment of Global Earthquake Response (PAGER)** uses XML, MySQL, and other geoinformatic components in the on-time relational database, to receive global information globally on seismological activity for measuring the severity of an earthquake, computing ground motion amplification maps, and ground shaking estimates within 30 min (Earle et al. 2009; Gundersen 2008).

Mapping and Analyzing Global Geospatial Data

Geospatial data can be hundreds of terabytes, or petabytes in size, use a variety of formats and projection methods, and

have multiple levels of resolution. These characteristics can impede data discovery from existing databases and require extensive data processing before applying analytical methods. For example, performing an analysis or building a model with data that has been referenced using a geographic coordinate system like WGS84 (global) and data using a projected coordinate system like NAD83 (south central Texas) can create dubious results if not rectified. Recent efforts to develop a geospatial platform for rapid retrieval and analysis of data from diverse, current, and historical databases are showing success. One such platform is the **Physical Analytics Integrated Repository and Services**, or **PAIRS** (Klein et al. 2015).

Geoinformatics: Discussion Topics

Open vs. Private Data. The digital landscape of geoscience crosses multiple disciplines both public and private facing. Data that are freely available to use is referred to as “open data,” following the concept of openness (Perens 2005) and its related concepts of open source software, open access, which contrast with “private” data that is proprietary. Open data provides scientists with a similar benefit experienced by programmers using open source technologies (i.e., large communities of talent working together). Before the advantages of open data can fully be harnessed, applications and separate datasets need to be interoperable and standards need to be addressed. Ongoing efforts are being made to standardize open data and make it more easily discoverable, supported by applications, and machine readable with minimal human intervention, as shown, for example, in the global dataset of wind and solar farms based on OpenStreetMap data (Wilkinson et al. 2016, 2019; Dunnett et al. 2020).

Private data are not necessarily subject to the same standards. Access to private data often requires legal non-disclosure or data use agreements with restrictions and prevents broader community support. The idea that knowledge growth can occur concurrently through public transparency and commercial privatization is problematic. The net effect can often result in a level of discord between industry and the public. Ultimately, the concept of open data and the privatization of data will hinge upon government policies (G8 2013; Obama 2013). The question of public transparency, knowledge growth, and free enterprise remains open, and many companies have identified ways to be commercially successful even with open source products, such as RedHat Linux (McMillan 2012).

Automation and Beyond. AI and automation are already influencing academic, governmental, and industrial institutions. The “Big Crew Change,” from baby boomer subject matter experts to GenX and Millennial young professionals, is well underway, and the geosciences are changing accordingly. Understanding the trend in technologies with the

purpose of predicting the future impact of these changes will help business leaders to define necessary skill sets (BHEF 2016). To be sustainable as a geoscientist without an education in data science and geoinformatics will be very challenging. Universities should incorporate these key disciplines into Earth science curricula in order to stay relevant and prepare for the inevitable transition from automation to robotics and autonomous computing. The question of the degree of automation and autonomy and the role of human supervision remains open.

Summary

Geoinformatics is about turning geoscience data and information into knowledge through the fusing of statistical and numerical methods with computer technologies. These processes are quantitative, which may seem unusual for a discipline whose origins are built largely on the pillars of qualitative analysis. Quantitative approaches, while not common in the early years, were highly impactful. From the beginnings of modern geoscience in the nineteenth century with Sir Charles Lyell, adoption of mathematics in the early to mid-twentieth century by pioneers like Griffiths, Krumbein, Sloss, and Matheron, to embracing geoscience software by researchers and industry practitioners in the mid-to-late twentieth century, the Earth sciences are squarely placed in the today's digital explosion. The construction and use of geoinformatic systems today are critical to solving twenty-first century geo-canonical problems. Understanding rocks and fluids at the nano-scale, modeling the impact of climate change, the prediction of earthquakes, and even data mining the vast storage of our historical qualitative data represent only some of the work ahead. To succeed, the geoscience community must recognize that our professional makeup must go well beyond our traditional domain specialities and include the continuous adaptation to geoinformatics which must be woven into the very nature of our common disciplines. As these systems merge with the expanding computational cloud and machine learning resources to form integrated geo-spatiotemporal data science and analytics, our future capabilities may be hard to imagine but will be impressive as they are brought to bear on the upcoming challenges.

Cross-References

- Artificial Intelligence in the Earth Sciences
- Cloud Computing and Cloud Service
- Data Life Cycle
- Data Mining
- Data Visualization

- Database Management System
- Deep Learning in Geoscience
- Digital Elevation Model
- Exploratory Data Analysis
- FAIR Data Principles
- Geocomputing
- Geomatics
- Machine Learning
- Mathematical Geosciences
- Metadata
- Multivariate Analysis
- Object-Based Image Analysis
- Ontology
- Open Data
- Quantitative Stratigraphy
- Self-Organizing Maps
- Spatial Analysis

Bibliography

- Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al (2016) TensorFlow: a system for large-scale machine learning. Proceedings of the 12th USENIX symposium on operating systems design and implementation 16 [Internet]. USENIX Association, Savannah, [cited 2019 Jan 26], pp 265–283. Available from: <https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi>
- Ackerly DD, Loarie SR, Cornwell WK et al (2010) The geography of climate change: implications for conservation biogeography: geography of climate change. *Divers Distrib* 16:476–487. <https://doi.org/10.1111/j.1472-4642.2010.00654.x>
- Agterberg FP (1974) *Geomathematics*. Elsevier, Amsterdam. 596 pp
- Allaby M (2013) *A dictionary of geology and earth sciences*. Oxford University Press. Retrieved 27 Aug 2020, from <https://www.oxfordreference.com/view/10.1093/acref/9780199653065.001.0001/acref-9780199653065>
- Aster R (2016) SAS: history of SAS versions. <http://www.globalstatements.com/sas/differences/>. Accessed Nov 2020
- Bajuk L, Howe C (2020) Why RStudio focuses on code-based data science [Internet]. R Studio Blog. [cited 2020 Nov 30]. Available from: <https://www.blog.rstudio.com/2020/11/17/an-interview-with-lou-bajuk/>
- Becker RA (1994) A brief history of S. computational statistics [Internet]. 9:81–110. Available from: <https://www.math.uwaterloo.ca/~rwdldfor/software/R-code/historyOfS.pdf>
- Broman K, Cetinkaya-Rundel M, Nussbaum A, Paciorek C, Peng R, Turek D, et al (2017) Recommendations to funding agencies for supporting reproducible research. American Statistical Association [Internet]. Available from: <https://www.amstat.org/asa/files/pdfs/pol-reproducible-research-recommendations.pdf>
- Burrows SM, Maltrud M, Yang X, et al (2020) The DOE E3SM v1.1 biogeochemistry configuration: description and simulated ecosystem-climate responses to historical changes in forcing. *Journal of Advances in Modeling Earth Systems* 12:e2019MS001766. <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001766>
- Business Higher Education Forum (2016) Creating a minor in applied data science | BHEF [Internet]. The Business Higher Education Forum, pp 1–16. Available from: <http://www.bhef.com/publications/creating-minor-applied-data-science>

- Cray – The Reader Wiki, Reader View of Wikipedia. Accessed 5 Nov 2020
- Cubitt JM, Henley S (1978) Statistics in geology. Dowden, Hutchinson & Ross, Stroudsburg. 340 pp
- Dunnett S, Soricichetta A, Taylor G, Eigenbrod F (2020) Harmonised global datasets of wind and solar farm locations and power. Scientific Data 7:130. <https://doi.org/10.1038/s41597-020-0469-8>
- Earle P, Wald D, Jaiswal KS, Allen TI, Hearne MG, Marano KD, Hotovec AJ, Fee JM (2009) Prompt Assessment of Global Earthquakes for Response (PAGER) – a system for rapidly determining the impact of earthquakes worldwide. U.S. Geological Survey Open-File Report 1131, 15 pp
- Evans & Sutherland – Wikipedia. https://www.en.wikipedia.org/wiki/Evans_%26_Sutherland. Accessed 29 Nov 2020
- Fagnan K, Nashed Y, Perdue G, Ratner D, Shankar A, Yoo S (2019) Data and models: a framework for advancing AI in science. USDOE Office of Science (SC) (United States) <https://doi.org/10.2172/1579323>
- FLOPS (2020) [Internet]. Wikipedia. [cited 2020 Nov 30]. Available from: <https://en.wikipedia.org/w/index.php?title=FLOPS&oldid=989051847>
- Frane JW (1976) The BMD and BMDP series of statistical computer programs. Commun ACM 19:570–576. 7, New York
- French RH, Podgornik R, Peshek TJ et al (2015) Degradation science: mesoscopic evolution and temporal analytics of photovoltaic energy materials. Curr Opin Solid State Mater Sci 19:212–226. <https://doi.org/10.1016/j.cossms.2014.12.008>
- GIS Geography (2020) The remarkable history of GIS. Accessed 16 Nov 2020
- Gohau G (1990) A history of geology. Rutgers University Press, New Brunswick. 259 pp
- Group of 8 (G8) (2013) Open Data Charter [Internet]. Available from: <https://www.gov.uk/government/publications/open-data-charter>
- Gundersen LC (2008) Answers to earth system science questions – the evolution of informatics at the U.S. Geological Survey. In: Shailaja RB, Sinha AK, Gundersen LC (eds) Geoinformatics 2008 – data to knowledge. scientific investigations report 2008–5172. U.S. Geological Survey, Reston, pp 2–3
- Hewlett Packard Computer Museum (2020) Technical desktops. HP Computer Museum. Accessed 16 Nov 2020
- History of Computing (2020) Osborn 1. Osborne 1 – Complete History of the Osborne 1 Computer. Accessed 16 Nov 2020
- Hu Y, Gunapati VY, Zhao P, Gordon D, Wheeler NR, Hossain MA, Peshek TJ, Bruckman LS, Zhang G, French RH (2017) A nonrelational data warehouse for the analysis of field and laboratory data from multiple heterogeneous photovoltaic test sites. IEEE J Photovoltaics 7:230–236. <https://doi.org/10.1109/JPHOTOV.2016.2626919>
- IBM (2003) IBM archives: IBM 5100 portable computer. http://www.ibm.com/ibm/history/exhibits/pc/pc_3.html. Accessed 16 Nov 2020
- Jatana N, Puri S, Ahuja M et al (2012) A survey and comparison of relational and non-relational database. Int J Eng Res 1:5
- Jouppi NP, Young C, Patil N, Patterson D, Agrawal G, Bajwa R, et al (2017) In-datacenter performance analysis of a tensor processing unit. Proceedings of the 44th annual international symposium on computer architecture [Internet]. Association for Computing Machinery, New York. [cited 2020 Nov 30]. pp 1–12. Available from: <https://doi.org/10.1145/3079856.3080246>
- Klein LJ, Marianno FJ, Albrecht CM et al (2015) PAIRS: a scalable geo-spatial data analytics platform. In: 2015 IEEE international conference on big data (big data). IEEE, Santa Clara, pp 1290–1298
- Knuth DE (1984) Literate programming. The Computer Journal [Internet]. 1 [cited 2017 May 23];27(2):97–111. Available from: <https://www.academic.oup.com/comjnl/article/27/2/97/343244/Literate-Programming>
- Kothe D, Lee S, Qualters I (2019) Exascale computing in the United States. Comput Sci Eng 21(1):17–29
- Krumbein WC, Sloss LL (1963) Stratigraphy and sedimentation. W.H. Freeman and Company, San Francisco. 600 pp
- LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature [Internet]. [cited 2016 Dec 21]. 521(7553):436–44. Available from: <https://www.nature.com/nature/journal/v521/n7553/abs/nature14539.html>
- Lovelace R, Nowosad J, Muenchow J (2019) Geocomputation with R [internet], 1st edn. Chapman and Hall/CRC, London. [cited 2020 Nov 30]. Available from: <https://www.geocompr.robinlovelace.net/>
- Ma X (2019) Geo-data science: leveraging geoscience research with geoinformatics, semantics and open data. Acta Geologica Sinica – English Edition 93:44–47. <https://doi.org/10.1111/1755-6724.14240>
- Matheron G (1965) Les variables régionalisées et leur estimation, une application de la théorie de fonctions aléatoires aux sciences de la nature: Masson et Cie. Editeurs, Paris. 305 p
- McMillan R (2012) Red Hat becomes open source's First \$1 Billion Baby. Wired [Internet]. [cited 2020 Nov 28]. Available from: <https://www.wired.com/2012/03/red-hat/>
- Merriam DF (2004) The quantification of geology: from abacus to Pentium. Earth Sci Rev 67:55–89. <https://doi.org/10.1016/j.earsci.2004.02.002>
- Merriam DF, Howarth RJ (2004) Pioneers in mathematical geology. Earth Sci Hist 23(2):314–324. Accessed 27 Aug 2020. <http://www.jstor.org/stable/24137098>
- Microsoft Windows – Wikipedia. https://www.en.wikipedia.org/wiki/Microsoft_Windows. Accessed 29 Nov 2020
- Moore WB, Simon JJ, Alexander GW (2017) Heat-pipe planets. Earth Planet Sci Lett 474:13–19. <https://doi.org/10.1016/j.epsl.2017.06.015>
- National Museum of American History (2020) Kaypro IV Portable Computer. https://americanhistory.si.edu/collections/search/object/nmah_1156670. Accessed 16 Nov 2020
- Nie_SPSS – Wikipedia. <https://www.en.wikipedia.org/wiki/SPSS>. Accessed 29 Nov 2020
- Obama BH (2013) Executive order – making open and machine readable the new default for Government Information [Internet]. [cited 2014 Jan 4]. Available from: <https://obamawhitehouse.archives.gov/open>
- Olea RA (1975) Optimal mapping techniques using regionalized variable theory. Series on spatial analysis, no.2. Kansas Geological Survey, Lawrence
- Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al (2019) PyTorch: an imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, dAlché-Buc F, Fox E, Garnett R, eds. Advances in neural information processing systems 32 [Internet]. Curran Associates, Inc., pp 8024–8035. Available from: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
- Peng RD (2011) Reproducible research in computational science. Science [Internet]. [cited 2014 Jun 12]. 334(6060):1226–1227. Available from: <http://www.sciencemag.org/content/334/6060/1226>
- Perens B (2005) The open source definition, by The Open Source Initiative. http://www.opensource.org/docs/definition_plain.php. [Internet]. [cited 2020 Nov 28]. Available from: <https://www.ci.nii.ac.jp/naid/10022604704/>
- Rhode Island Computer Museum (2020) Wang 2200 VP WCS/20. In: Large systems collection, Rhode Island Computer Museum. <https://sites.google.com/a/ricomputermuseum.org/home/collections-gallery/equipment>. Accessed 16 Nov 2020
- Runnels MJ (2016) Science base: sciencebase catalog item REST services. <https://my.usgs.gov/confluence/display/sciencebase/ScienceBase+Catalog+Item+REST+Services>. Accessed 16 Nov 2020
- Sampson RJ (1975) Surface II user's manual: Kansas geological survey series in spatial analysis. University of Kansas, Lawrence. 206 p

- Sanan P, Tackly P, Gara T, Gauss BJP, May DA (2017) StagBL: a scalable, portable, high-performance discretization and solver layer for geodynamic simulation. 20th EGU General Assembly, EGU2018 Proceedings, Vienna, Austria, 1248 pp
- Schulthess TC, Bauer P, Wedi N, Fuhrer O, Hoeffler T, Schär C (2019) Reflecting on the goal and baseline for exascale computing: a roadmap based on weather and climate simulations. *Comput Sci Eng* 21(1):30–41
- Shiryayev A (1989) Kolmogorov: life and creative activities. *Ann Prob* 17:866–944.
- Silicon Graphics – Wikipedia. https://www.en.wikipedia.org/wiki/Silicon_Graphics. Accessed 29 Nov 2020
- Software framework – Wikipedia. https://www.en.wikipedia.org/wiki/Software_framework. Accessed 29 Nov 2020
- Wang C (2018) Information extraction and knowledge graph construction from geoscience literature. *Comput Geosci* 112:112–120
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* [Internet]. [cited 2020 Apr 13]. 3(1):1–9. Available from: <https://www.nature.com/articles/sdata201618>
- Wilkinson MD, Dumontier M, Jan Aalbersberg I, Appleton G, Axton M, Baak A, et al (2019) Addendum: the FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* [Internet]. Nature Publishing Group. [cited 2020 Apr 13]. 6(1):1–2. Available from: <https://www.nature.com/articles/s41597-019-0009-6>
- Zhang X, Song W, Liu L (2014) An implementation approach to store GIS spatial data on NoSQL database. In: 2014 22nd international conference on geoinformatics. IEEE, Kaohsiung, pp 1–5. <https://doi.org/10.1109/GEOINFORMATICS.2014.6950846>

Geologic Time Scale Estimation

Felix M. Gradstein¹ and Frits Agterberg²

¹Founding Member and Past Chair, Geologic Time Scale Foundation, Emeritus Natural History Museum, University of Oslo, Oslo, Norway

²Central Canada Division, Geomathematics Section, Geological Survey of Canada, Ottawa, ON, Canada

Definition

The Geologic Time Scale (GTS) is the framework for deciphering and understanding the history of our planet. The steady increase in data, development of better methods and new procedures for actual dating and scaling of the rocks on Earth, and a refined relative scale with more defined units pose challenges to the methodology employed to calculate the GTS. This review explains the qualitative and quantitative methods used in GTS2020 with special emphasis on cubic splining. Relative to its ancestor GTS2012, the qualitative and quantitative methodology for GTS2020 is improved, uncertainty reduced, and stratigraphic resolution increased.

Introduction

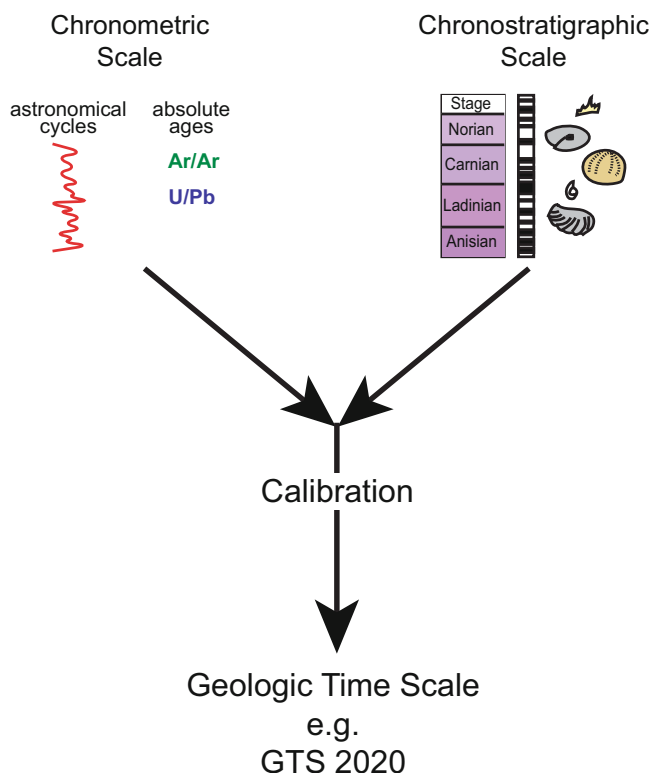
The Geologic Time Scale (GTS) is the framework for deciphering and understanding the long and complex history of our planet, Earth. As Arthur Holmes, the father of the Geologic Time Scale, once wrote (Holmes 1965, p. 148): “To place all the scattered pages of earth history in their proper chronological order is by no means an easy task.” Ordering these pages, and understanding the physical, chemical, and biological processes that acted on them since Earth appeared and solidified more than 4 billion years ago, requires a detailed and accurate time scale. The time scale is the tool “par excellence” of the geological trade, and insight in its construction, strength, and limitations greatly enhances its function and utility. This calibration to linear time of the succession of events recorded in the rocks on Earth has three components:

1. The international stratigraphic divisions and their correlation in the global rock record
2. The means of measuring linear time or elapsed durations from the rock record
3. The methods of joining the two scales, the stratigraphic one and the linear one

For clarity and precision in international communication, the rock record of Earth’s history is subdivided into a “chronostratigraphic” scale of standardized global stratigraphic units, such as “Carboniferous,” “Eocene,” “*Zigzagiceras zigzag* ammonite zone,” or “polarity Chron M19r.” Unlike the continuous ticking clock of the “chronometric” scale (measured in years before the year 2000 AD), the chronostratigraphic scale is based on relative time units in which global reference points at boundary stratotypes define the limits of the main formalized units, such as “Permian.” The chronostratigraphic scale is an agreed convention, whereas its calibration to linear time, using radiometric and cyclostratigraphic methods, is a matter for discovery or estimation (Fig. 1).

In contrast to the Phanerozoic that has an agreed upon chronostratigraphic scale with formal stage boundary stratotypes, Precambrian stratigraphy is formally classified chronometrically, i.e., the base of each Precambrian eon, era, and period is assigned a numerical age. Below, we review geomathematical methods used in the Phanerozoic, spanning the last 538 million years. Most information and figures (except for new Fig. 5 and new Table 2) are from Agterberg et al. (2020), Agterberg et al. (2021), and Gradstein et al. (2020), and these will not be quoted again.

Figure 2 shows the current Phanerozoic Geologic Time Scale (GTS2020) with the Paleozoic, Mesozoic, and Cenozoic Eras and their subdivision into more than 10 periods and over 100 stages. The reason that quantitative methodology is



Geologic Time Scale Estimation, Fig. 1 The construction of a geologic time scale is the merger of a chronometric scale (measured in years) and a chronostratigraphic scale (formalized definitions of geologic stages, biostratigraphic zonation units, magnetic polarity zones, and other subdivisions of the rock record)

employed for geologic time scale building is the simple fact that only nine out of 100+ stage boundaries have a precise and accurate age date.

Methods and Ages

The methods used for the construction of Geologic Time Scale 2020 (GTS2020) integrate different techniques depending on the quality of data and methods available for different intervals, as shown schematically in Fig. 3. In this review we focus on cubic splining with composite standards from Ordovician through Devonian and cubic splining with orbital tuning and magnetostratigraphy for part of the Cretaceous and Jurassic.

GTS2020 time scale construction may be summarized as follows:

- Step 1. Construct an updated global chronostratigraphic scale for the Earth's rock record from standard outcrop sections and cored ocean drill cores.
- Step 2. Scale the updated chronostratigraphic scale with magnetostratigraphy or a composite standard technique.

The latter commonly takes average zone thickness from many sections as directly proportional to zone duration. It is applied for Paleozoic periods.

- Step 3. Identify key linear-age calibration levels for the chronostratigraphic scale using radiogenic isotope age dates, and/or apply astronomical tuning to cyclic sediment.
- Step 4. Interpolate the combined chronostratigraphic and chronometric scale, for example with a cubic spline that fits closest to data points with the lowest possible error. Such splines effectively bridge gaps in data along the linear scale or along the chronostratigraphic scale.
- Step 5. Calculate or estimate error bars on the combined chronostratigraphic and chronometric information to obtain a geologic time scale with estimates of uncertainty on stage boundary ages.

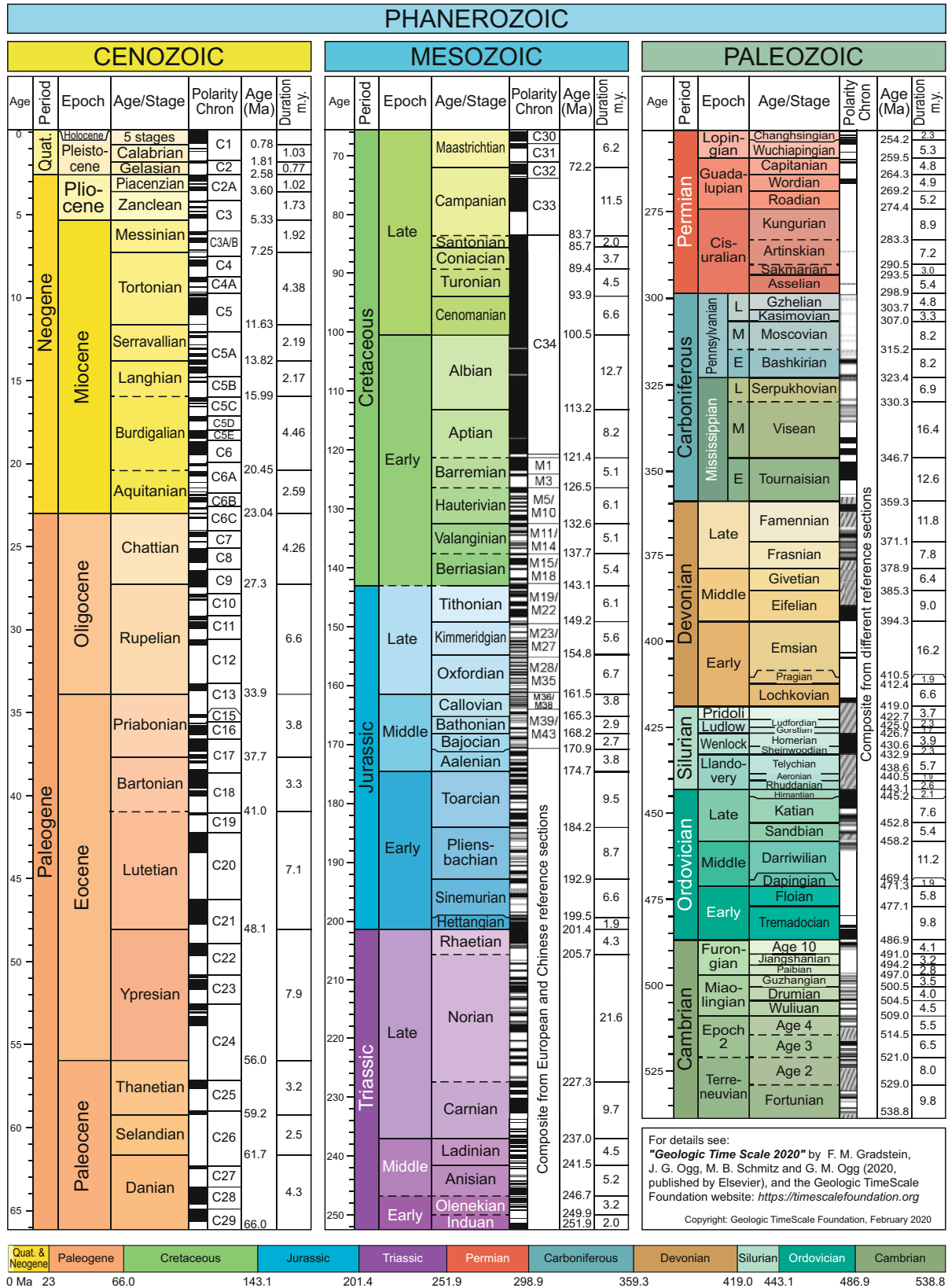
The first step, integrating multiple types of stratigraphic information in order to construct the chronostratigraphic scale, is the most time-consuming; it summarizes and synthesizes centuries of detailed geological research, while reconciling it with the most up-to-date information. The second step has lacked progress, with Periods like Cambrian, Devonian, Triassic, and part of Jurassic not yet having a composite standard of bio- and other events. The third step, identifying which radiogenic isotope and cycle-stratigraphic studies are to be used as the primary constraints for assigning linear ages, is the one that has been evolving most rapidly over the last decade. Historically, time scale building went from an exercise with few and relatively inaccurate radiogenic isotope dates, as used by Holmes (1947, 1960), to one with many dates with greatly varying analytical precision (like GTS89), to one with a majority of accurate and often precise dates (like GTS2020). The new philosophy, which was started with GTS2004 and continued in GTS2012 and GTS2020, is to select analytically precise radiogenic isotope dates with high stratigraphic resolution. More than 330 radiogenic isotope dates were thus selected for their reliability and stratigraphic importance to calibrate the geologic record in linear time.

The uncertainty on older stage boundaries systematically increases owing to potential systematic errors in the different radiogenic isotope methods, rather than to the analytical precision of the laboratory measurements. In this connection it is good to remember that biostratigraphic error is fossil event and fossil zone dependent, rather than age dependent.

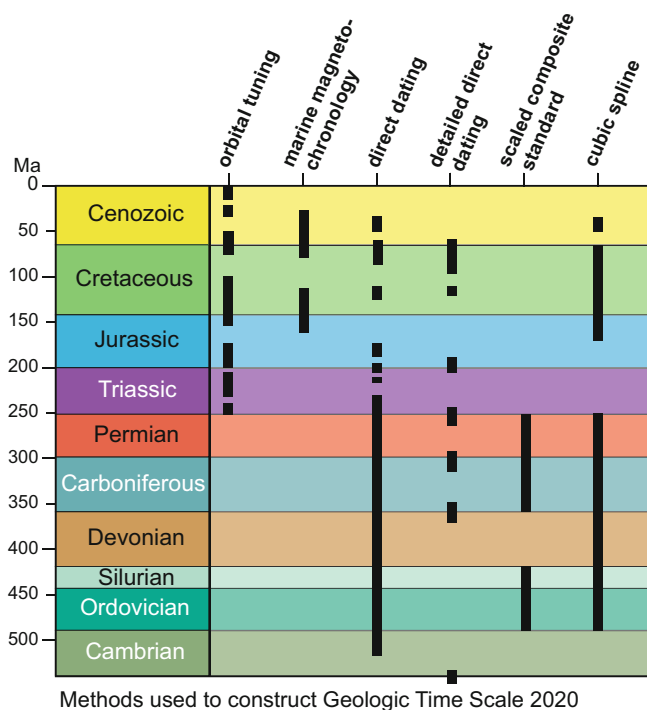
Ages and durations of Cenozoic stages derived from orbital tuning are considered to be accurate to within a precession cycle (~20 kyr) assuming that all cycles are correctly identified, and that the theoretical astronomical-tuning for progressively older deposits is precise.



GEOLOGIC TIME SCALE 2020



Geologic Time Scale Estimation, Fig. 2 The new Geologic Time Scale 2020 (GTS2020)



Geologic Time Scale Estimation, Fig. 3 Methods used to construct Geologic Time Scale GTS2020 integrate different techniques depending on the quality of data available within different stratigraphic intervals

Conventions and Standards

Ages are given in years before “Present” (BP). To avoid a constantly changing datum, “Present” was fixed as AD 1950 (as in ^{14}C determinations), the date of the beginning of modern isotope dating research in laboratories around the world. For most geologists, this offset of official “Present” from “today” is not important. However, for archeologists and researchers into events during the Holocene (the past 11,500 years), the offset between the “BP” convention from radiogenic isotope laboratories and actual total elapsed calendar years becomes significant. The offset between the current year and “Present” has led many Holocene specialists to use a “BP2000,” which is relative to the year AD 2000. This practice is used in GTS2020 also.

For clarity, the linear age in years is abbreviated as “a” (for annum), and ages are measured in *ka*, *Ma*, or *Ga* for thousands, millions, or billions of years before present. Elapsed time or duration is abbreviated as “yr.” (for year), and longer durations in *kyr* or *myr*. Therefore, the Cenozoic began at 66 *Ma*, and spans 66 *Myr* (to the present day, defined as the year 2000 AD).

The *Ma* and *Myr* practice is confusing and inconsistent both internally and with respect to SI (Le Système international d’unités). As Holden et al. (2011) elegantly clarify (on behalf of IUPAC and IUGS), the same unit is used for absolute and relative measurements. This is in compliance

with quantity calculus, and its unit assignment for continuous and interval scales. Elapsed time or duration should also be abbreviated as in *ka* or *Ma*. Therefore, the Cenozoic began at 66 *Ma*, and spans or lasted 66 *Ma*. This is similar to the use of m (meters) for both absolute depth/distance and a depth/distance difference, and the use of $^{\circ}\text{C}$ for both temperature and temperature difference. For example, we say that the interval between two log markers in a borehole is 450 m thick, and starts 900 m below ground level. Despite this solution to an often fuzzy and confusing debate with respect to the notation of age and duration units in earth science, we (for now) stick to the same format as used in GTS2004, with both *Ma*, *ka* and *Myr*, *kyr* units.

The uncertainties in computed ages or durations are expressed as standard deviation (1-sigma or 68% confidence) or 2-sigma (95% confidence). The uncertainty is indicated by “ $\pm$ ” and will have implied units of thousands or millions of years as appropriate to the magnitude of the age. Therefore, an age cited as “124.6 $\pm$ 0.3 *Ma*” implies a 0.3 *Myr* uncertainty (2-sigma, unless specified as 1-sigma) on the 124.6 *Ma* date. We present the uncertainties ($\pm$) on summary graphics of the geologic time scale as 2-sigma (95% confidence) values.

Geologic time is measured in years, but the standard unit for time is the second *s*. Because the Earth’s rotation is not uniform, this “second” is not defined as a fraction (1/86400) of a solar day, but as the *atomic second*. The basic principle of the atomic clock is that electromagnetic waves of a particular frequency are emitted when an atomic transition occurs. In 1967, the Thirteenth General Conference on Weights and Measures defined the *atomic second* as the duration of 9,192,631,770 periods of the radiation corresponding to the transition between two hyperfine levels of the ground state of cesium-133. This value was established to agree as closely as possible with the solar-day second. The frequency of 9,192,631,770 hertz (Hz) which the definition assigns to cesium radiation was carefully chosen to make it impossible, by any existing experimental evidence, to distinguish the atomic second from the ephemeris second based on the Earth’s motion. The advantage of having the atomic second as the unit of time in the International System of Units is the relative ease, in theory, for anyone to build and calibrate an atomic clock with a precision of 1 part per 10^{11} (or better). In practice, clocks are calibrated against broadcast time signals, with the frequency oscillations in hertz being the “pendulum” of the atomic time keeping device. One year is approximately 31.56 mega seconds (1 a $\approx$ 31.56 Ms).

The McKerrow Method

McKerrow et al. (1985) described the following type of method to construct a numerical time scale for the Ordovician, Silurian, and Devonian. Use was made of an iterative

construction involving a sequence of diagrams wherein the isotopic age of the sample was plotted along the x -axis and its stratigraphic age along the y -axis. They stated (p.73):

Most graphs are constructed with definite numerical scales along both the x and y axes; this is not the case with fig. 1, where only the horizontal (x) axis is numerical. The vertical (y) axis is a stratigraphic time scale, showing periods, series, stages and zones; the precise duration of each of these time divisions is unknown. In fact, the whole object of this documentation is to determine, as far as possible with the evidence available, what estimates can be given on the duration of these stratigraphic divisions. Thus, in the course of preparing this figure, we have constructed a series of graphs, each with slightly differing vertical scales, until we obtained a scale which allowed a straight line to pass through almost all the rectangles representing the analytical errors (2σ) and the stratigraphic uncertainties in the data we use.

In a later paper on the Ordovician time scale, Cooper (1999) used 14 analytically reliable and stratigraphically controlled high-resolution TIMS U–Pb zircon dates and a single Sm–Nd date. Adopting a modified version of the McKerrow method, Cooper plotted these Ordovician dates along a relative time scale that was then re-proportioned as necessary to achieve a good fit with a straight line obtained by linear regression. This method of relative shortening and lengthening of parts of the Ordovician time scale was based mainly on a comparison of sediment accumulation rates in widely different regions and, to some extent, on empirical re-proportioning. Agterberg (2002) subjected Cooper's (1999) data to spline fitting and found that the optimum smoothing factor corresponds to a straight-line fit. He then used Ripley's MLFR (Maximum Likelihood for Functional Relationship, Ripley and Thompson 1987) method to fit a straight line in which stratigraphic uncertainty was considered as well.

Smoothing Splines

A cubic-smoothing spline $f(x)$ is a sequence of cubic polynomials connecting “nodes” along the x -axis. These nodes are often set at the points where there are observed values. Spline curves are continuous at the nodes both in value of $f(x)$ and its first derivative $f'(x)$. Using index values $i = 1, 2, \dots, n$, a spline is fully determined from the values (x_i, y_i) , the standard deviations of the observed values $s(y_i)$, and a smoothing factor (SF) representing the square root of the average value of the squares of scaled residuals $r_i = (y_i - f(x_i))/s(y_i)$. All values (x_i, y_i) receive “weights” $w_i = 1/s^2(y_i)$. Special steps are taken to determine $f(x)$ at the first and last value of a sequence (cf. de Boor 1978; Wang 2011). In general, if all $s(y_i)$ values are unbiased, $SF \approx 1$, or SF is equal to a value slightly less than 1 (cf. Agterberg 1994, p. 874). If SF significantly exceeds 1, this suggests that some or all variances used are too small

(under-reported). Thus, the spline-fitting method can provide an independent method of assessing mutual consistency and average precision of published 2-sigma error bars.

The method of “leaving-out-one” cross-validation can be used to determine the optimum smoothing factor. In this method, all observed dates y_i , between the oldest and youngest one are successively left out from spline fitting with pre-selected trial values of SF. The result is $(n - 2)$ spline curves for each SF tried. The cross-validation value for any SF is the sum of squares of deviations between y_i and estimated values on the $(n - 2)$ spline curves with the same x_i values as y_i . The best SF has the smallest cross-validation value. It is noted that even if the cross-validation pattern shows a well-developed minimum at a value not close to 1, adoption of the optimum SF value instead of $SF = 1$ generally constitutes only a minor improvement of the spline curve. In Paleozoic applications, the optimum SF is generally smaller than 1. Setting $SF = 1$ would result in slightly smoother spline curves.

Unless its unit is consistently proportional to geologic time measured in millions of years, the numerical time scale resulting from spline fitting is not linearly related to the initial relative time scale. The latter is based on chronostratigraphic information such as relative thicknesses of biozones in Paleozoic periods or geographic locations of geomagnetic inversion points along ocean floor spreading profiles in the Late Mesozoic, to name a few examples. It is, however, linearly related to geologic time in millions of years. This allows for gradual changes over time in the original hypothetical process on the basis of which the initial relative time scale is constructed. For example, deviations from a straight line on a fitted spline curve may represent corrections of changes in rate of sedimentation, evolutionary change, or seafloor spreading.

In GTS2004, distances between successive points along the x -axis in x - y plots were not zero or close to zero. In the spline fitting, separate cubic polynomials are fitted between pairs of points, taking care that the resulting spline curve is continuous not only in age but also in its first derivative (changes in age). If distances between successive points are either zero or negligibly small, cubic polynomials cannot be fitted. Situations of this type arise when successive ash layers, which are close to one another in the field, are dated separately. This can result in problems in the spline fitting, especially when the leaving-one-out method is used for cross-validation. Such problems can be avoided by calculating the weighted average of dates with exactly or approximately the same values along the chronostratigraphic scale.

Smoothing Spline Software

In our GTS2020 applications, use was made of the *R*-program *smooth-spline* as implemented by B.D. Ripley and

M. Maechler in R. Stats Package version 3.6.0, with uncertainties based on total variances $s_t^2(y)$ or $s^2(y)$ if stratigraphic uncertainty could be neglected (also see next section). This program is freely available on the internet and widely used. Originally, the technique used now was based on code in the GAMBIT FORTRAN program by Hastie and Tibshirani (1990) using a smooth-spline function similar to the one described in Chambers and Hastie (1992).

The default application in R *smooth-spline* is application of a technique called generalized cross-validation (GCV). This technique was originally introduced by Wahba (1985) as a possible refinement of leaving-one-out cross-validation (CV). Differences between CV and GCV are discussed by Wang (2011). Usually the two methods give approximately the same results. However, in several of our applications to Paleozoic datasets, CV provided better results than GCV. This is because it is more robust in situations where relatively many points along the x-axis are close together or coinciding, even after application of the weighted average method described in the previous section. In such situations, CV produces a smooth curve whereas the GCV result may default to a set of straight-line segments connecting neighboring points. Best-fitting smooth curves were obtained by CV in all Paleozoic applications for GTS2020. Because of sparsity of dates for the Devonian, the spline curve for this period was assigned more uncertainty associated with it than the spline curves for the other periods.

As noted before, in some of our *smooth-spline* applications the initial output contained a warning that results (automatic *smooth-spline* estimation of CV) were not necessarily accurate because locally one or more successive x-values were equal or nearly equal to one another. In such situations, we have combined successive values (with the same position along the chronostratigraphic axis) with one another and the *smooth-spline* warning sign disappears. The method used to combine values with different values and their variances is based on elementary statistics (cf. Agterberg et al. 2020).

Devonian Example

Figure 4 shows the Devonian spline obtained for GTS2020 (Becker et al. 2020). It has significantly more associated uncertainty than the GTS2020 spline curves for the other Paleozoic periods. Table 1 shows corresponding input and output information. Almost all age dates are from U-Pb analyses using the TIMS method, and have less than 0.5 Ma external uncertainty. A number of the high-resolution age determinations are on ash layers with the same location along the horizontal (chronostratigraphic) scale. These ages were combined with one another, which reduces the 30+ radiogenic isotope age dates to 19. The chronostratigraphic

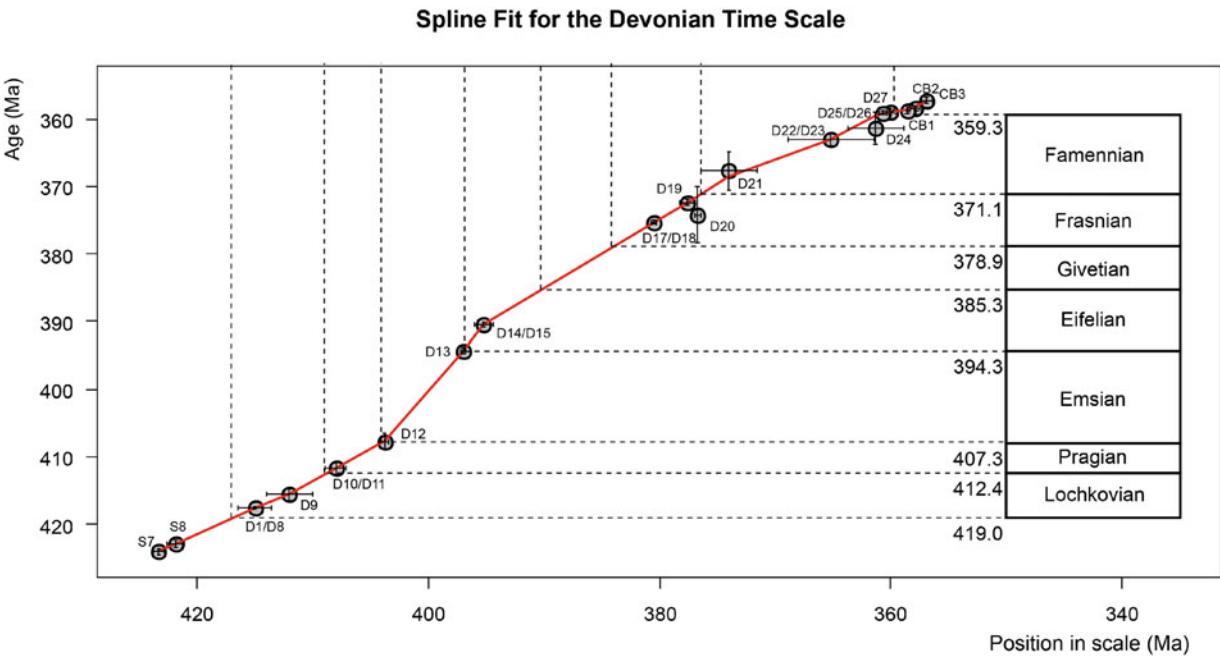
scale is based on thickness of over 20 Devonian conodont zones tentatively converted into Ma units (“Becker scale”). Several of the spline input values have significant uncertainties along the Becker scale, the result of uncertainty in linking some lower-resolution age dates to a specific conodont zone. Thus, total uncertainty associated with every input value is shown with a cross in Fig. 4. In general, these crosses should intersect the spline curve because they represent 95% confidence intervals along the two scales. Along the vertical scale, the underlying frequency distribution is assumed to be normal (Gaussian). For the horizontal scale it is assumed to be rectangular in shape. The weights in Table 1 assigned to the input data were obtained from variances $s_t^2(y) = s^2(x) + s^2(y)$. For other Paleozoic, $s^2(y)$ was used because stratigraphic uncertainty could be neglected.

The spline of Fig. 4 is based on 19 age points. Because there are 7 Devonian stages the number of age points expected to occur in two stages is $2 \times 19/7 = 5.43$. The Givetian and about half of the Frasnian and half of the Eifelian are without any age points. Using the binomial probability distribution, it can be shown that the probability of two adjoining stages being without any age points is about 0.0003. This is unlikely to occur if the age points are randomly distributed over geologic time. It is probable that within as well as immediately before and after the Givetian no or fewer ash layers were deposited. Such a relative lack of input dates would result in relatively greater uncertainty of the estimated spline values. In Agterberg et al. (2020), it is discussed in detail that uneven frequency distribution of input ages through time constituted a third source of uncertainty for Devonian stage boundary age determinations in GTS2012, in addition to the age determination and the chronostratigraphic uncertainties. Normally, the 95% half-confidence interval of a best-fitting spline curve is approximately hyperbolic in shape, widening toward the period boundaries, but for the Devonian there is extra widening of it at and near the Givetian because of the uneven frequency distribution of dates through time.

Curve fitting for the other Paleozoic periods is rather straightforward in that various techniques (smoothing spline, polynomial regression, or LOESS) yield more or less the same solutions. Moreover, 2-sigma values for successive age determinations are similar and it is possible to interpolate between these values to obtain reasonably good estimates of the 95% confidence intervals for the estimated stage boundary ages.

Paleozoic “Super-Spline”

The final GTS2020 age estimates for the Paleozoic stage boundaries are shown in Table 2 with their 2-sigma uncertainties. In this section a single spline curve is fitted to all age



Geologic Time Scale Estimation, Fig. 4 Devonian spline curve fitting results (based on GTS2020 Figure 22.14, Gradstein et al., 2012)

Geologic Time Scale Estimation, Table 1 Comparison of GTS2020 Devonian age determinations with estimated spline values; $2\sigma'$ represent estimated spline-value uncertainty

			Devonian		Spline	
No.	Age (Ma)	$\pm 2\sigma$	x-axis	Weight	Age (Ma)	$\pm 2\sigma'$
CB3	357.26	0.42	356.9	15.4	357.30	0.87
CB2	358.43	0.42	357.9	17.4	358.37	0.85
CB1	358.71	0.42	358.5	21.1	358.69	0.85
D27	358.89	0.48	360.1	15.4	358.90	0.84
D25/D26	359.11	0.30	360.7	38.0	359.15	0.84
D24	361.30	2.40	361.5	0.5	359.81	0.83
D22/D23	363.06	0.88	365.3	3.2	363.03	0.68
D21	367.70	2.80	375.5	0.4	369.13	0.98
D19	372.36	0.41	377.5	13.9	372.36	1.15
D20	374.20	4.20	376.7	0.2	371.15	1.16
D17/D18	375.32	0.26	380.3	52.3	375.38	1.10
D14/D15	390.47	0.34	395.9	18.2	390.58	1.45
D13	394.29	0.47	397.1	17.1	394.28	0.91
D12	407.75	1.16	402.0	2.9	407.73	0.97
D10/D11	411.64	0.88	408.0	3.3	411.73	1.27
D9	415.60	0.80	412.0	2.0	414.77	1.37
D1/D8	417.53	0.19	415.0	29.5	417.53	1.42
S8	422.91	0.49	421.8	9.4	422.90	1.57
S7	424.08	0.53	423.3	11.0	424.09	1.72

Source: GTS2020 Table 14.3

determinations for the Paleozoic by combining the chronostratigraphic scales for different periods. GST2020 results were obtained from spline curves fitted to the Ordovician-Silurian (OS), Devonian, and Carboniferous-Permian (CP), respectively. The Ordovician-Silurian and

Carboniferous-Permian splines are shown in Figs. 20.15c and 23.8 of Gradstein et al. (2020), respectively. The three separate horizontal chronostratigraphic scales for these splines are different but can be combined with one another into a single scale as follows.

Geologic Time Scale Estimation, Table 2 Comparison between GTS2012, GTS2020, and “Super-spline” age estimates (Ma). Last column shows difference between GTS2020 and “Super-spline”

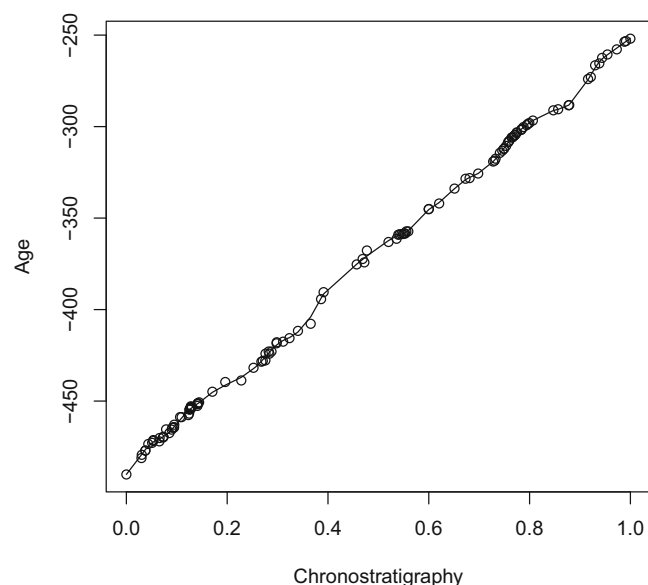
Period	Age/stage	GTS2012	GTS2020	2-Sigma	Super-spline	Difference
Permian	Changhsingian	254.2	254.2	0.4	254.4	−0.2
	Wuchiapingian	259.81	259.5	0.4	259.5	0
	Capitanian	265.14	264.3	0.4	264.3	0
	Wordian	268.8	269.2	0.4	269.3	−0.1
	Roadian	272.3	274.4	0.4	274.4	0
	Kungurian	279.33	283.3	0.4	283.3	0
	Artinskian	290.06	290.5	0.4	290.5	0
	Sakmarian	295.53	293.5	0.4	293.5	0
	Asselian	298.88	298.9	0.4	298.9	0
Carboniferous	Gzhelian	303.67	303.7	0.4	303.7	0
	Kasimovian	306.99	307	0.4	307	0
	Moscovian	315.16	315.2	0.4	315.2	0
	Bashkirian	323.23	323.4	0.4	323.5	−0.1
	Serpukhovian	330.92	330.3	0.4	330.4	−0.1
	Visean	346.73	346.7	0.4	346.8	−0.1
	Tournaisian	358.94	359.3	0.3	358.8	0.5
Devonian	Famennian	372.24	371.1	1.1	371.8	−0.7
	Frasnian	382.69	378.9	1.2	378.1	0.8
	Givetian	387.72	385.3	1.2	384.9	0.4
	Eifelian	393.25	394.3	1.1	393.8	0.5
	Emsian	407.57	410.5	1.1	410.3	0.2
	Pragian	410.78	412.4	1.1	413	−0.6
	Lochkovian	419.2	419	1.8	418.1	0.9
Silurian	Pridoli (Epoch)	422.96	422.7	1.6	422.2	0.5
	Ludfordian	425.57	425	1.5	424.1	0.9
	Gorstian	427.36	426.7	1.5	425.8	0.9
	Homerian	430.45	430.6	1.3	430.3	0.3
	Sheinwoodian	433.35	432.9	1.2	432.9	0
	Telychian	438.49	438.6	1.1	438.6	0
	Aeronian	440.77	440.5	1	441	−0.5
Ordovician	Rhuddanian	443.83	443.1	0.9	443.1	0
	Hirnantian	445.16	445.2	0.9	445.2	0
	Katian	452.97	452.8	0.7	452.8	0
	Sandbian	458.36	458.2	0.7	458.2	0
	Darriwilian	467.25	469.4	0.9	469.3	0.1
	Dapingian	469.96	471.3	1	471.3	0
	Floian	477.72	477.1	1.2	477	0.1
	Tremadocian	485.37	486.9	1.5	486.3	0.6

The Devonian chronostratigraphic scale of Fig. 4 is taken as a starting point. In order to make the OS and CP chronostratigraphic scales compatible with it, two pairs of values were used: the base OS with age of 491.30 Ma on the GTS2020 time scale and CONOP value of 3.96 on the GTS2020 chronostratigraphic scale, and top CP with GTS2020 age of 251.22 Ma and 2658.97 on its GTS2020 chronostratigraphic scale. Because of this assumption, the original OS chronostratigraphic scale x' can be made compatible with the Devonian chronostratigraphic scale x by the linear transformation: $x = -489.37 - 0.7385 \cdot x'$. Likewise, the original CP chronostratigraphic scale can be subjected to the linear

transformation $x = -425.38 - 0.0655 \cdot x'$. These two linear transformations result in a single chronostratigraphic scale for the entire Paleozoic.

Figure 5 shows the result of spline fitting applied to the transformed dataset. The initial choice of using the Devonian chronostratigraphic scale x is arbitrary because any other linear scale could have been used for ensuring compatibility of the three original Paleozoic chronostratigraphic scales. For this reason, the unit for the horizontal scale in Fig. 5 was linearly transformed setting 0 for top of the Permian and 1 for base of the Ordovician.

The “super-spline” of Fig. 5 is used to obtain new estimates of all Paleozoic stage boundaries shown in Table 2 for comparison with the GTS2020 Paleozoic stage boundaries. Differences between the two types of estimates are given in the last column of Table 2 for comparison with GTS2020 2-sigma uncertainty estimates. These results show that only one of the estimated Paleozoic super-spline falls outside the ± 2 -sigma error band for its GTS2020 stage boundary



Geologic Time Scale Estimation, Fig. 5 Paleozoic “super-spline” as obtained by using the R-program *smooth-spline*

Geologic Time Scale

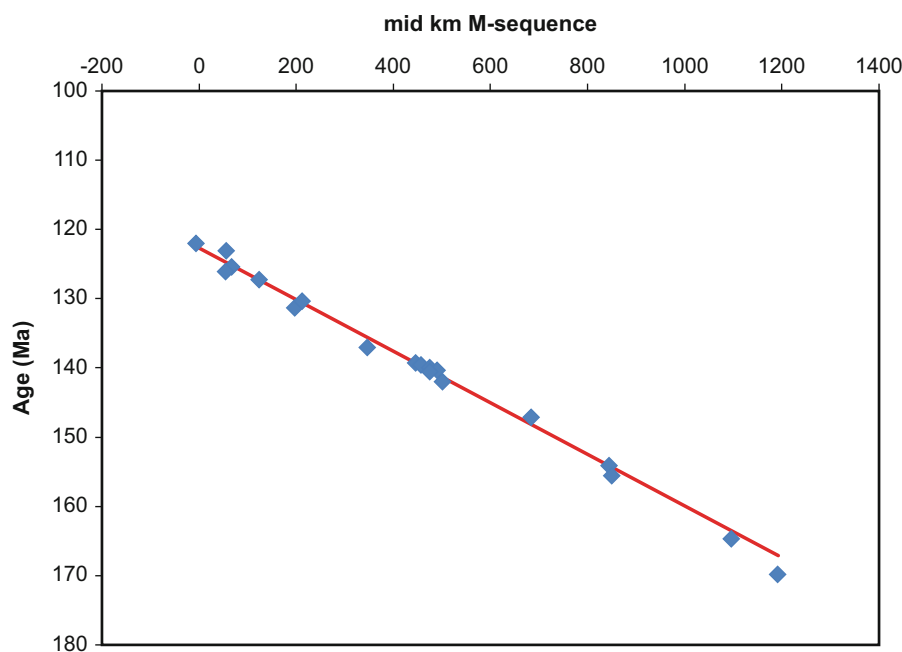
Estimation, Fig. 6 Best-fitting spline curve for the Late Jurassic – Early Cretaceous

estimate (base Tournaisian = Devonian-Carboniferous boundary). This indicates that cubic spline fitting is a flexible statistical method which is not subject to significant changes in number of age determinations per unit of time as occurred during the Paleozoic. The standard deviation of the differences in the last column of Table 2 is 0.80. This would be equivalent to an overall 2-sigma value of 1.6 Ma exceeding most 2-sigma values for the three separate splines used for GTS2020. It remains an open question whether or not the super-spline values of Table 2 are slightly better than the GTS2020 estimates.

Cretaceous and Jurassic Periods Cubic Spline

Spline fitting as applied to Paleozoic Periods in GTS2020 was also used to help obtain the GTS2020 Cretaceous and Jurassic time scales. The critical factor for the Early Cretaceous scale in GTS2020 is that a rather high-resolution U-Pb radiometric, geomagnetic, and cyclostratigraphic dataset is now available for the Tithonian through Barremian, not known during construction of GTS2012 (see discussion in Gradstein et al. 2020, Chapter 27).

Figure 6 is a plot of radiometric age against mid km M-sequence values that span 1198.2 km in total. In order to incorporate stratigraphic uncertainty, $s(x)$ was estimated for each distance value (x) by multiplying its sigma value by the ratio of total time interval (= 47.8 my) and the span of 1198.2 km in order to make use of the total variance equation: $s_t^2(y) = s^2(x) + s^2(y)$. The resulting smoothing spline is



almost exactly a straight line representing constant seafloor spreading. The spline curve in Fig. 6 is approximately a straight line with the equation $y = 0.0372 \cdot x + 122.77$, a result that can also be obtained by using the method of ordinary least squares. Degree of fit for this line could not be improved significantly by including higher-order terms in the polynomial equation fitted by the method of least squares.

Already Kent and Gradstein (1985) in their Cretaceous and Jurassic geochronology study for the Decade of North America Geology publications arrived at constant and linear spreading of the M-sequence as a reasonable template to interpolate the Oxfordian through Barremian time scale. Despite use of few tie points for the age versus M-sequence km plot, the Jurassic-Cretaceous boundary was interpolated by Kent and Gradstein (1985, 1986) at 144 Ma, only slightly older than the current 143.1 ± 0.6 Ma.

GTS2012 and GTS2020 differ from earlier geological time scales in that the ages post-Cretaceous stage boundaries are based on astrochronology-geochronology intercalibration. Increasingly, new information on Milankovitch cycles is becoming available for earlier stages. Ages of Late Cretaceous stage boundaries in GTS2020 are entirely based on Milankovitch cycles. The spline curve in Fig. 6 was used to obtain estimates of the Late Jurassic and Early Cretaceous stage boundaries. However, additional information on durations of Milankovitch cycles available for most of these stage boundaries was used to refine the estimates based on the spline curve of Fig. 6 (cf. Agterberg et al. 2020).

Summary

In our GTS2020 applications to estimate the age of stage boundaries use was made of the *R*-program *smooth-spline* with cross-validation (CV). Best-fitting smooth curves were obtained by CV in all Paleozoic applications for GTS2020 using a large number of radiogenic isotope age dates along the linear scale axis and detailed composite standards to independently scale the stages and zones along the chronostratigraphic axis. A specially constructed Paleozoic “super-spline” that combined the three individual spline curves for Ordovician-Silurian, Devonian, and Carboniferous-Permian did not significantly change the age estimates in GTS2020, and it remains an open question whether or not the super-spline values are slightly better than the GTS2020 estimates. Spline fitting as applied to Paleozoic Periods in GTS2020 was also used to help obtain the GTS2020 Cretaceous time scale. The critical factor for the Early Cretaceous scale in GTS2020 is that a rather high-resolution U-Pb radiometric, geomagnetic, and cyclostratigraphic dataset is now available for the Tithonian through Barremian, not known during construction of GTS2012. Information on durations of Milankovitch

cycles available for most of the Cretaceous stage boundaries was used to refine the estimates based on the spline curve.

Cross-References

- [Geomathematics](#)
- [Quantitative Stratigraphy](#)
- [Splines](#)

References

- Agterberg FP (1994) Estimation of the Mesozoic geological time scale. *Math Geol* 26:857–876
- Agterberg FP (2002) Construction of geological time scales. *Terra Nostra-Schriften der Alfred-Wegener-Stiftung*:227–232
- Agterberg FP, Gradstein FM, Da Silva AC (2020) Geomathematical and statistical procedures. In: Gradstein FM, Ogg JG, Schmitz MD, Ogg GM (eds) *The geologic time scale 2020*. Elsevier, Amsterdam, pp 402–424
- Agterberg FP, Gradstein FM, Gale AS (2021) Estimation of the ages of Devonian and cretaceous stage boundaries in the geologic time scale. *Boletín Geológico y Minero* 131(2). <https://doi.org/10.21701/bdgeom.121.2.006>
- Becker RT, Marshall JEA, Da Silva A-C (2020) Devonian. In: Gradstein FM, Ogg JG, Schmitz MD, Ogg GM (eds) *The geologic time scale 2020*. Elsevier, Amsterdam, pp 733–810
- Chambers JM, Hastie TJ (1992) *Statistical models in S*. Wadsworth and Brooks/Cole, London
- Cooper RA (1999) The Ordovician time scale – calibration of graptolite and conodont zones. *Acta Univ Carol Geol* 43(1/2):1–4
- Gradstein FM, Ogg JG, Schmitz MD, Ogg GM (2020) *The geologic time scale 2020*. Elsevier, Amsterdam, p 1357
- Hastie TJ, Tibshirani RJ (1990) *Generalized adaptive models*. Chapman/Hall, London
- Holden NE, Bonardi ML, De Bièvre P, Renne PR, Villa IM (2011) IUPAC-IUGS common definition and convention on the use of the year as a derived unit of time (IUPAC recommendations 2011). *Pure Appl Chem* 93:1159–1162
- Holmes A (1947) The construction of a geological time-scale. *Trans Geol Soc Glasgow* 21:117–152
- Holmes A (1960) A revised geological time-scale. *Trans Edinb Geol Soc* 17:183–216
- Holmes A (1965) *Principles of physical geology*. Nelson Printers, London, p 1288
- Kent DV, Gradstein FM (1985) A cretaceous and Jurassic geochronology. *Geol Soc Am* 96:1419–1427
- Kent DV, Gradstein FM (1986) A Jurassic to recent geochronology. In: Vogt PR, Tucholke BE (eds) *The geology of North America. Vol. M, the Western North Atlantic region*. Geological Society of America, Boulder, pp 45–50
- McKerrow WS, Lambert RSJ, Cocks LRM (1985) The Ordovician, Silurian and Devonian periods. In: Snelling, N.J. (ed), *the chronology of the geological record*. *Geol Soc London, Mem* 10:73–80
- Ripley BD, Thompson M (1987) Regression techniques for the detection of analytical bias. *Analyst* 112:177–183
- Wahba G (1985) A comparison of GCV and GNL for choosing the smoothing parameters in the generalized spline smoothing problem. *Ann Stat* 4:1378–1402
- Wang Y (2011) *Smoothing splines. Methods and applications, monographs on statistics and applied probability, vol 121*. CRC Press, Boca Baton, 370 pp

Geomathematics

Frits Agterberg
Central Canada Division, Geomathematics Section,
Geological Survey of Canada, Ottawa, ON, Canada

Definition

In its broadest sense, “geomathematics” includes all applications of mathematics to studies of the Earth excluding mathematics in geophysics. A synonym used more frequently is “mathematical geoscience.” Authors, teachers, and editors using the term “geomathematics” include Griffiths (1966a), Osborne (1969), Craig and Labovitz 1981, Merriam 1988, Zhao (1988), Bonham-Carter and Cheng 2008, and Agterberg (1974, 2014). These authors confined the term to applications involving geological studies of parts of the Earth’s crust. The modeling of geological processes that took place millions of years ago almost invariably requires consideration and estimation of uncertainties which are to be captured numerically by using probabilities. For this reason, most geomathematical applications continue to make extensive use of methods originally developed by mathematical statisticians. Recently, “big data” also has come to the foreground involving important new contributions by computer scientists (Zhao and Agterberg 2018).

Introduction

This author became involved in geomathematics because of his interest in mathematical statistics as a graduate student at the University of Utrecht. He joined the Geological Survey of Canada in 1962 as a “petrological statistician.” In 1965, he became the “geomathematician” in a section called “Geomathematics and Data Processing,” and in 1967, head of a “Geomathematics Section” in Ottawa which he headed until his retirement in 1996, although by then the section’s name had been changed into “Mathematical Applications in Geology.” As a term, “geomathematics” has never become as firmly established as the terms “geophysics” or “geochemistry.” “Mathematical geoscience” (formerly also known as “mathematical geology”) is a more widely used synonym of geomathematics. Outside the field of geology, Freedman et al. 2015 use the term “geomathematics” for mathematical applications restricted to the atmosphere. Also, in a different field of science, photogrammetrical engineers in France had the idea of abbreviating “géomatématique” to “géomatique” (Paradis 1981). This term was translated into “geomatics” in English and, since the 1980s, has become widely used for the treatment of (non-geological) geographical data and information.

The aim of this entry is to provide a historical perspective of geomathematics. Involvement by mathematical statisticians is discussed. Subsequently, applications more widely practiced in a number of different geoscience fields will constitute the main part of this entry. On the whole, the remarks to be made are of a historical nature although new challenges for future developments are outlined as well.

Relation of Geomathematics to Mathematical Statistics

Because of limited exposure of bedrock consisting of different rock types that were formed at different times during the history of the Earth, quantification of geological data generally is difficult. Originally most geologists, therefore, developed qualitative methods of approach using two- and three-dimensional graphics without the use of mathematics for geological problem-solving. According to Ronald Fisher (1954), geology with Lyell (1833) was originally evolving as a more quantitative science but, rapidly, opposition against this development grew to the extent that Lyell’s tables and statistical arguments for his subdivision of the Tertiary, which comprised 60 pages, were omitted from later editions of his book entitled *Principles of Geology*.

Until about 50 years ago, most geologists argued that the use of mathematical formulae is not appropriate for connecting the many different observations that can be made in outcrops of bedrock. They used qualitative conceptual models for guidance in their endeavors. As in the past, geological observations still have to be combined with one another to produce geological maps and their three-dimensional downward projections into the Earth’s crust. A good conceptual understanding of the mechanisms underlying the many different types of geological processes involved, therefore, remains essential. The results of different types of geological processes are often superimposed upon one another.

In spite these difficulties, the use of mathematics in geology gradually became more widespread along with the advance of scientific methods including the determination of chemical composition and numerical age of rock samples, and with the advent of digital computers. Mathematical geoscience commenced in earnest with the work of Krumbein (1936, 1939), Chayes (1956), Kolmogorov (1951), and Vistelius (1949). Geomathematical research by these pioneers and their successors included the modeling of processes that took place in geological time measured in millions of years. Merriam and Howarth (2004) arranged for the publication of biographical articles on the pioneers Matheron, Griffiths, Chayes, Reymont, Krumbein, and Vistelius in a special edition of *Earth Sciences History*.

An important development in geomathematics was the establishment in 1968 of the International Association for Mathematical Geology (IAMG). The initiative to create this new organization was taken by Richard Reymont of the University of Uppsala and Daniel Merriam at the Kansas Geological Survey. IAMG history is reviewed in several chapters of the handbook (Daya Sagar et al. 2018, eds.) that was published on the golden IAMG anniversary. The founders conceived this organization to be the offspring of two parents: International Union of Geological Sciences (IUGS) and International Statistical Institute (ISI). As a successful example to follow, they took the biometric society of quantitative biologists which was already in existence and continues to have a strong mathematical component.

Involvement by Mathematical Statisticians

When consulted by Reymont and Merriam, mathematical statisticians including John Tukey and M.S. Bartlett strongly supported the initiative to establish the IAMG. The mathematical statistician Geoffrey Watson became the first IAMG Vice President. He served on the executive committee of the organization. In 2007, the IAMG was renamed International Association for Mathematical Geosciences in recognition of the fact that many members are in disciplines other than geology.

Geomathematics is fortunate that influential mathematical statisticians of the last century were keenly interested in geology. As explained by Fisher Box (1978, p. 440), Ronald Fisher commenced giving regular talks on continental drift lamenting that geophysicists and geologists did not take seriously the theory of continental drift proposed by Alfred Wegener in 1912. Fisher (1954) made an important contribution to this theory by publishing a paper on statistical analysis of measurements of magnetization directions of basaltic lava flows (Fisher 1954). This technique allowed reconstruction of the movements of tectonic plates through geologic time. Geoscientists only accepted these ideas in the mid-1960s.

In 1966, the GSC allowed me to participate in the Advanced Statistical Seminar at the University of Wisconsin. During the Icebreaker, I was introduced to John Tukey who told me about his interest in geology. At this seminar, he presented “The Fast Fourier Transform, for Fun and Profit” (cf. Cooley and Tukey 1965). Back in Ottawa, I received a box filled with about 2000 IBM cards for running the FFT in one, two, or three dimensions on our mainframe computer. During the next 25 years, Tukey commented on my projects at the GSC in three of the approximately 800 publications he authored or co-authored (cf. Agterberg 2018a). Like Matheron, he was recognized at the 2001 ISI Congress in Seoul as one of the greatest mathematical statisticians alive during the second part of the twentieth century.

With Watson who had become Chair of the Princeton University Statistics Department, where Tukey was a professor, he attended the 1969 Geostatistics Colloquium organized by Dan Merriam in Lawrence, Kansas, that also had Matheron, Krumbein, and Serra as participants. Watson owned a cottage on Blood Hill near Elizabethville in the Adirondacks, New York State, not too far from Ottawa. In those days, the GSC maintained a pool of cars with the words “Geological Survey of Canada” in big letters on the sides. I could use one if these cars to visit Watson during weekends. On one occasion I drove Geof and some of his family members to Princeton where Tukey spotted us on the campus. He started laughing and pointing his finger at Watson suggesting that Geof had become a “geologist.” Watson stimulated me to improve my mathematical skills. Pointing out some errors in a review of Agterberg (1974) he had, somewhat sarcastically, remarked that one could see I was not trained as a mathematical statistician. However, he would have granted me an MSc degree in this discipline.

Subsequently, I worked hard on my mathematics. In 1983, I organized a geomathematical workshop at the GSC in Ottawa at which Geof Watson, Fred Bookstein, Jean Serra, and Benoit Mandelbrot were among the presenters. Mandelbrot who had coined the word “fractal,” like Matheron, originally was a student of Paul Lévy at the École Polytechnique in Paris. Both Matheron’s and Mandelbrot’s theoretical developments contain elements of Lévy’s approach to mathematical statistics. Other participants included the directors of Carleton University’s Centre of Mathematical Statistics who shortly afterwards invited me to become an Adjunct Professor in their Mathematics Department. Personally, I have always felt that mathematical statistics offered more challenges than conventional geology, although this remains a scientific discipline in its own right.

Relation of Geomathematics to the Geosciences

Applications in nine different geoscience subfields will be discussed in the next subsections concerned with, respectively: geostatistics, mineral resource estimation, compositional data analysis, directional features, mathematical morphology, quantitative stratigraphy, geologic time scale estimation, and fractals and multifractals and lognormal and Pareto frequency distributions. These are all areas where mathematics makes major contributions.

Geostatistics

Current applications of geostatistics and kriging are adequately explained in other entries in this encyclopedia. Some historical comments are as follows: Originally, Georges Matheron (1955a, 1963) made an important contribution to geoscience by introducing the idea of regionalized random

variables that randomly change their values in three-dimensional space. It is not widely known that Matheron (1955b) started his career in 1954 as a geologist in the French Geological Survey (BRGM). One of his first publications (Matheron 1955b) is concerned with the Gara Gjbilet oolitic iron deposit in Algeria. It is a standard geological publication with a folded geological map in the back and detailed descriptions of the stratigraphy, structure, and genesis of this deposit of Early Devonian age. It can be assumed that, indirectly, the spatial continuity of bedrock probably led Matheron to largely disregard the use of most statistical methods that were in existence at that time.

Originally writing in French, Matheron coined the term “géostatistique,” later translated into “geostatistics” in English. This term became more widely used than “geomathematics.” Although geostatistics is a form of mathematical geoscience, its practitioners are to a large extent mining engineers who apply geostatistical techniques for problem-solving in three-dimensional space, generally without consideration of the various processes of change that may have taken place during geologic time when the objects of study were created. Matheron (1955a)’s first geostatistical paper accompanies a translation into French of Krige (1951)’s pioneering quantitative valuation of Witwatersrand gold mines in South Africa. Somewhat later this led Matheron (1963) to introduce the term “krigeage” that in English became “kriging.”

Mineral Resource Estimation

After geostatistics, that is mainly concerned with ore reserve estimation, mineral resource estimation continues to be a topic of primary interest in geomathematics. Originally, Allais (1957) assumed that (1) the spatial distribution of large mineral deposits in the Sahara Desert satisfies the Poisson distribution for number of deposits per unit area and (2) the total dollar values of these deposits satisfy a lognormal distribution. Griffiths (1966b) pointed out that mineral deposits tend to cluster and proposed that the negative binomial distribution provides a better model than the Poisson distribution. Various early models along these lines were reviewed by Harris (1984) including a novel approach by Brinck (1974) that can be characterized as multifractal (see section on Fractals and Multifractals). It is noted here that, contrary to the Poisson or negative binomial distribution, fractal or multifractal distributions for point patterns do not have a constant mean value for number of points per unit of area. This topic will be discussed in more detail in the section on fractals and multifractals.

Other methods of mathematical statistics modified for use in the geosciences continue to play an important role in quantitative mineral resource evaluation. An objective of this branch of geoscience is to produce mineral potential maps representing, in one way or another, the probabilities

of occurrences of undiscovered mineral deposits of different types and their characteristic features such as size and grade. There usually are great uncertainties associated with the results of these efforts. It is well-known that the process of searching for an economically viable new ore or hydrocarbon discovery often resembles the process of looking for a needle in a haystack.

However, different types of mineral deposits including oil and gas fields are the end products of processes that not only required favorable environments for hosting the ore but also probably left important clues in the vicinities of the deposits themselves. These clues include geophysical and geochemical anomalies. It is the task of the quantitative geoscientist to evaluate in probabilistic terms the mineral potential of large or small study areas. The geo-input for this work consists of geological, geophysical, geochemical, and remote sensing data. These need to be transformed into variables relevant to the occurrences of the known deposits so that probabilities of occurrence of both known and unknown deposits can be estimated. One method for this is Weight-of-Evidence modeling (Agterberg 1989; Bonham-Carter 1994; Cheng 2008). This method quickly results in probabilistic maps with potential targets for further mineral exploration in places with no or fewer than expected ore deposits. Finally, properties of the known deposits such as their sizes and grades could be used to estimate additional expected values that amplify the maps representing the probabilities of occurrence. This presents a new topic for further research (cf. Agterberg 2014, Fig. 4.16).

Compositional Data Analysis

Geological variables, especially for rock composition data, are subject to various constraints that affect the shapes of their frequency distributions and their relations with other variables. Closed-number systems such as the major oxides measured on rock samples provide a primary example as first pointed out by Chayes (1971) showing that silica (SiO_2) content generally is negatively correlated with other major oxides in collections of rock samples, simply because it usually is the dominant rock constituent, and increases in concentration values for other major oxides must be accompanied by decreases in silica. Pearson (1897) had already pointed out that spurious correlations between variables can arise if they are functions of random variables. Aitchison (1986) developed the powerful new methodology of compositional data for variables in closed-number systems. The book by Pawlowsky-Glahn and Buccianti (2011) contains many more recent examples of application.

Directional Features

Almost invariably when one inspects an outcrop of bedrock anywhere, some of the geological features that can be observed show directional features that can be measured. Planar features such as bedding or schistosity have a strike

and dip; linear features such as minor fold axes or non-spherical pebbles have a direction of plunge and dip. These directions often are displayed on geological maps or collectively in diagrams for larger areas. Not surprisingly, much of the statistical theory for treatment of directional features has been developed on the basis of geological data. An early example of this is the study of preferred orientation of pebbles in a conglomerate by Krumbein (1939), who measured the direction in which each pebble was oriented on a scale from 0° to 360°, then doubled the angle, took the average for all pebbles considered, and then divided this average by 2 to obtain a “best” estimate of preferred orientation. This indirect procedure avoids the problem that the direction of elongation of an elongated pebble cannot be distinguished from its counterpart in the opposite direction.

Later, when methods of statistical inference including analysis of variance had become available, geologists, including Potter and Olson (1954), applied these methods to their directional data. However, the methods they used were based on the assumption that the measurements are normally distributed. A problem of application of such methods to directional features is that the circle is used for measurement so that the system is not open but closed. Personally, I became involved in the statistical analysis of directional features when as a graduate student I commenced work on minor fold axes in quartz phyllites in the Italian Alps. Measurements collected in small areas of about 1 km² approximately satisfy a two-dimensional normal distribution on the sphere, which presents a problem of closure analogous to the closure problem for 2-D directional features on the circle. Results of this approach are shown in my PhD thesis (Agterberg 1961) that later was made available on the internet by the University of Utrecht. Subsequent refinements of the methods used to construct regional patterns in this publication are summarized in Agterberg (2014, Chapter 8).

The PhD thesis helped me to obtain a postdoctorate fellowship at the University of Wisconsin in Madison in 1961 where my main task was to help a graduate student to statistically analyze directional measurements of sedimentary features (Agterberg and Briggs 1963). William Krumbein was a reviewer of this paper in which we applied conventional analysis of variance to the data. In a footnote in the manuscript, I had pointed out that this procedure was permissible because for increasingly small variance the circular normal distribution approaches the normal (Gaussian) distribution. Krumbein felt that this statement was a significant finding which deserved further explanation in a section rather than a footnote. Two of his PhD students (Jones and James 1969) continued the use of conventional analysis of variance. However, it was not known to us at the time that Watson (1960) had developed methods of statistical inference including variance analysis providing good approximations based on the isotropic 2-D frequency distribution proposed by Fisher

(1954). Numbers of degrees of freedom in Watson’s methods remain approximate but are better than if the 2-D Gaussian distribution would be used as the basic statistical model. Statistical analysis of spherical data is discussed in detail by Fisher et al. (1987).

Mathematical Morphology

In 1968, Georges Matheron established the Centre de Morphologie Mathématique in Fontainebleau, as a research institute of the École des Mines de Paris. Jean Serra was his close collaborator. Watson has done much to make Matheron’s work in the fields of geostatistics and mathematical morphology better known in the English-speaking world. He persuaded Matheron (1975) to write his book on random sets and integral geometry. At the time Watson told me that there would be only three people in the world able to understand this book from beginning to end.

Since the upper part of the Earth’s crust consists of numerous geological bodies that seem to be in part randomly located with respect to one another, it is somewhat surprising that mathematical geologists have paid relatively little attention to mathematical morphology. The examples in publications by Sagar (2013, 2018) show that good new results can be obtained by increased usage of this methodology.

Quantitative Stratigraphy

Stratigraphy is the branch of geology concerned with the study of rock layers (strata) and layering (stratification) which is a common property of sedimentary and some types of volcanic rocks. Stratigraphy has two related subfields: lithostratigraphy (lithologic stratigraphy) and biostratigraphy (study of layers characterized by fossils). It was mentioned before that Lyell (1833) used a statistical method to subdivide in what was then termed the Tertiary into stages. For large samples, he determined the proportions for species still alive today. For example, the “Pleistocene” contains most species that had not become extinct. This was the earliest example of quantitative biostratigraphy. Early examples of quantitative lithostratigraphy are reviewed by Schwarzacher (1975). This field covers a wide range of methods of time series analysis applied to lithostratigraphic sections.

My book on automated stratigraphic correlation (Agterberg 1990) is mainly concerned with methods to correlate stratigraphic sections at different geographic locations (e.g., using microfossils in drill cores from deep wells drilled in sedimentary basins). One computer program for this method developed in close collaboration with Felix Gradstein was called RASC for ranking and scaling of biostratigraphic events (Gradstein and Agterberg 1982). Among paleontologists, it has led to the term “rascing” (cf. Agterberg and Gradstein 1996), that continues to be used.

Dan Merriam helped to establish several quantitative projects within the framework of the International Geological

Correlation Programme (IGCP) co-sponsored by UNESCO and the IUGS. It included IGCP Project 163 (1977–1984): “IGBA” (IGneous data Base) led by Felix Chayes. IGCP Project 148 (1976–1983): “Quantitative Stratigraphy” was originally led by John Cubitt, but in 1977, I took over as its director. Every year we held an international meeting and there were several national meetings as well. There also was an international quantitative stratigraphy short course informally known as the “traveling circus” that eventually was eventually given in nine different countries. These activities have helped to spread the use of methods of quantitative stratigraphy.

Geologic Time Scale Estimation

The international geologic time scale (GTS), which provides ages measured in millions of years (Ma) with their uncertainties for the boundaries between the approximately 100 stratigraphic stages that have been defined and mutually correlated worldwide, is based on geomathematical methods (Gradstein et al. 2020). The chronogram method originally used by Harland et al. (1982, 1990) for this purpose is suitable for estimation of the age of chronostratigraphic boundaries from a radiometric database when the age estimates of most rock samples used for age determination are subject to significant relative uncertainty. Inconsistencies in the vicinity of chronostratigraphic boundaries can then be ascribed to imprecision of the age determination method.

Cox and Dalrymple (1967) originally developed an approach for estimating the ages of Cenozoic chron boundaries from inconsistent K–Ar age determinations of basaltic rocks. Harland et al. (1982, 1990) adopted this method in their calculations of ages of stage boundaries. The basic principle of the approach is as follows: assuming a hypothetical trial age for an observed chronostratigraphic boundary, rock samples from above this boundary should be younger, and those below it should be older. An inconsistent date is either an older date for a rock sample known to be younger than the trial date, or a younger date for a sample known to be older. The difference between each inconsistent date and the trial age can be standardized by dividing it by the standard deviation of the inconsistent date. Thus, relatively imprecise dates receive less weight than more precise dates. Standardized differences between inconsistent dates and trial age can be squared and the sum of squares (written as E^2) determined for all inconsistent dates corresponding to the same trial age.

The time scale chronograms constructed by Harland et al. (1982) were U-shaped plots of E^2 against different trial ages spaced at narrow time intervals. Optimum choice of age was set at the trial age where E^2 was a minimum. Agterberg (1988) made the following improvement to this method. In addition to inconsistent dates, there are generally more consistent dates for any trial age selected for a chronostratigraphic boundary. The statistical maximum likelihood method was used to

combine consistent with inconsistent dates resulting in an improved estimate of the age of the chronostratigraphic boundary considered. It involves using an iterate process clearly explained by Rao (1973). In this type of calculation, inconsistent dates receive more weight than consistent dates. Generally, the improvement resulting from using consistent dates in addition to the inconsistent dates is relatively minor. However, when there are relatively few dates, use of the consistent dates yields significantly better results. The log-likelihood function is beehive-shaped. For examples, see Gradstein et al. (1995).

A general disadvantage of chronogram methods is that the relative stratigraphic position of the sample is generalized with respect to stage boundaries that are relatively far apart in time. Relative stratigraphic position of one sample with respect to another within the same stage is not considered. A better approach is to incorporate relative stratigraphic positions of fewer samples for which more precise age determinations are available. Current geochronological (U–Pb and Ar–Ar) dating methods have become about 10 times as precise as the earlier methods. It implies that weights in statistical methods given to more recent age determinations are now roughly 100 times greater than in the 1980s. In 1983, the traveling circus mentioned in the preceding section was at the Indian Institute of Technology in Kharagpur. The lecturers included Geof Watson, who in 2 h filled an extra wide blackboard entirely with equations on the relationship between kriging and interpolation splines. It is doubtful that anybody in the audience (including me) could understand what he was talking about. Later I spent significant time comprehending his subsequent paper on the subject (Watson 1984) and I commenced using smoothing splines for geological time scales (Agterberg 1988). Early history of time scale construction methods is reviewed in more detail in Agterberg (2014).

Splining was used extensively for the international geologic times scales published during the past 20 years (Gradstein et al. 2020). Initially, computer programs used to fit smoothing splines were based on FORTRAN code published by de Boor (1978). The leaving-one-out jackknife method was employed to estimate the optimum smoothing factor. This method was also used in GTS2020 but now use is made of the R-program *smooth-spline* as implemented by B.D. Ripley and M. Maechler in R.Stats Package version 3.6.0 made generally available in 2019 (cf. Agterberg et al. 2020).

Fractals and Multifractals

The term “fractal” was introduced by Mandelbrot (1977) to characterize natural phenomena with non-integer dimensions. Other influential books with geoscience applications on this subject were Federer (1988), Korvin (1992), Stoyan and Stoyan (1994), and Turcotte (1997). Later it was found that many phenomena originally thought to be fractals are

multifractals that are physically intertwined fractals with different dimensions (see, e.g., Mandelbrot 1989, or Lovejoy and Schertzer 2007). In his doctoral thesis, Qiuming Cheng (1994; also see Cheng and Agterberg 1996) showed that point patterns previously thought to have a constant mean number of points per unit of area can be modeled as multifractals. In Agterberg (2014, 2018b), it is illustrated that worldwide patterns of ore deposits of different genetic types compiled by Singer and Menzie (2004) for so-called permissive tracts are probably multifractal although they can be approximated by fractals as well.

Mandelbrot (1989) has pointed out that the so-called model of de Wijs (1951) was the first multifractal conceived. This model was taken by Brinck (1974) to describe the worldwide distributions of uranium and several other metals. Later it was used by Agterberg (2018b) for the same purpose.

Lognormal and Pareto Frequency Distributions

Both the lognormal and the Pareto frequency distribution have played important roles in the geosciences, particularly in geochemistry and economic geology. Both are “stable” forms that arise as the end products of generating processes that may involve other types of frequency distributions (Lévy 1925; also see Mandelbrot 1977). In the past, authors have preferred one type of frequency distribution above the other. For example, Turcotte (1997) favors the Pareto distribution, whereas Ahrens (1953) proposed the lognormal as a fundamental law of geochemistry. Both models provided excellent degree of fit for relatively small data sets. During the past 20 years, a new model called the Pareto-lognormal distribution has been developed that is lognormal for most of the data but Pareto in the tails (cf. Kleiber and Kotz 2003).

Patiño Douce (2018) has compiled a large data base of metallic mineral deposits. Various metals including Cu, Zn, Pb, Ni, Mo, Au, and Ag are approximately lognormally distributed but have upper and lower tails that are clearly Pareto type (see, e.g., Agterberg 2020). A major difference between the Pareto-lognormal distributions for metallic mineral deposits and the Pareto-lognormal distributions for incomes described by Kleiber and Kotz (2003) is that in the upper tails where there are fewer than expected large mineral deposits which are smaller than the largest deposits that satisfy the Pareto model. A possible explanation for this departure from Pareto-lognormality is that the largest deposits have better developed geophysical and other anomalies, and were economically the most desirable targets (Agterberg 2020).

Summary

Although the term “geomathematics” has been used by a number of authors, a better-known synonym is “mathematical

geoscience.” In this entry, the emphasis was on historical uses of mathematics in several subfields of the Earth Sciences. It has been pointed out that geologists were slow to adopt mathematical methods in their approach to develop complete two- and three-dimensional representations of the upper parts of the Earth’s crust. However, with the advent of new computer-based methods for graphics and “big data” analysis, the quantitative approach to geological problem-solving is being practiced increasingly widely.

Cross-References

- Compositional Data
- De Wijs Model
- Geomatics
- Geostatistics
- Kriging
- Mathematical Geosciences
- Mathematical Morphology
- Maximum Likelihood
- Mineral Prospectivity Analysis
- Multifractals
- Point Pattern Statistics
- Quantitative Stratigraphy
- Splines

References

- Agterberg FP (1961) Tectonics of the crystalline basement of the dolomites in North Italy. Kemink, Utrecht. <https://dspace.library.uu.nl/handle/1874/316591>
- Agterberg FP (1974) Geomathematics – mathematical background and geo-science applications. Elsevier, Amsterdam
- Agterberg FP (1988) Quality of time scales – a statistical appraisal. In: Merriam DF (ed) Current trends in geomathematics. Plenum, New York, pp 57–103
- Agterberg FP (1989) Computer programs for mineral exploration. Science 245:76–81
- Agterberg FP (1990) Automated stratigraphic correlation. Elsevier, Amsterdam
- Agterberg FP (2014) Geomathematics: theoretical foundations, applications and future developments. Springer, Heidelberg
- Agterberg FP (2018a) Origin and early development of the IAMG. In: Sagar BSD, Cheng Q, Agterberg F (eds) Handbook of mathematical geosciences – fifty years of IAMG. SpringerOpen, Cham, pp 901–913
- Agterberg FP (2018b) Can multifractals be used for mineral resource appraisal? Jour Geochem Explor 189:54–63
- Agterberg FP (2020) Multifractal modeling of worldwide and Canadian metal size-frequency distributions. Nat Resour Res. <https://doi.org/10.1007/s11053-019-09410.1>
- Agterberg FP, Briggs G (1963) Statistical analysis of ripple marks in Atokan and Desmoinesian rocks the Arkoma Basin of central Oklahoma. J Sediment Petrol 33:393–410
- Agterberg FP, Gradstein FM (1996) Recent developments in the rasing of biostratigraphic events. Paleontol Soc 8:3

- Agterberg FP, Gradstein FM, Da Silva AC (2020) Geomathematical and statistical procedures. In: Gradstein FM, Ogg JG, Schmitz MD, Ogg GM (eds) *The geologic time scale 2020*, vol 14. Elsevier, Chapter, Amsterdam, p 1
- Ahrens LH (1953) A fundamental law of geochemistry. *Nature* 172:1148
- Aitchison J (1986) *The statistical analysis of compositional data*. Chapman and Hall, London
- Allais M (1957) Method of appraising economic prospects of mining exploration over large territories: Algerian Sahara case study. *Manag Sci* 3:285–347
- Bonham-Carter GF (1994) *Geographic information systems for geoscientists: modelling with GIS*. Pergamon, Oxford
- Bonham-Carter GF, Cheng Q (eds) (2008) *Progress in geomathematics*. Springer, Heidelberg
- Brinck JW (1974) The geochemical distribution of uranium as a primary criterion for the formation of ore deposits. In: *Chemical and physical mechanisms in the formation of uranium mineralization, geochronology, isotope geology and mineralization*, vol 374. International Atomic Energy Proceedings Series STI/PUB/, Vienna, pp 21–32
- Chayes F (1956) Petrographic modal analysis. Wiley, New York
- Chayes F (1971) Ratio correlation. University of Chicago Press, Chicago
- Cheng Q, Agterberg FP (1996) Multifractal modeling and spatial statistics. *J Math Geol* 28:1–16
- Cheng Q (2008) Non-linear theory and power-law models for information integration and mineral resources quantitative assessments. In: Bonham-Carter GF, Cheng Q (eds) *Progress in geomathematics*, Springer, Heidelberg, pp 195–225
- Cooley JW, Tukey JW (1965) An algorithm for the machine calculation of complex Fourier series. *Math Comput* 19:297–307
- Cox AV, Dalrymple GB (1967) Statistical analysis of geomagnetic reversal data and the precision of potassium-argon dating. *J Geophys Res* 72(10):2603–2614
- Craig RG, Labovitz ML (eds) (1981) *Future trends in geomathematics*. Pion, London
- De Boor C (1978) *A practical guide to splines*. Springer, New York
- De Wijs HJ (1951) Statistics of ore distributions. *Geol Mijnb* 13: 365–375
- Federer J (1988) *Fractals*. Plenum, New York
- Fisher Box J (1978) R.A. Fisher: the life of a scientist. Wiley, New York
- Fisher RA (1954) Expansion of statistics. *Am Sci* 42:275–282
- Fisher NI, Lewis T, Embleton JJ (1987) *Statistical analysis of spherical data*. Cambridge University Press, Cambridge
- Freedman W, Nashed MZ, Sonar T (eds) (2015) *Handbook of geomathematics*, 2nd edn. Springer, Heidelberg
- Gradstein FM, Agterberg FP (1982) Models of Cenozoic foraminiferal stratigraphy – northwestern Atlantic margin. In: Cubitt JM, Reymont RA (eds) *Quantitative stratigraphic correlation*. Wiley, Chichester, pp 119–173
- Gradstein FM, Agterberg FP, Ogg JG, Hardenbol J, van Veen P, Thierry T, Huang Z (1995) A Triassic, Jurassic and Cretaceous time scale. In: Berggren WA, Kent DV, Aubry M-P, Hardenbol J (eds) *Geochronology, time scales and global stratigraphic correlations: a unified temporal framework for a historical geology, Society of economic paleontologists and mineralogists special publication*, vol 54, pp 95–128
- Gradstein FM, Ogg JG, Schmitz MD, Ogg GM (2020) *The geologic time scale 2020*. Elsevier, Amsterdam
- Griffiths JC (1966a) Future trends in geomathematics. *Bul Miner Ind Exp Stn, Pa State Univ* 35(5):1–8
- Griffiths JC (1966b) Exploration for mineral resources. *Oper Res* 14: 189–209
- Harland WB, Armstrong RI, Cox AV, Craig LA, Smith AG, Smith DG (1990) *A geologic time scale 1989*. Cambridge University Press, Cambridge
- Harland WB, Cox AV, Llewellyn PG, Pickton CAG, Smith AG, Walters R (1982) *A Geologic Time Scale*. Cambridge University Press, Cambridge
- Harris DP (1984) *Mineral resources appraisal*. Clarendon Press, Oxford
- Jones TA, James WR (1969) Statistical analysis of orientation data. *J. Int. Assoc Math Geol* 1:129–136
- Kleiber C, Kotz S (2003) *Statistical distributions in economics and actuarial sciences*. Wiley, Hoboken
- Kolmogorov AN (1951) Solution of a problem in probability theory connected with the problem of the mechanism of stratification. *Am Mat Soc Trans* 53:8
- Korvin G (1992) *Fractal models in the earth sciences*. Elsevier, Amsterdam
- Krige DG (1951) A statistical approach to some basic valuation problems on the Witwatersrand. *J South Afr Inst Min Metall* 52:119–139
- Krumbein WC (1936) Applications of logarithmical moments to size frequency distributions of sediments. *J Sediment Petrol* 6:35–47
- Krumbein WC (1939) Preferred orientation of pebbles in sedimentary deposits. *J Geol* 47:637–706
- Lévy P (1925) *Calcul des Probabilités*. Gauthier Villars, Paris
- Lovejoy S, Schertzer D (2007) Scaling and multifractal fields in the solid earth and topography. *Nonlinear Process Geophys* 14:465–502
- Lyell C (1833) *Principles of geology*. Murray, London. (+ 109 pp. supplements)
- Mandelbrot BB (1977) *Fractals – form, chance, and dimension*. Freeman, New York
- Mandelbrot BB (1989) Multifractal measures, especially for the geophysicist. *Pure Appl Geophys* 131:5–42
- Matheron G (1955a) Application des méthodes statistiques à l'évaluation des gisements. *Annales des Mines*, décembre 1955:50–72
- Matheron G (1955b) Les gisements de Gara Djebilet. *Bulletin Scientifique et Économique du B.R.M.A.*, No 2, mai 1955, published by Bureau de Recherches Minières de l'Algérie, 53–63
- Matheron G (1975) *Random sets and integral geometry*. Wiley, New York
- Merriam DF (ed) (1988) *Current trends in Geomathematics*. Plenum, New York
- Merriam DF, Howarth RJ (2004) Pioneers in mathematical geology. *Earth Sci Hist* 23(2):314–432
- Osborne RH (1969) Undergraduate instruction in Geomathematics. *J Geol Educ* 17(4):12–124
- Paradis M (1981) De l'arpentage à la géomatique. *Le Géomètre Canadien* 35(3):262
- Patiño Douce AE (2018) Metallic mineral resources in the twenty first century. II constraints on future supply. *Nat Resour Res* 25:97–124
- Pawlowsky-Glahn V, Buccianti A (eds) (2011) *Compositional data analysis – theory and applications*. New York, Wiley
- Pearson K (1897) On a form of spurious correlation which may arise when indices are used in the measurements of organs. *Proc R Soc London* 60:489–498
- Potter PE, Olson JS (1954) Variance components of cross-bedding direction in some basal Pennsylvanian sandstones of the eastern Interior Basin: geological application. *J Geol* 62:50–73
- Rao CR (1973) *Linear statistical inference and its applications*. Wiley, New York
- Sagar BSD (2013) *Mathematical morphology in Geomorphology and GISci*. CRC Press, Boca Raton, 546 p
- Sagar BSD (2018) *Mathematical morphology in geosciences and GISci: an illustrative review*. In: Sagar BSD, Cheng Q, Agterberg F (eds) *Handbook of mathematical geosciences – fifty years of IAMG*. Cham, SpringerOpen, pp 703–740
- Sagar BSD, Cheng Q, Agterberg F (eds) (2018) *Handbook of mathematical geosciences – fifty years of IAMG*. SpringerOpen, Cham, 914 p
- Schwarzacher W (1975) *Sedimentation models and quantitative stratigraphy*. Elsevier, Amsterdam, 382 p

- Singer DA, Menzie DW (2004) Quantitative mineral resource assessments. Oxford University Press, New York, 219 p
- Stoyan D, Stoyan H (1994) Fractals, random shapes and point fields. Wiley, Chichester, 389 p
- Turcotte DL (1997) Fractals and Chaos in geology and geophysics. Cambridge University Press, Cambridge, 398 p
- Vistelius (1949) On the question of the mechanism of formation of strata. Dokl Akad Nauk SSSR 65:191–194
- Watson GS (1960) More significance tests on the sphere. Biometrika 47: 87–91
- Watson GS (1984) Smoothing and interpolating by kriging and with splines. Math Geol 16(6):601–615
- Zhao P (1988) Review of geomathematical applications for mineral resources evaluation in China. In: Chung CF, Fabbri AG, Sinding-Larsen R (eds) Quantitative analysis of mineral and energy resources, NATO ASI series C: mathematical and physical sciences, vol 223. Springer, Dordrecht, pp 79–87
- Zhao P, Agterberg F (2018) Foreword. In: Sagar BSD, Cheng Q, Agterberg F (eds) Handbook of mathematical geosciences – fifty years of IAMG. SpringerOpen, Cham, pp vii–x

Geomatics

Rifaat Abdalla

Department of Earth Sciences, College of Science, Sultan Qaboos University, Al-Khoudh, Muscat, Oman

Definition

Geomatics is defined as a systematic, multidisciplinary, integrated science that is dealing with systems and technologies for collecting, storing, integrating, modeling, analyzing, retrieving, transforming, displaying, and distributing spatially georeferenced data from different measurements with well-defined precision and accuracy characteristics and continuity and in a digital format. The term geomatics consists of two parts: geo, which means Earth, and “matics” in the sense of informatics. Hence, geomatics refers to the science of geoinformatics. Geomatics is based on a number of basic sciences and technologies, which include computer science, remote sensing, geospatial information systems (GIS), photogrammetry, cartography, global navigation satellite system (GNSS), and decision support systems (DSS) (Barrile et al. 2019).

Geomatic Components

Land Surveying

Survey science is the science that professionals and researchers use to locate the natural and human features present on or under the Earth’s surface and the representation

of these appearances on traditional (printed) or digital maps (using the computer).

There are several divisions of the types of survey applications, whether in terms of the field of use or in terms of the purpose of the survey work or in terms of the survey device used and the survey sections (Deren 2017), which has classified survey into terrestrial survey with subcategories of geodetic and plane survey, photogrammetry, hydrographic survey, and astronomical survey. Figure 1 summarizes the classification of survey operations.

Survey Measurements

Survey measurements are of three types, namely, distance measurements, angle measurements, and contour measurements (elevation). The distance measurement could be measured with a tape or electronically, while the angles can be measured with the help of theodolite equipment. Contours can be measured with level and height and through the leveling process.

Geodesy

Geodesy is the science of measuring and mapping the Earth’s surface. Geodesy is also concerned in determining the shape and size of the Earth, based on accurate measurements such as distance measurements, star positioning, and gravitational force measurements. Geodesy can be categorized into (1) mathematical, (2) satellite, (3) ground measurements, and (4) astronomical.

Photogrammetry

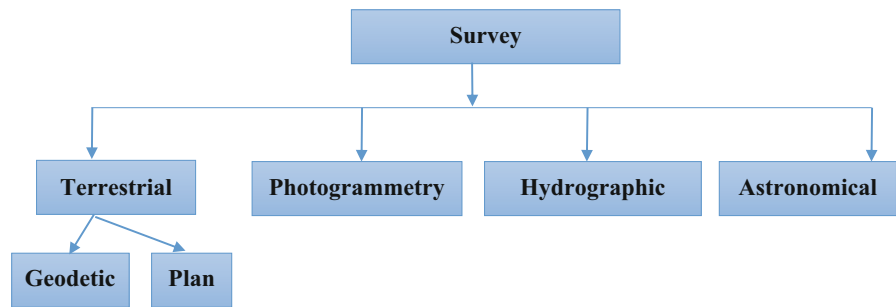
Photogrammetry is a branch of surveying that allows measuring landmarks without approaching them, as measurements are made through images, whether aerial photos or satellite images. According to Smith et al. (2016), photogrammetry is characterized by (a) making accurate measurements, (b) covering a large area of the Earth’s surface (producing maps from aerial photography is less time-consuming and less costly than a ground survey), (c) making stereograms and topographic mapping, (d) tracking time changes, (e) clarifying features and features that the human eye cannot see, (f) highlighting underground features, and (H) accentuating spatial features in remote areas.

Digital photogrammetry is a technological development of traditional photographic scanning methods based on computers in the stages of imaging, analysis, and measurement of aerial photos, and it is characterized by several characteristics, including high spatial resolution, image acquisition speed, automatic operation and analysis, fast image reproduction, and lower cost (Woodget et al. 2017).

Remote Sensing

Remote sensing attempts for deriving information about a specific target or location located at a distance from the

Geomatics, Fig. 1 Classification of survey operations



sensor. Remote sensing is characterized by the ability to sense the same geographical area repeatedly over a period of a span of time ranging from several hours to several weeks. This allows for obtaining updated imagery periodically, thus allowing for following up the dynamic geographical phenomena.

Remote sensing information acquisition consists of three stages, data collection, data processing, and data interpretation. Based on the type of sensor, it is classified into passive sensors, where the source of energy is transmitted from the sensor, and active sensors, in which the sensor uses its own energy. The remote sensors can be divided in terms of the number of electromagnetic energy ranges that are dealt with and recorded, so the sensors can be divided into the following (Steenweg et al. 2017): optical, where there are four types of sensors including panchromatic (black and white), multispectral, super spectral, and hyperspectral. Radio detection and ranging (radar) is a type of the active sensors and is divided into single frequency and multifrequency.

Global Navigation Satellite Systems (GNSS)

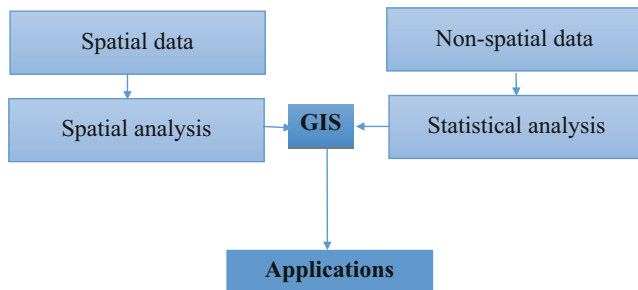
GNSS or global navigation satellite system is the standard generic term for satellite navigation systems that provide autonomous geospatial positioning with global coverage. It offers three-dimensional coordinates of fixed or moving targets anywhere on the Earth and under any climatic conditions. GNSS is integration of different positioning systems, which are the global positioning system (GPS), the global navigation satellite system (GLONASS) developed by Russia, and GALELIO, which is developed by the European Space Agency. The basic idea in these systems depends on the satellites sending radio waves that are received and recorded through special ground receivers and then calculating the coordinates of the position of the receiver, and these calculations depend on knowing the coordinates of the satellite itself, depending on the prior knowledge of the satellite's orbit in space. For instance, the GPS location coordinates are determined by calculating the distance between the device and the satellite by knowing the time between the moment the signal is sent from the satellite and the moment it is received in the receiver (Pasku et al. 2017). The GPS as a major component

of the GNSS has three segments: the space segment, the control segment, and the user segment. It offers two main types of services: standard positioning service and the precise positioning service.

Among the many important applications of the GPS are (1) producing accurate and digital topographic and detailed maps, (2) hydrography and nautical chart development, (3) determining the ground adjustment marks for images and videos, (4) spatial data collection for geographic information systems, and (5) portable map systems.

Geographic Information Systems (GIS)

GISs are an effective set of tools for collecting, storing, analyzing, and displaying spatial data using specific data models that replicates the real world. These models can be for a specific purpose or for a general purpose. The representation of spatial data depends on (1) spatial data that includes the location or coordinates attributed to a specific system, (2) nonspatial data which provide attribute information related to features such as area and type, and (3) spatial relationships between representations or topology of features. GIS is subdivided into different types including (Jia et al. 2017) land information systems (LIS), topographic information systems (TIS), and building information modeling (BIM). There are many advantages of GIS, including (1) merging spatial and nonspatial information into a single database, (2) high ability to analyze spatial and nonspatial data, (3) the ability to visualize spatial information, (4) helping to make decisions, and (5) data documentation with specific specifications. Data sources for GIS are from paper and digital maps, remotely sensed imagery, photogrammetry data, nautical charts, and bathymetric surveys in addition to nonspatial data from statistical sources. The GIS represents the phenomena in a certain spot on the surface of the Earth through several layers. Each layer represents a specific type of geographical phenomena, for example, when representing a neighborhood of a particular city, streets are drawn in a layer, buildings in the second layer, and trees in the third layer. If all these layers are visualized at the same time, a representation of the real scene in this area is obtained. There



Geomatics, Fig. 2 GIS data types and products

are two broad data models in GIS: (a) vector data, which can be found in form of point, line, and polygon, and (b) raster data, which can be formed in pixel-based products. GIS data components and analysis are shown in Fig. 2.

Cartography

Cartography is the science of providing a description of the shape and size of the Earth and its natural and human features and showing this description on maps (Craib 2017). Cartographic tasks are (1) drawing, draw the features and geographical features with precision; (2) representation, the accurate appearances of features (projection); (3) generalization, the selection of features and best mode for representation; and (4) map design, finding the most appropriate way to allow for visualizing the features easily.

According to cartography, a map is a flat and symbolic approximate representation of the surface of the Earth or parts of it. This requires models to represent the shape of the Earth through specific symbols. The elements of map content include (A) basic elements (the title of the map, map legend, coordinate grid, and map projection) and (B) secondary elements (data sources, graphic charts, tables, inset maps, date of the map, and producer of the map). Key components of a map are based on survey (measurement) and on the type and selection of spatial data. It also depends on data standardization and data generalization, converting the points of the surface of the Earth to its equivalent on a reference surface by relying on reference coordinate systems and representing the reference surface of the Earth on the map through mathematical models that allow the representation of landmarks or points on the map or on the computer screen. Perception of the map through symbols is important in allowing for real-world representation.

Digital cartography is distinguished from traditional cartography by the main characteristic of using digital data. This allows for data storage, manipulation, and easier and faster presentation. It provides the ability to control the scale of the map to achieve the desired level of detail. Digital cartography deals with 3D coordinates. It enables the user to quickly automate selection, classification, and statistics and spatial analysis of targets or features. Users are able to add additional

dimensions like the fourth dimension (time) to spatial variables by dealing with different data for the same spatial region, which allows tracking the temporal changes of spatial phenomena in an accurate digital manner.

Innovations in Geomatics

New technologies have influenced the shape of geomatics in the last decade. Although the domain of geomatics in itself is a modern domain, many advancements in technologies have further contributed to the innovation in geomatics. This can be summarized by more portable and handheld geomatic devices, for visualization, field data collection, and integration with other equipment in the field. The second leap in geomatic technologies is advanced integration in instrumentation. Unmanned aerial vehicles (UAVs) can now perform sophisticated geomatic operations through different onboard geomatic sensors, whether it be imaging sensors or surveying sensors, to provide timely solutions on field operations. The third leap on geomatic innovation is the advancement in hydrography and underwater imaging. Hydrography tools and equipment are getting more integrated, more robust, and more affordable, which allow for additional advancement in geomatic applications. Figure 3 is showing new advanced mobile handheld devices for field surveying that integrates GIS, GPS, remote sensing, and land surveying devices in one portable instrument.

Geomatic Applications

There are many applications for geomatics. It can be used in surveying, to produce ground constant network maps whether in the form of single or multiple drawing scale with different geodetic references. The advanced geomatic technology applies high-accuracy and high-precision special surveying equipment, for instance, portable thermal imaging devices and portable laser range finders. Specific applications of remote sensing are numerous; some of the most important are production of detailed and contour maps, urban studies (land use) transportation network planning, follow-up on natural disasters, climatic studies, searching for natural resources, and geological studies. Remote sensing military applications monitor what is happening on the ground in a specific area on a daily basis (or less). The use of UAV drones provides fast acquisition and high-resolution imagery, while some satellites record videos, not just visuals. Remote sensing is one of the most accurate and fastest ways to obtain spatial information. When starting precision agriculture project, the use of remote sensing in determining and studying the qualitative distribution of soil use of RS to develop high-resolution DEM for the region during work. Remote sensing

Geomatics, Fig. 3 Showing new advanced mobile handheld devices for field surveying



can also be used to monitor crop growth throughout the growing season. Using remote sensing is effective to determine proper watering times and plant need for water. Remote sensing is used to monitor soil moisture condition and to track weather, rain, and drought characteristics. Remote sensing is also usable in detection of plant diseases using near-infrared imaging. Estimation of the crop size at the end of the growing season can be achieved using remote sensing.

There are numerous applications for GIS; some of these applications are for (1) creating maps, utilities and services, public facilities, and planning; (2) urban and regional planning; (3) Earth science and the exploration for natural resources; (4) environment-related application; (5) agriculture-related projects; (6) historical, archaeological, and tourism applications; and (7) military applications. Real-time security applications such as police-ambulance-civil defense use real-time detection of sites. It allows for determining the shortest/fastest route for rescue vehicles and observes for huge crowds. Specific applications such as precision agriculture depend on the availability of modern, accurate maps at a

scale suitable for its purposes. Detailed and topographic maps in addition to updated maps of everything that is happening on the ground, after each quarterly crop: update topographic maps (terrain). It depends on the availability of spatial and nonspatial information in a timely manner about plants, soil, and climate in order to provide the farmers with the necessary instructions on soil management and the best time for agricultural management operations to reach the best yield.

Summary

Geomatics is interdisciplinary science that incorporates computer science, where all data sources, processing, and algorithms are based upon. Geodesy, which provides approximation for the shape and area of the Earth or large portions of it. Surveying, whether land surveying or hydrographic surveying in which measurements about the Earth surface and subsurface are taken. Cartography is the science

of making map, whether in the old artistic way or in the new digital way of composing maps. Photogrammetry is the science of taking measurements out of remotely sensed high-resolution aerial photos or imagery. It is of two parts, softcopy and hardcopy photogrammetry. Remote sensing is the science of extracting imagery from the Earth surface, without being in a direct contact with it. GNSS is related to the global navigation satellite system, which comprises the global positioning system (GPS) and other systems that provide accurate location data. Laser scanning system provides 3D surface capture. Geographic information system (GIS) provides a system for managing, processing, and manipulation of georeferenced data. Decision support system (DSS) is used to support decision-makers with efficient and informed aid for effective decision-making process.

Cross-References

- [Digital Geological Mapping](#)
- [Remote Sensing](#)
- [Spatial Data](#)

Bibliography

- Barrile V, Fotia A, Bilotta G, De Carlo D (2019) Integration of geomatics methodologies and creation of a cultural heritage app using augmented reality. *Virtual Archaeol Rev* 10(20):40–51
- Bock Y, Melgar D (2016) Physical applications of GPS geodesy: a review. *Rep Prog Phys* 79(10):106801
- Craib RB (2017) Cartography and decolonization. *Decolonizing the map: cartography from colony to nation*, 11–71
- Deren L (2017) From geomatics to geospatial intelligent service science. *Acta Geodaetica Cartographica Sinica* 46(10):1207
- Jia P, Cheng X, Xue H, Wang Y (2017) Applications of geographic information systems (GIS) data and methods in obesity-related research. *Obes Rev* 18(4):400–411
- Pasku V, De Angelis A, De Angelis G, Arumugam DD, Dionigi M, Carbone P, Ricketts DS (2017) Magnetic field-based positioning systems. *IEEE Commun Surv Tutor* 19(3):2003–2017
- Smith MW, Carrivick JL, Quincey DJ (2016) Structure from motion photogrammetry in physical geography. *Prog Phys Geogr* 40(2):247–275
- Steenweg R, Hebblewhite M, Kays R, Ahumada J, Fisher JT, Burton C, Brodie J (2017) Scaling-up camera traps: monitoring the planet's biodiversity with networks of remote sensors. *Front Ecol Environ* 15(1):26–34
- Wang X, Xie H (2018) A review on applications of remote sensing and geographic information systems (GIS) in water resources and flood risk management
- Woodget AS, Austrums R, Maddock IP, Habit E (2017) Drones and digital photogrammetry: from classifications to continuums for monitoring river habitat and hydromorphology. *Wiley Interdiscip Rev Water* 4(4):e1222

Geomechanics

Pejman Tahmasebi

College of Engineering and Applied Science, University of Wyoming, Laramie, WY, USA

Definition

Geomechanics is a branch of science where the mechanical behavior of rock and soil is studied. More precisely, the deformation of rock/soil is investigated in geomechanics through which one can better manage the risk. This field brings the geology and mechanics of materials together to analyze the physical behaviors and deformations in an analytical way.

Introduction

Geomechanics influence can be seen in a wide range of fields such as radioactive waste disposal, carbon dioxide sequestration in the subsurface to avoid the negative effect of these gas on the atmosphere, natural gas storages, energy resources, mining, and natural hazards (e.g., earthquake, landslide, volcano, and flood). Understanding the mechanics of geomaterials and their application in the service of humanity is vital. In geomechanics, we are interested to describe the behavior of geological environments to internal and external forces, which is often achieved by describing the state of stress and strain history.

It also has been widely shown that the lack of geomechanical considerations has a significant drawback while its absence can cause damages, including financial losses and/or humans life. There are many historical accidents that could be avoided if they were studied geomechanically before it happens. An example is the St. Francis Dam disaster (Begnudelli and Sanders 2007; Rogers 2006); this gravity curved dam was built in 1926 and the building process took 2 years. The height of this dam was 205 feet. In 1928, the dam failed, 432 people died, and property damage was estimated to be around seven million dollars. The insufficient foundation stress support and reactivation of a landslide near the dam were recognized as the cause of the failure, which could be estimated using accurate geomechanical modeling. A couple of other disasters led engineers to recognize the importance of geomechanical studies during the design of projects.

In a similar fashion, subsurface geomechanics has also witnessed huge progress, which refers to cases where the

depth of the case study does not exceed 10 km. Knowledge of the stress in such depths is the essential element that leads engineers to have comprehensive geomechanical models. If the stress concentrated around, for example, the well/tunnel becomes higher than the rock's strength, the rock system will fail. The use of geomechanics, thus, is to estimate where and when the failure happens. Consequently, the associated risk and recommendations can be provided to lower the threat. Similarly, a fault can slip when effective normal stress becomes lower than the exerted shear normal stress. These variations of stress can be initiated from a variety of sources, such as deposition, fluid injection/withdrawal, thermal effects, and other tectonic variations, which altogether vary with time. Thus, geomechanical behaviors should be studied with time and they manifest a dynamic process. The same conclusion is also valid for geomaterials' strength properties as they change with time as well. Given the above-described situation with geological environments, pre-existing stresses always exist. Therefore, it is critical that such variabilities (e.g., stress, strength, pressure) are characterized before making any intervention, which is often referred to as geological history. Based on mentioned facts, we have to know the fundamental stresses to enhance our knowledge for modeling purposes. In subsurface systems, such elements always experience compressive stress, which depends on factors such as pore pressure, depth, and the geological characteristics of soil and rock layers above the element. Briefly, three characteristics are describing the stress field:

- Stress at depth: it is considered as a basic knowledge that could help engineers solve practical problems in geomechanics.
- In situ stress field at depth and scales relation: with analyzing different data from different sources, these scales important can be found.
- Many numerical and experimental techniques can help us measure and estimate stress magnitude in the stress fields.

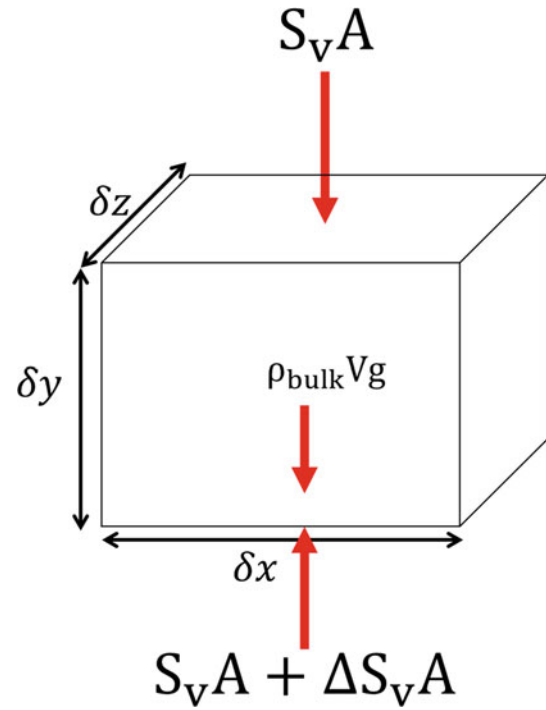
In this section, the subsurface stresses and pore pressure and stress tensor will be discussed briefly.

Subsurface Stresses and Pore Pressure

Subsurface Stresses

Stress in rock systems varies with depth. Unlike fluid pressure, the stress in rocks depends on many elements such as the properties of the upper layers and other geological conditions (i.e., tectonic stresses). In general, stress (σ) can be formulated as a force (F) acting on an area (A):

$$\sigma = \frac{F}{A}. \quad (1)$$



Geomechanics, Fig. 1 Schematic of forces on a small element in subsurface systems

In the element that is shown in Fig. 1, with volume $\delta x \times \delta y \times \delta z$, the summation of all vertical forces is zero in an equilibrium condition, $\sum F = 0$. Thus:

$$S_v A + \rho_{bulk} \delta x \delta y \delta z g - S_v A - \Delta S_v A = 0 \quad (2)$$

$$\Delta S_v = \rho_{bulk} g \delta z \quad (3)$$

which in turn:

$$\int_0^{S_v(z)} dS_v = \int_0^z \rho_{bulk} g \delta z. \quad (4)$$

where S_v is total vertical stress, ρ_{bulk} is the density of bulk material (density of water and rock), and g is the gravitational acceleration.

Effective stress (σ_v) is defined as the difference between total vertical stress and pore pressure P_p , which is mostly compressive. Effective stress is an important quantity using which one can analyze the stress in rocks and soils better. The effective stress is given by:

$$\sigma_v = S_v - P_p \quad (5)$$

Pore Pressure

In geomechanics, in particular for geosystems, one must always consider the effect of fluid on the total stress. Depending on the amount of fluid, pore pressure can make strengthen or weaken the system. For example, exceeding a certain amount of fluid can result in liquefaction and large geological disasters. Similarly, granular particles can stick together when a small amount of fluid exists. An example of this situation can be found in castle structures generated using sand on beaches. Thus, understanding the effect of pore pressure on stress is very critical. Generally speaking, pore pressure, in large-scale systems, can be divided into two groups: hydrostatic and non-hydrostatic. In many cases, estimation of the pore pressure is challenging since there are lots of pores connected. In lower depths, pore pressure increases with the depth hydrostatically. It should be noted that the stress can also vary with the rock properties and the type of materials. Similarly, the rock type can also control the pore pressure. For instance, low permeability barriers can cause overpressure. Pore pressure measurement development with depth is shown in Fig. 2 graphically. The overpressure parameter (λ_p) is defined as the ratio of pore pressure to the total vertical stress:

$$\lambda_p = \frac{P_p}{S_v} \quad (6)$$

Vertical Stress

The vertical stress solely is not enough to evaluate the state of solid. But, the stress applied to an element can be defined in form of principal stress. The principal stress is all stresses applied perpendicular to each other in three directions. If there

is no shear stress on the plane, the stress is named principal stress. In most cases, the total vertical stress is considered principal stress. In this particular situation, the other principal stresses are horizontal. All stresses can be formed as a symmetric 3×3 matrix, which is called a stress tensor. Since the stress tensor is symmetric, there are three eigenvalues, and they are called principal stresses since there is no shear stress on those planes. Mostly, the stress elements should be rotated to reach the principal stress.

Research Path of Geomechanics

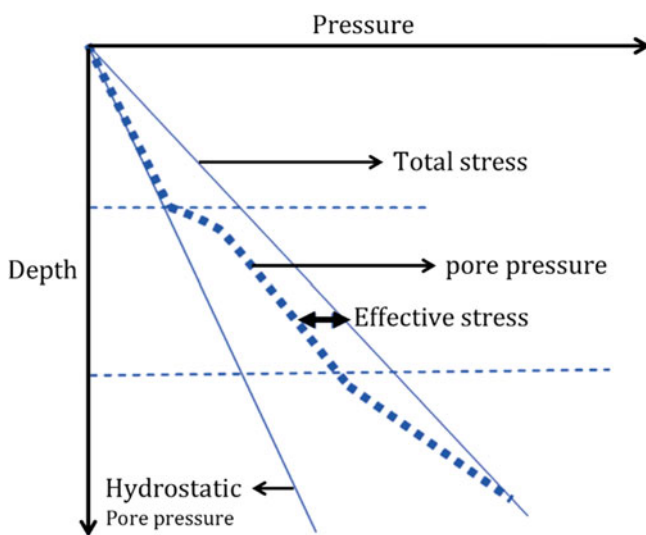
The importance of geomechanics for problems related to geomaterials has been shown extensively, which has led to conducting more research both experimentally and numerically. Most of the initial studies and analytical solutions were developed for particular problems, and they often contain various user/material-dependent parameters. With the introduction of high-speed computers, numerical investigations have become more popular. Indeed, the recent progress in computational capabilities has led to developing methods that were not applicable before for large-scale systems, such as the finite element method (FEM), finite volume method (FVM), etc. Although an enormous improvement can be seen in the experimental techniques, numerical simulations are developing rapidly.

One of the fields that has witnessed the abovementioned advancement is developing various Lagrangian approaches. The discrete element method (DEM) is one of the well-known Lagrangian techniques used in geomechanics where each particle's trajectory is estimated and forces applied on particles are calculated based on Newton's second law (O'Sullivan 2011, 2015). To achieve a more accurate and comprehensive representation of particles and the mechanical behavior of the system, a wide range of forces are added to the DEM formulations, including cohesion, and DLVO theory.

As discussed earlier, the pore pressure effect should not be neglected as it can tremendously change the spectrum of final deformation. Computational fluid dynamic (CFD), which is an Eulerian method, has been added to the Lagrangian techniques (i.e., DEM) and is known as CFD-DEM coupling (Tahmasebi and Kamrava 2019; Zhang and Tahmasebi 2019; Zhao and Shan 2013). Moreover, other important physics can also be added to the above framework and develop THMC (thermo-hydro-mechanical-chemical) modeling (Zhang et al. 2015).

Conclusion

Geomechanical modeling, in particular those related to sub-surface systems, is one of the complex and important systems, which affect humans' life, and their accurate modeling can be to reduce the risk while reducing the cost. This branch of



Geomechanics, Fig. 2 Pore pressure development with depth

science and engineering has been applied in many fields related to geomaterials. To have a fundamental knowledge of geomechanics, one should understand the stress for the target system. For example, CO₂ leakage and sequestration are some of the topics that require an extensive application of geomechanical modeling, and it can help to reduce the deformation and safe preserving of carbon in subsurface systems. With the recent demand for electric cars and also renewable energy, mining will also experience more challenges in terms of slope modeling when more minerals are extracted.

Bibliography

- Begnudelli L, Sanders BF (2007) Simulation of the St. Francis dam-break flood. *J Eng Mech* 133:1200–1212
- O'Sullivan C (2015) Advancing geomechanics using DEM. In: 3rd international symposium on geomechanics from micro to macro. Taylor & Francis Group, London (pp. 21–32)
- O'Sullivan C (2011) Particle-based discrete element modeling: geomechanics perspective. *Int J Geomech* 11:449–464
- Rogers JD (2006) Lessons learned from the St. Francis Dam failure
- Tahmasebi P, Kamrava S (2019) A pore-scale mathematical modeling of fluid-particle interactions: Thermo-hydro-mechanical coupling. *Int J Greenh Gas Control* 83:245–255. <https://doi.org/10.1016/j.ijggc.2018.12.014>
- Zhang R, Winterfeld PH, Yin X, Xiong Y, Wu Y-S (2015) Sequentially coupled THMC model for CO₂ geological sequestration into a 2D heterogeneous saline aquifer. *J Nat Gas Sci Eng* 27:579–615. <https://doi.org/10.1016/j.jngse.2015.09.013>
- Zhang X, Tahmasebi P (2019) Effects of grain size on deformation in porous media. *Transp Porous Media* 129:321–341. <https://doi.org/10.1007/s11242-019-01291-1>
- Zhao J, Shan T (2013) Coupled CFD-DEM simulation of fluid-particle interaction in geomechanics. *Powder Technol* 239:248–258. <https://doi.org/10.1016/j.powtec.2013.02.003>

Geoscience Signal Extraction

Frits Agterberg

Central Canada Division, Geomathematics Section,
Geological Survey of Canada, Ottawa, ON, Canada

Definition

In various geoscience applications, a series of observed data can be considered to be the sum of a signal and random noise. Usually, the noise can be eliminated in order to retain the signal. “White” noise results in a value less than one at the origin of the correlogram, or as the so-called “nugget effect” in the variogram of geostatistics. “Red” noise is best evaluated in the frequency domain where white noise shows as a horizontal line whereas red noise results in a straight line that dips in the direction of the frequency axis. In most applications of geoscience signal extraction, the signal occurs at or

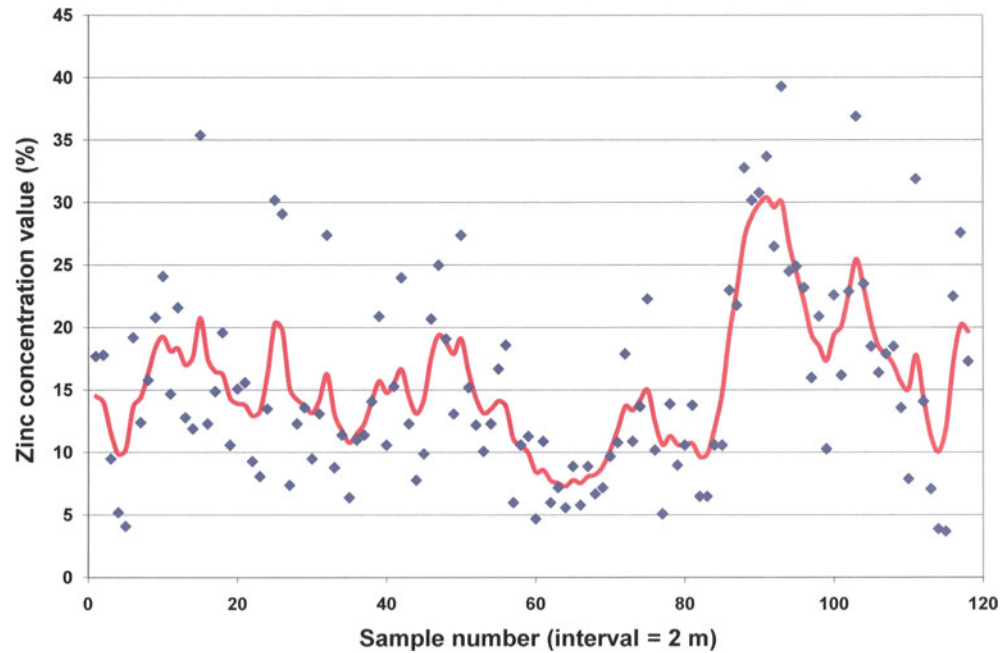
around a single point. However, especially in stratigraphy, the signal can be sinusoidal. A primary example of this is the Milankovitch cycle that not only controlled the occurrences of ice ages but also those of many older periodical stratigraphic phenomena as well. Finally, the signal can be a meaningful pattern involving many points at which observations are obscured by superimposed random noise. Examples of these different types of signals will be given in this entry.

Introduction

Suppose a series of observed values (x_k) is the sum of two series: signal (s_k) and white noise (n_k) with $k = 1, 2, \dots, n$. The autocovariance functions of such series can be written as $\Gamma_x(h)$, $\Gamma_s(h)$, and $\Gamma_n(h)$ satisfying the relations: $\Gamma_n(h) = 0$ if $h \neq 0$ and $\Gamma_x(h) = \Gamma_s(h) + \Gamma_n(h)$. Suppose that $\Gamma_x(0) = 1$ and $\Gamma_n(0) = c$. Let $F(h)$ denote a filter to which the record must be subjected in order to obtain the signal: $s_k = \int_{-\infty}^{\infty} F(h)x(k+h)dh$. Then, $\Gamma_s(h) = \int_{-\infty}^{\infty} F(h+\tau)\Gamma_s(\tau)d\tau + \int_{-\infty}^{\infty} F(h+\tau)\Gamma_n(\tau)d\tau$ or $\Gamma_s(h) = \int_{-\infty}^{\infty} F(h+\tau)\Gamma_s(\tau)d\tau + (1-c)F(h)$. Fourier transformation of both sides gives: $G(\omega) = \Phi(\omega) \cdot G(\omega) + (1-c)\Phi(\omega)$ where $G(\omega) = \int_{-\infty}^{\infty} \Gamma_h \cos \omega h dh$ and $\Phi(\omega) = \int_{-\infty}^{\infty} F_h \cos \omega h dh$. It follows that: $\Phi(\omega) = \frac{G(\omega)}{G(\omega) + (1-c)}$. The filter $F(h)$ can then be found by taking the inverse Fourier transform: $F(h) = \int_{-\infty}^{\infty} \Phi(\omega) \cos \omega h d\omega$. When $\Gamma_x(h) = \exp(-a \cdot |h|)$, then: $G(\omega) = \frac{2ac}{a^2 + c^2}$ and $F(h) = \frac{ac}{\pi p^2(1-c)} \int_{-\infty}^{\infty} \frac{1}{\frac{\omega^2}{p^2} + 1} \cos \omega h d\omega$ where $p = [a^2 + 2ac/(1-c)]^{0.5}$. It follows that: $F(h) = q \cdot \exp(-p \cdot |h|)$ where $q = ac/[(1-c) \cdot p]$. Two methods most frequently used to help extract the signal are using the autocovariance function and spectral analysis. The latter method has the advantage that its successive values are stochastically independent. In geostatistics, it is common to use the variogram which is closely related to the autocovariance.

The autocovariance of an ordered sequence of random variables is defined as $\Gamma_h = E[(X_k - \mu) \cdot (X_{k+h} - \mu)]$. The sample autocovariance for n values can be estimated by: $C_h = \frac{1}{n-h} \sum_{k=1}^{n-h} (x_k - \bar{x})(x_{k+h} - \bar{x})$ where the mean value $\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$. The autocorrelation function satisfies $\rho_h = \Gamma_h/\Gamma_0 = \Gamma_h/\sigma^2$, and the sample autocorrelation coefficient is $r_h = C_h/C_0$. Together, the r_h values form a so-called “correlogram.” Wiener’s (1933) theorem of autocorrelation states that the power spectrum satisfies: $\varphi(\omega) = \int_{-\infty}^{\infty} \rho_h e^{-i\omega h} dh$. Since ρ_h is even, $\varphi(\omega) = \int_{-\infty}^{\infty} \rho_h \cos \omega h dh$. After some manipulation, it follows that: $F(h) = q \cdot \exp(-p \cdot |h|)$ where $q = ac/[(1-c) \cdot p]$ (cf. (Yaglom 1962; Agterberg 1967). In the next section, the power spectral density $P(f)$ will be computed as:

Geoscience Signal Extraction,
Fig. 1 Pulacayo Mine zinc concentration values for 118 2-m channel samples along horizontal drift (original data from de Wijs 1951). Sampling interval is 2 m. “signal” was retained after filtering out “white noise”. (Source: Agterberg 2012, Fig. 1)



$$P(f) = C_0 + \sum_{k=1}^m W(k) C_k \cos 2\pi k f \quad \text{with } W(k) = \frac{1}{2} (1 + \cos \pi k / m).$$

The weighting function $W(k)$ defines a cosine-shaped lag window for the autocovariance function whose use is called “hanning” (cf. Blackman and Tukey 1958).

The purpose of hanning is that individual values $P(f)$ can be considered to estimate a smoothed version of the underlying true spectrum. For more explanations of why this procedure can be useful, see Blackman and Tukey (1958). The power spectrum $P(f)$ can be regarded as a decomposition of all variability in a series in terms of components of the variance for narrow frequency bands. Smoothing operations such as hanning significantly improve their estimation from the autocovariances by eliminating distortions.

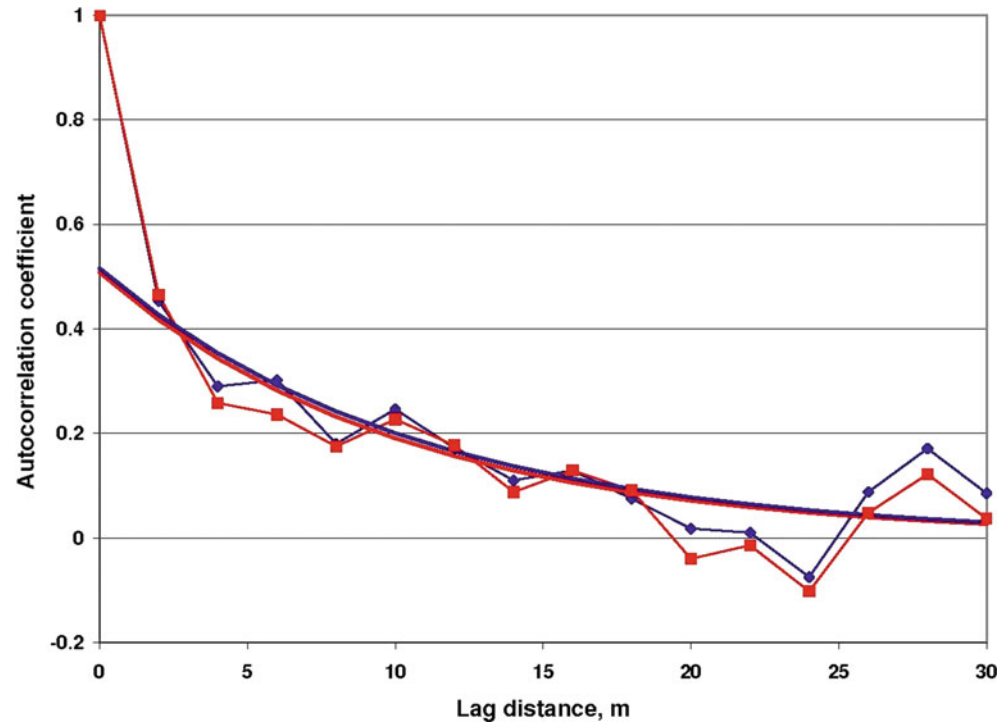
Pulacayo Mine and Main KTB Borehole Examples

De Wijs (1951) used a series of 118 zinc concentration values from samples taken at a regular interval along a horizontal drift in the Pulacayo Mine in Bolivia (Fig. 1). Average zinc value is 15.61%. This series has been used extensively in later studies by many authors including Matheron (1962), Tukey (1970), Chen et al. (2007), Lovejoy and Schertzer (2007), Cheng (2014), and Agterberg (2014). The geological setting of the Pulacayo Mine and genesis of this sphalerite-quartz ore deposit are briefly described in a scientific communication by Pinto-Vásquez (1993).

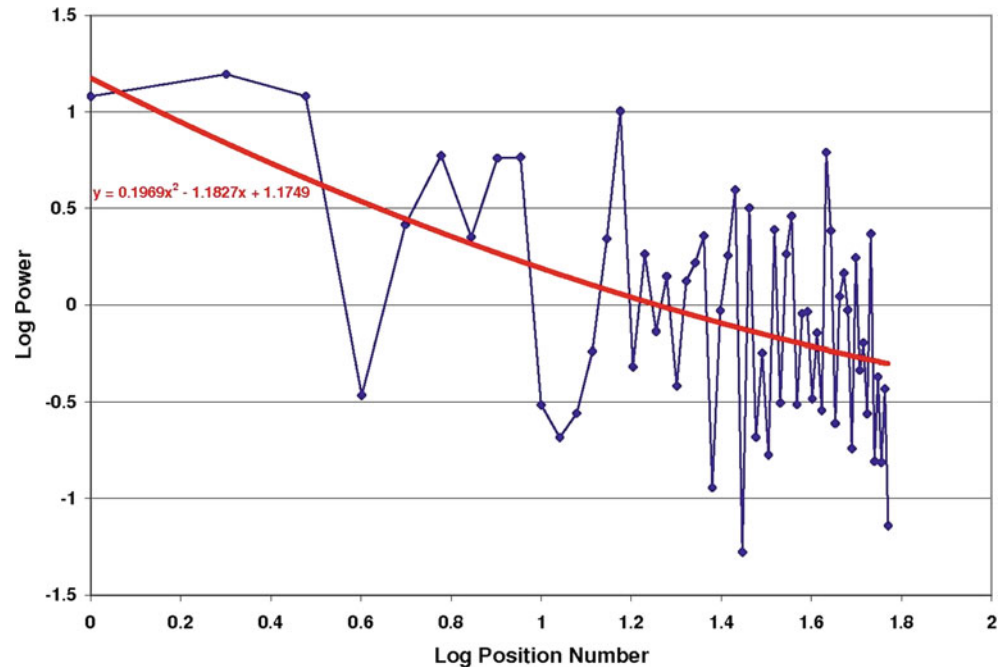
The zinc values used by de Wijs (1951) are for channel samples cut at 2-m intervals across the steeply dipping Tajo vein along a horizontal drift on the 446-m level. These channel samples were cut over a standard width of 1.30 m corresponding to expected mining width. Figure 1 shows both the original zinc values and the signal extracted from them. The method used for this smoothing was described in detail in Agterberg (1974). It assumes that each zinc value is the sum of a “signal” component for continuous change along the series plus a random white-noise component. Together these two components were assumed to produce an autocorrelation function of the type $\rho_h = c \cdot \exp(-ah)$ where h represents distance along the drift (see Fig. 2) and $c < 1$ and a are constants. Several authors including Matheron (1962) preferred to work with logarithmically transformed metal concentration values. The best-fitting autocorrelation functions for original and logarithmically transformed zinc values are practically the same for the Pulacayo deposit. Figure 3 indicates that the noise also can be assumed to be red (cf. Tukey 1970), because white noise would have resulted in an approximately horizontal line in the frequency domain.

Figure 4 shows 2706 copper concentration values taken at 2-m intervals along the Main German Continental Deep Drilling Program Borehole (abbreviated to KTB, cf. Goff and Hollinger 1999), along with mean values for 101-m long segments of drill core. Statistical analysis for this example is discussed in more detail in Agterberg (2012). The long series of copper values was divided into three subseries, each about 2 km in length. Figure 5 shows that, over short distances, the logarithmically transformed (original) copper values satisfy

Geoscience Signal Extraction, Fig. 2 Estimated autocorrelation coefficients for original data shown in Fig. 1 (diamonds) and logarithmically transformed zinc concentration values, shown together with best-fitting negative exponential autocorrelation functions. The curves for original and logarithmically transformed data coincide approximately. (Source: Agterberg 2012)



Geoscience Signal Extraction, Fig. 3 Periodogram of the 118 zinc concentration values with quadratic curve fitted by least squares (Logarithms base 10). The negative exponential curve of Fig. 1 would show as a horizontal line in this kind of plot. Obviously, the assumption that the noise is “red” instead of “white” is better because it would result in a straight line with a negative dip. (Source: Agterberg 2012, Fig. 25)

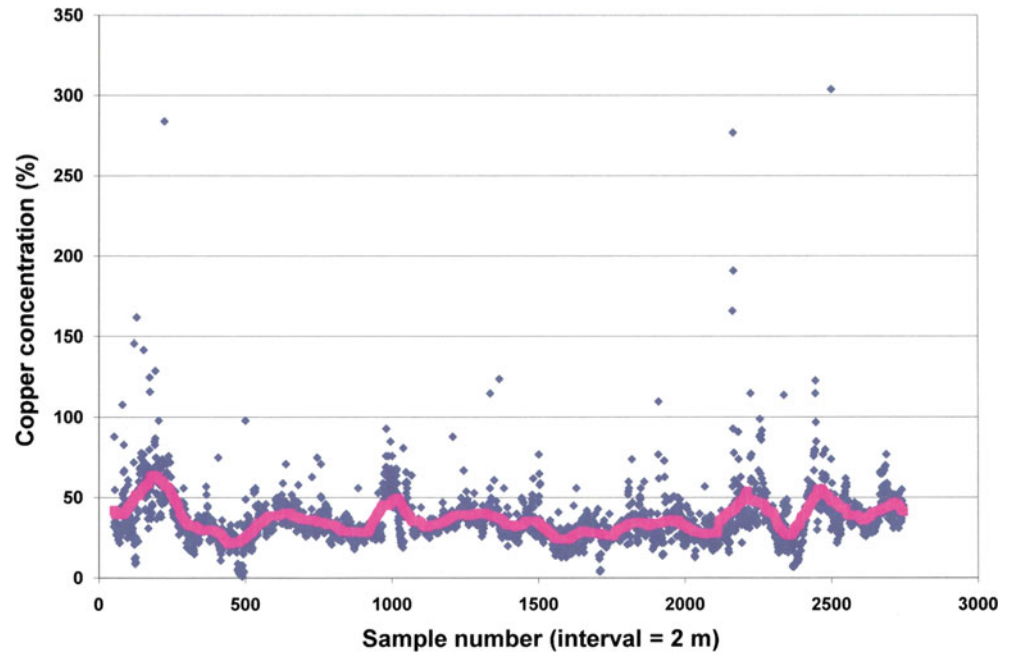


approximately a signal-plus-white-noise model of the type $\rho_h = c \cdot \exp(-ah)$ with approximately the same value of a along the entire length of this deep borehole. To some extent, the white noise in this example is due to the measurement errors of the copper content determination method used.

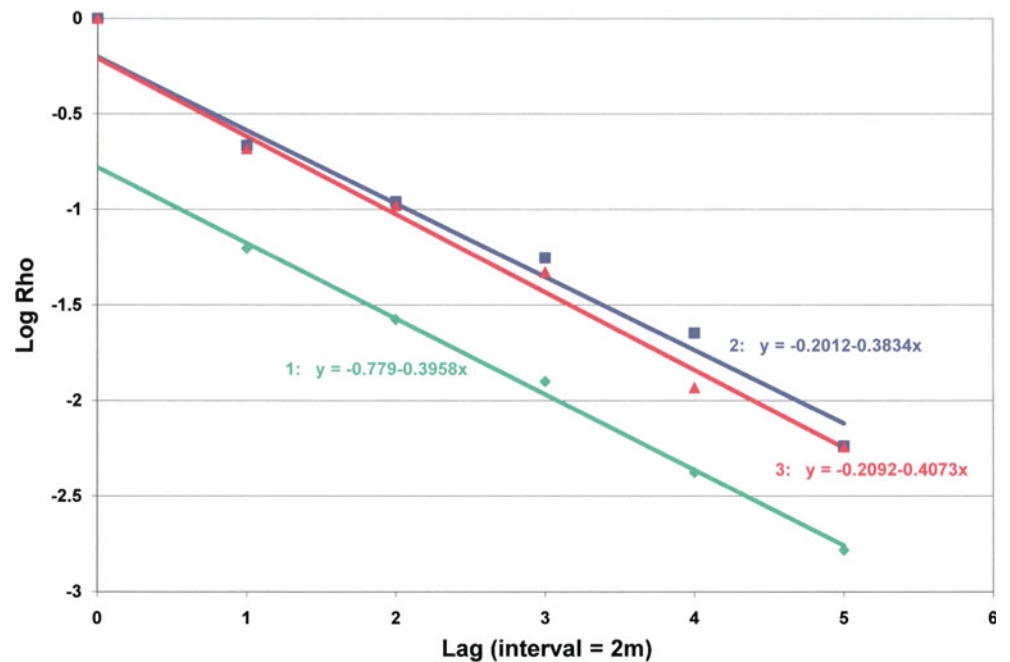
Spectral Analysis: Glacial Lake Barlow-Ojibway Example

Glacial varves that consist of alternate silt and clay layers, because of their presumed annual nature (silt deposited during

Geoscience Signal Extraction, Fig. 4 Copper concentration (ppm) values along Main KTB bore-hole together with average copper concentration values for 101 m long segments of drill-core. Locally the original data deviate strongly from the moving average. (Source: Agterberg 2012, Fig. 15)



Geoscience Signal Extraction, Fig. 5 Correlograms for three segments of series of original copper values shown in Fig. 4. Series 2 (for depths between 2 and 4 km) and series 3 (for depths between 4 and 5.54 km) exponential autocorrelation functions similar to the one for depths from 0.05 to 2 km except for difference in Log Rho intercept. (Source: Agterberg 2012, Fig. 16)



the summers; clay during the winters), have attracted the attention of geologists as a geochronological tool. Anderson and Koopmans (1965) were among the first to apply spectral analysis to varves.

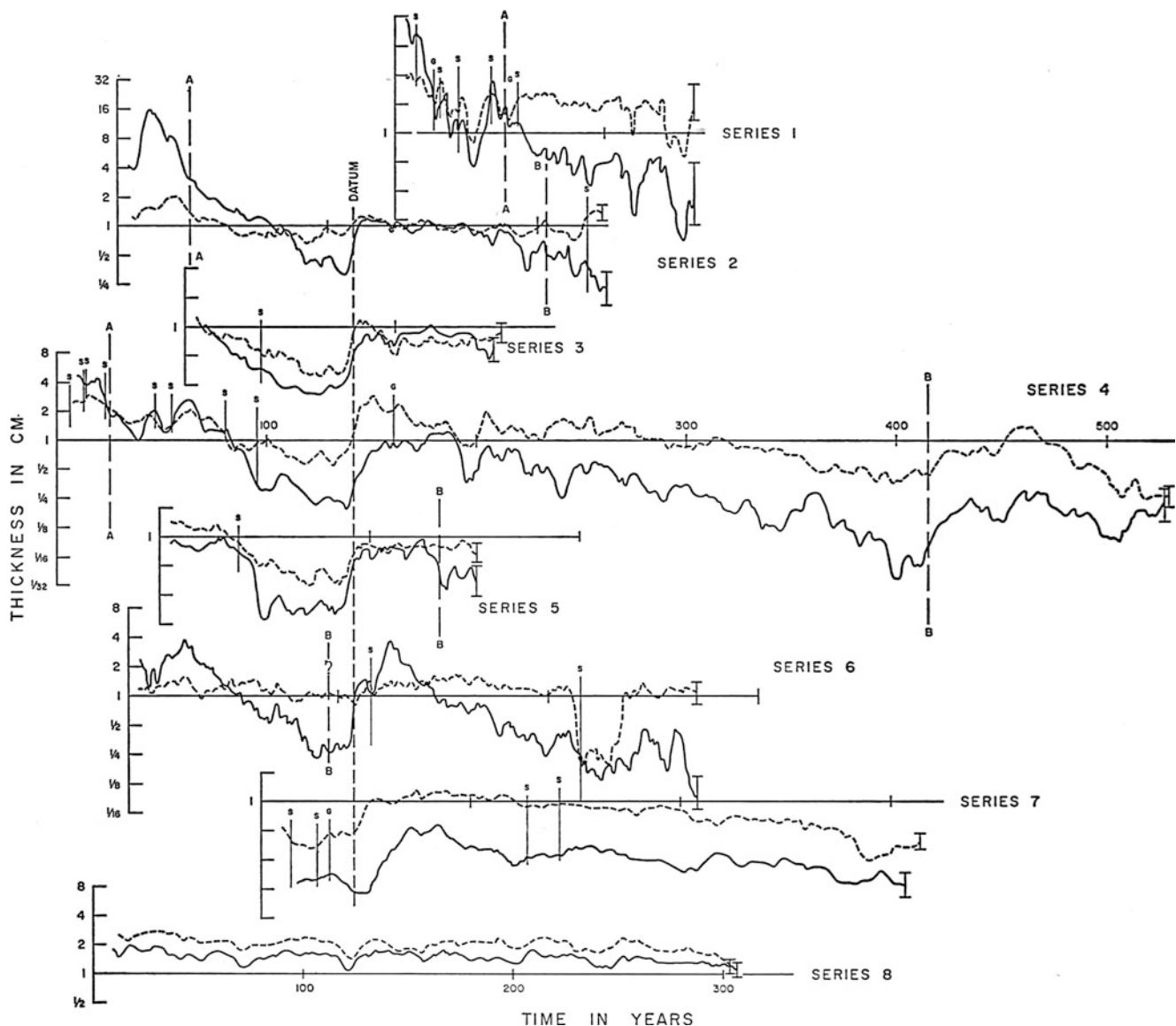
Lake Barlow-Ojibway was a late-glacial water body that formed approximately 11,000 years ago during the retreat of the Late Wisconsin ice sheet in northern Ontario and western Quebec. The lake, in its maximum extent, measured about

960 km in the east-west and 240 km in the north-south direction. The most extensive deposit formed in it was the sheet-like body of varves which extends uninterruptedly across the Hudson Bay – St. Lawrence divide. More detailed geological background with references to earlier work is given in Agterberg and Banerjee (1969). Kuenen (1951) had proposed a mechanism of varve deposition by annual

turbidity currents with rapid silt deposition during the summer followed by slow clay deposition during the winter.

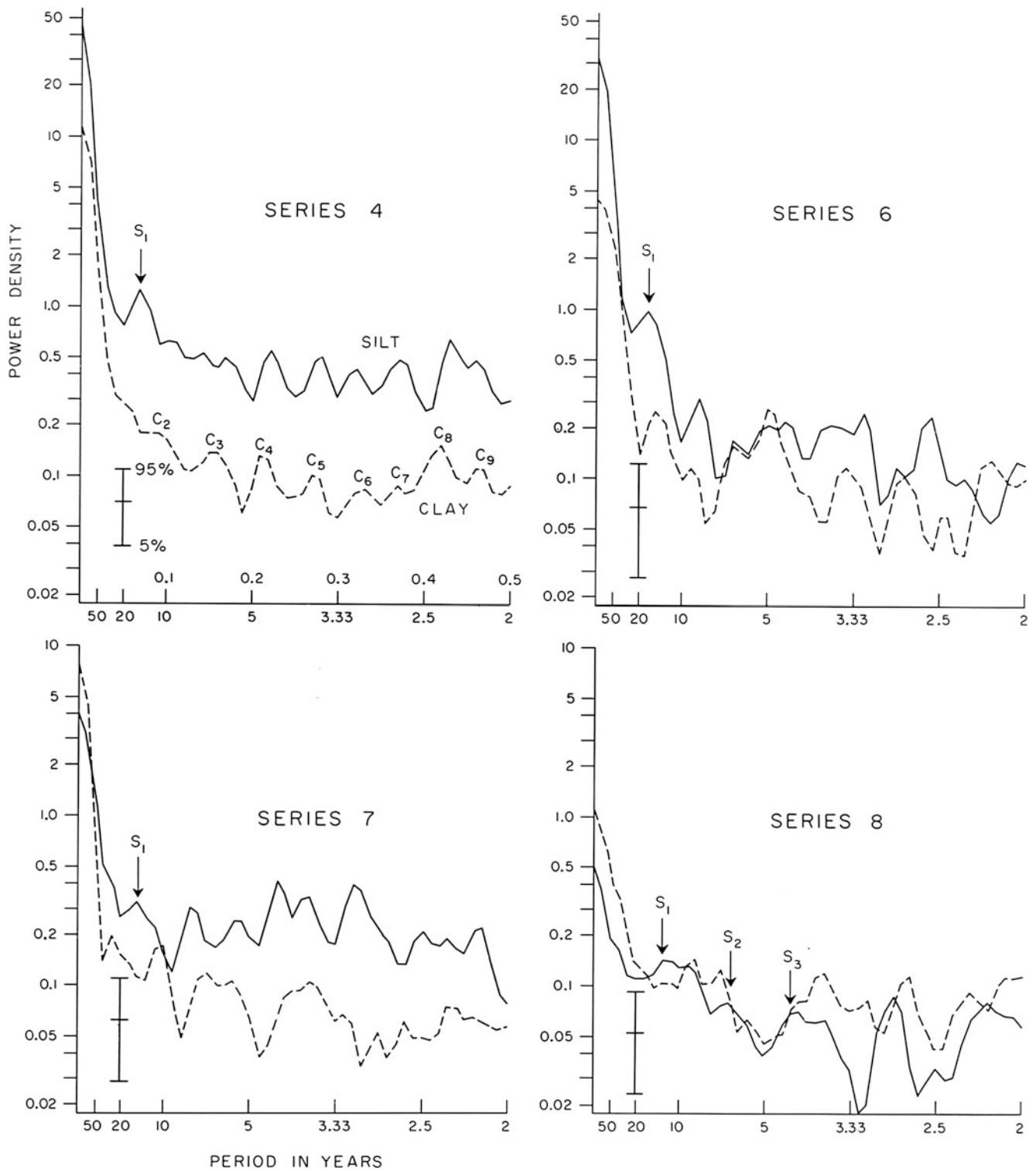
In total, Agterberg and Banerjee (1969) statistically analyzed 4310 thickness measurements from eight sections (Fig. 6). Six of these sections were aligned with respect to one another on the basis of the “datum” which is a relatively abrupt increase in thickness of the varves. This datum coincides with the Cochrane readvance of the land-ice about 8000 years ago (Antevs 1925; Hughes 1955). In the longest series (No. 4 with 537 silt-clay couplets), silt layers near the bottom are about ten times as thick as those near the top. The clay layers near the top are about three times thinner than those near the bottom. A logarithmic transformation was

desirable in order to ensure that variations in thickness are of the same magnitude in parts of the series that are relatively thick and relatively thin, respectively. For example, the sum of squared silt thickness data, on which measures of variability such as the variance are based, amounts to 1037.2 cm^2 for series 4. The first 100 values in this series contribute 91.99% and the last 100 values only 0.02% to this total sum of squares that is based on all 537 data. If no logarithmic transformation would be applied to the observations, results from a statistical study would be largely determined by variations in the thicker parts of the series, whereas variations in the thinner parts of the series would have a negligibly small effect on the end results. The logarithmic transformation stabilizes the



Geoscience Signal Extraction, Fig. 6 Filtered curves for varve thickness data, series 1-8, Lake Barlow-Ojibway. Solid lines denote silt thickness; broken lines are for clay thickness. (a) A is boundary between sandy and silty facies; (b) B is boundary between silty and diamictic

facies, s: slumped intervals, G gaps in records with 5–10 varves missing. Width of 95% confidence belt is shown at end of each series. (Source: Agterberg and Banerjee 1969, Fig. 7)



Geoscience Signal Extraction, Fig. 7 Power spectra for four longest series ($n > 250$). Frequency scale is shown in spectrum for series 4 only. 95% confidence belts are also shown. (Source: Agterberg and Banerjee 1969, Fig. 5)

variance. The ratio of the variances for the first and last sets of 100 silt values for series 4 is 33.6, but after logarithmic transformation, it reduces to 1.5.

The varve time series are not stationary in that, for example, both mean silt and clay layer thickness tend to decrease with time. As a rule, such decreases are stronger in the silt than in the clay. Observed varve series are incomplete because

thicknesses could not be measured in places where slumped layers occur. These places are shown as “s” in Fig. 6. Some gaps where a larger number of varves is probably missing are identified as “G” in this figure. In total, as much as 10% of total number of varves may be missing in some series including series 4.

Varve-thickness data, like tree-rings, provide a sensitive tool to measure very small temperature fluctuations on Earth. Power spectra are shown in Fig. 7 for the four longest series ($n > 250$). All spectra commence with a strong peak close to the origin for low-frequency waves representing long-term variations or “trends.” For 20-year and lesser periods, the spectra tend to flatten out. It is likely that the peaks for shorter periods are not exclusively caused by random variations but are partly due to weak periodical phenomena. The first peak for silt (S_1 in Fig. 7) occurs at a period of about 14 years in all four spectra. Along with its 14-year multiples, it also shows up clearly in the correlogram for residuals from the Series 4 linear trend (see Fig. 8).

A 14-year periodicity of this type is not known to occur in glacial deposits elsewhere in the world. A possible explanation was offered by Schove (1972). This author suggested climatological variations due to solilunar cycles as the cause. Brier (1968) had shown that solilunar cycles produce significant tidal effects in the atmosphere at 13.5 and 27 years. According to Schove (1972), silt components of varves are suitable for study of solilunar cycles as affect in the summer months of June and July.

The 13.5-year (163 calendar months) cyclicity is the beat period between the calendar month (30.44 days) and the synodic month (29.5 days). These two cycles are in phase with one another every 163 months, and both then are approximately in phase with the cycle arising from the anomalistic month (27.55 days). The synodic cycle is the period from one new moon to the next and the anomalistic cycle that from one

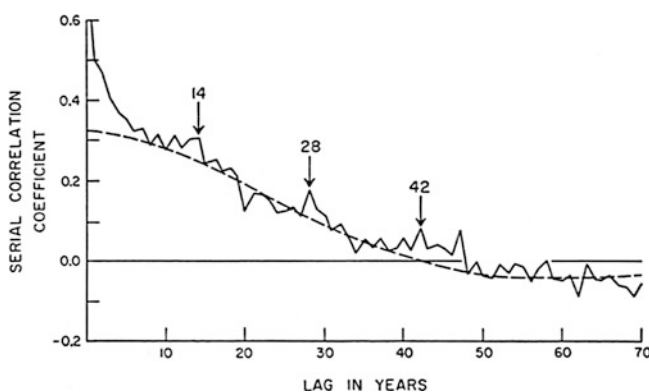
perigee to the next. These two cycles are most important in determining the magnitude of the solilunar gravitational tide, which has significant effects on monthly weather conditions according to Brier (1968).

Peaks in power spectra often are followed by secondary peaks that tend to be equally spaced along the frequency scale. These may represent “harmonics” reflecting deviations from sinusoidal shape of the periodic phenomenon (*cf.* Schwarzacher 1967). This phenomenon is most conspicuous for clay in series 4 where there are nine peaks, all of which are multiples of frequency 0.05 or a value slightly larger than 0.05. This probably reflects the 22-year and 11-year sunspot cycles that have been found in varves elsewhere in the world (Schove 1983; Berry 1987) and in other laminated sediments, e.g., in marl-limestone couplets belonging to the Late Campanian *Radotruncana calcarata* Zone (Wagreich et al. 2012). Sunspot cycles have variable duration, but on average, they last 11 years. The 22-year cycle is due to alternation of stronger and weaker 11-year cycles.

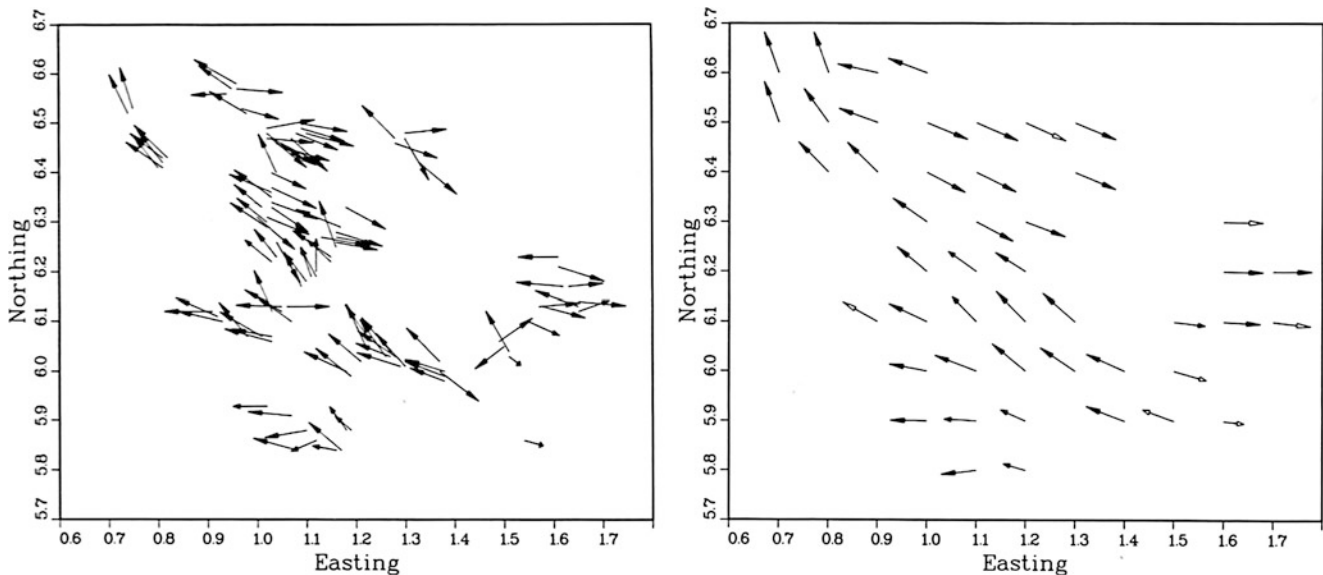
Unit Vector Fields

In the final example, the signal is a two-dimensional pattern that only becomes clearly visible after the elimination of white noise. It is concerned with so-called B-axes for Hercynian minor folds and schistosity planes in quartzphyllites belonging to the crystalline basement of the Italian Dolomites. During Alpine orogeny, these *s*-planes were not only refolded, but there were also large-scale sliding movements along the Hercynian schistosity resulting in various kinds of Alpine rotations of the Hercynian B-axes. Elimination of local random disturbances helps to outline the resulting regional deformation patterns. This example is for measurements of B-axes in quartzphyllites in the San Stefano area east of the Dolomites.

Figure 9 (after Agterberg 1985) shows original measurements or average B-axes for (100 m × 100 m) squares on the left side in comparison with unit vector means on the right side that are based on much larger samples obtained by combining all measurements from within circular neighborhoods with radii of 1 km (solid arrow heads) or 2 km (open arrowheads). Local random deviations from the regional pattern are eliminated when the mean vectors are used because these are based on many more original measurements. Structural interpretation of the change in average dip of the B-axes is as follows. The ESE-striking schistosity-planes and the B-axes are Hercynian in age. During Alpine orogeny, the San Stefano crystalline containing them formed the core of an anticlinal structure in which the Hercynian schistosity planes were reactivated. The trace of the resulting Alpine anticlinal structure with NW-SE trending axial plane intersects the approximately WNW-ESE striking Hercynian schistosity planes.



Geoscience Signal Extraction, Fig. 8 Correlogram for residuals from linear trend, silt, series 4. The 14-year periodicity that shows as a weak oscillation is indicated by arrows. Fitted curve satisfies stochastic differential equation of second order Markov process. (Source: Agterberg and Banerjee 1969, Fig. 9)



Geoscience Signal Extraction, Fig. 9 San Stefano area. Unit of distance for coordinates of the grid is 10 km. Length of arrow is proportional to cosine of angle of dip (Maximum arrow length is 1 km). Left side: B-axes and unit vector means in (100 m × 100 m)

cells. Unit vector means of B-axes located within circles with 1 km radius (solid arrow heads) and 2 km radius (open arrow heads). (Source: Agterberg 2012, Fig. 4)

Summary and Conclusions

Geoscience signal extraction is one of the most important statistical methods in geoscience. For many different reasons including measurement errors, various features are obscured when they are observed at the surface of the Earth. White and red noise are the main causes of this obscuration. In this entry, various methods were discussed to combine multiple observations to eliminate random noise in order to extract meaningful signals. These signals either occur at or in the vicinity of points in geological time or three-dimensional space; they can be sinusoidal or regional in their extent. The examples given were (1) element concentration values in orebodies and ordinary rocks, (2) cyclical phenomena in sedimentary rocks, and (3) orogenic deformation patterns in structural geology.

Cross-References

- [Spectral Analysis](#)
- [Variogram](#)

Bibliography

Agterberg FP (1967) Mathematical models in ore evaluation. *J Can Oper Res Soc* 5:144–158

- Agterberg FP (1974) *Geomathematics – mathematical background and geo-science applications*. Elsevier, Amsterdam, 596 pp
- Agterberg FP (1985) Spatial analysis in the earth sciences. In: Glaesner PS (ed) *The role of data in scientific progress. Proceedings of 9th international CODATA conference*. Elsevier, Amsterdam, pp 9–18
- Agterberg FP (2012) Unit vector field fitting in structural geology. *Zeitschr D Geol Wis* 40(4/6):197–211
- Agterberg FP (2014) *Geomathematics: theoretical foundations, applications and future developments*. Springer, Cham/Heidelberg, 553 pp
- Agterberg FP, Banerjee I (1969) Stochastic model for the deposition of varves in glacial Lake Barlow-Ojibway, Ont., Canada. *Can J Earth Sci* 6:625–652
- Anderson RY, Koopmans LH (1965) Harmonic analysis of varve time series. *J Geophys Res* 68:877–893
- Antevs E (1925) Retreat of the last ice-sheet in eastern Canada. *Geol Surv Can Mem* 146:142 pp
- Berry PAM (1987) Periodicities in the sunspot cycle. *Vistas Astr* 30(2): 87–108
- Blackman RB, Tukey JW (1958) *The measurement of power spectra*. Dover, New York, 190 pp
- Brier GW (1968) Long-range prediction of the zonal westerlies and some problems in data analysis. *Rev Geophys* 6:525–551
- Chen Z, Cheng Q, Chen J, Xie S (2007) A novel iterative approach for mapping local singularities from geochemical data. *Nonlinear Process Geophys* 14:317–324
- Cheng Q (2014) Generalized binomial multiplicative cascade processes and asymmetrical multifractal distributions. *Nonlinear Process Geophys* 21:472–482
- De Wijs HJ (1951) Statistics of ore distribution I. *Geologie en Mijnbouw* 13:365–375
- Goff JA, Hollinger K (1999) Nature and origin of upper crustal seismic velocity fluctuations and associated scaling properties; combined stochastic analyses of KTB velocity and lithology. *J Geophys Res* 104:13,169–13,182

- Hughes OL (1955) Surficial geology of Smooth Rock and Iroquois Falls map area, Cochrane District, Ontario. Unpubl PhD thesis, University of Kansas, Lawrence
- Kuenen PH (1951) Mechanics of varve formation and the action of turbidity currents. *Geol Före Förh* 73:69–84
- Lovejoy S, Schertzer D (2007) Scaling and multifractal fields in the solid earth and topography. *Nonlinear Process Geophys* 14:465–502
- Matheron G (1962) *Traité de géostatistique appliquée*. Mém BRGM 14, Paris, 333 pp
- Pinto-Vásquez J (1993) Volcanic dome associated precious and base metal epithermal mineralization at Pulacayo, Bolivia. *Econ Geol* 88:697–700
- Schove DJ (1972) A varve teleconnection project. *Proceedings VIII Congr INQUA*, Paris 1969, 928–935
- Schove DJ (1983) Sunspot cycles. *Hutchison Ross*, Stroudsburg, 410 pp
- Schwarzacher W (1967) Some experiments to simulate the Pennsylvanian rock sequence of Kansas. *Kans Geol Surv Comput Contrib* 18: 5–14
- Tukey JW (1970) Some further inputs. In: Merriam DF (ed) *Geostatistics*. Plenum, New York, pp 163–174
- Vaughan S, Bailey RJ, Smith DG (2011) Detecting cycles in stratigraphic data: spectral analysis in the presence of red noise. *Paleoceanography* 26:PA 4216. <https://doi.org/10.1029/2011PA002195>
- Wagreich M, Hohenegger J, Heuher S (2012) Nannofossil biostratigraphy, strontium and carbon isotope stratigraphy, cyclostratigraphy and an astronomically calibrated duration of the Late Campanian. *Cretac Res* 38:80–96
- Wiener N (1933) *The Fourier integral and certain of its applications*. Dover, New York, 201 pp
- Yaglom AM (1962) *An introduction to the theory of stationary random functions*. Prentice-Hall, Englewood Cliffs, 235 pp

Geosciences Digital Ecosystems

Stefano Nativi¹ and Paolo Mazzetti²

¹Joint Research Centre (JRC), European Commission, Ispra, Italy

²Institute of Atmospheric Pollution Research, Earth and Space Science Informatics Laboratory (ESSI-Lab), National Research Council of Italy, Rome, Italy

Synonyms

Digital ecosystem

Definition

A “Geosciences Digital Ecosystem” is a digital ecosystem whose functional utility is the generation and sharing of knowledge on the Earth.

In other words, a “Geosciences Digital Ecosystem” is a system of systems that applies the digital ecosystem paradigm

to model the complex collaborative and competitive social domain dealing with the generation of knowledge on the Earth planet.

Stemming from the concept of natural ecosystem (Blew 1996), the Digital Ecosystem paradigm focuses on a holistic view of a diversity of autonomous organizations (i.e., corresponding to the “biotic” component), sharing (as the “abiotic” component) a common digital environment and a set of digital resources to survive, thrive, and coevolve. Digital ecosystems, as their digital counterparts, aim at emulating the self-organizing properties of biological ecosystems, which are considered to be robust, self-organizing, and scalable architectures that can automatically solve complex, dynamic problems (Briscoe and De Wilde 2006).

An archetypal example of geoscience digital ecosystem is represented by a set of public and private autonomous stakeholders, which deal with the generation of information and knowledge on the Earth, loosely cooperating on the Internet environment and complementing each other by sharing datasets, analytical software, and computing capacities (Nativi et al. 2021).

Essential Concepts and Components

Ubiquitous connectivity, datafication, sensing, and computing virtualization are the new and powerful capacities offered by a digitally transformed society. Industry, science, academia, and humanity must leverage these new instruments to further develop distributed knowledge systems and engage multiple, diverse, and autonomous stakeholders, for a better monitoring and understanding of the Earth systems and phenomena. This innovative paradigm, consisting of heterogeneous entities interacting through loosely coupled connections to generate and share knowledge on our environment is known as geosciences digital ecosystem. As in the case of the more mature business and software digital ecosystems (Moore 1996; Jansen et al. 2013), a geoscience digital ecosystem can be a temporary collaboration (for a recognized and timely limited common value) and, sometimes, an opportunistic one.

The geosciences digital ecosystem represents an innovative model to develop distributed knowledge platforms to study the Earth behavior, by engaging multiple and diverse stakeholders interacting in a common and powerful virtual environment, the cyber-physical sphere. It replaces and brings up to date the concepts of Spatial Data Infrastructures (SDI) we saw developing in the 1990s and 2000s. The geosciences digital ecosystem paradigm promises to address the traditional SDI technological limitations (e.g., technology evolution and the pragmatic interoperability) (Nativi et al. 2021) and more importantly the non-technological ones: the data strategies diversity, the different degrees of openness

characterizing the infrastructure constituent systems, the services and products ownership concerns, the pursuing of different enterprise utilities, and the business models heterogeneity. With the advent of big data and the consequent data-driven artificial intelligence algorithms and cloud computing, there is no longer a distinction between the local/physical digital environment and the network/virtual one as most organizations store and process their data on the network itself, which has become the common and virtual development platform for all the stakeholders.

By applying the digital ecosystem paradigm, complex and distributed systems of systems (such as the digital twins of the Earth and the knowledge platforms) can manage the challenges posed by their evolutionary nature and create the necessary conditions to be sustainable and thrive. The ecosystem model encapsulates the need for the diversity of actors and expertise, focusing on synergies and relationships, and favoring the capacity to produce common outcomes for the benefit of Society – what ecologists call the ecosystem services – over time.

In the geosciences domain, digital ecosystems are called to enable the coevolution (i.e., the complex interplay between competitive and cooperative business strategies) of geosciences public and private organizations around the new opportunities and capacities offered by the digital transformation of society – Internet, big data, and computing virtualization processes represent some of the main engines of innovation, rising an entirely new type of geosciences ecosystems. A successful geosciences digital ecosystem enables the implementation of big data value chains by bringing together data owners, data analytics companies, skilled data professionals, cloud service providers, companies from the user industries, venture capitalists, entrepreneurs, research institutes, and universities. Geosciences digital ecosystems promote the necessary actions to allow these stakeholders to interact seamlessly within the digital market, leading to business opportunities, easier access to knowledge and resources (European Commission 2015). Three framework requirements are important to enable the formation and effective functioning of a digital ecosystem (Scott 2013):

1. The regulatory requirement – laws, policies, standards, and agreements that control the structure of the ecosystem components and their relations.
2. The institutional or organizational cultural requirement – the set of common social and cultural values that characterize the ecosystem. These values inevitably propel and restrain the behaviors of actors in the ecosystem.
3. The digital capabilities requirement – the key enabling technologies and interoperability services (e.g., computing, analytical, and communication services) that are utilized to introduce new actors (i.e., components) to the ecosystem.

Ecosystem Principles, Patterns, and Governance

To achieve an effective geospatial digital ecosystem, a typical implementation framework must consider a set of principles, development patterns, and governance styles (Nativi et al. 2021).

Evolvability and Resilience: A geosciences digital ecosystem operates in a highly dynamic environment in which technology, policies, and use needs are in constant evolution. Some of the recommended patterns are as follows:

- Implement a *High Flexibility and Modularity* level to decouple the enterprise digital systems (constituting the ecosystems) from the geosciences applications (e.g., digital twins) enabled by the ecosystem.
- Keep *Independence* from any specific provider, technology, or license – to secure industrial competitiveness and defend the value that the ecosystem may have at local/regional level, creating a social benefit.
- Preserve and facilitate the *Coevolution* of the ecosystem components, preserving their “diversity.” This increases the ecosystem resilience.
- Provide *Equal Opportunities* of access to the ecosystem and affordability for small organizations and across the digital value chain.
- Provide a *Metasystemic governance* of the ecosystem to govern its evolutions, adaptations, mutations, and strains.

Emergent Behavior (Maier 1998) of the Geosciences Digital Ecosystem As a Whole: For the enterprise systems belonging to an ecosystem, the aim is to create a value that is greater than (or different from) the value that they would have without being part of the ecosystem. Such utility can emerge autonomously, as the result of unpredictable interactions, or it can be planned, facilitated, and governed by an internal or external agent – e.g., normative and regulatory frameworks, and management intervention. To carefully plan and manage a geoscience digital ecosystem, it is useful to consider the following patterns (Nativi et al. 2021):

- *Belonging* versus *Autonomy*: A digital ecosystem is based on the paradox of autonomous and yet cooperative entities (DiMario et al. 2009). To belong to a digital ecosystem, an enterprise system must accept a certain level of reduction in its degree of autonomy and optimization, which can vary over time.
- *Common* versus *Enterprise value*: The emergence of a common ecosystem value must preserve (and hopefully reinforce) the utility of the contributing enterprise systems. This is achieved by allowing the enterprise systems to maintain a good degree of autonomy and diversity.
- *Critical mass*: A critical mass of services and users is needed to guarantee a return-of-investments and facilitate the ecosystem resilience.

- *Cost-effectiveness*: Duplication of efforts and resources must be avoided by: making use of shared components, leveraging economies of scale, and using standard and widely adopted technologies.

Enterprise Systems Dispersion: The enterprise systems, constituting a geoscience digital ecosystem, are generally disperse (i.e., geographically distributed and heterogeneous) and cope with big data. Therefore, the resulting ecosystem must deal with the challenges characterizing the “big data” cyber-physical realm and implement appropriate strategies to manage volume, velocity, variety, and value of data from multiple sources in a scalable way. To do that, presently, network computing is one of the most critical areas. Therefore, the digital ecosystem must optimize the nexus of four (often conflicting) minimization factors: data movement minimization, data replication minimization, analytics time lapse minimization, and energy consumption minimization, while preserving distributed data ownership. Useful patterns to be considered are:

- *Computing continuum* (osmotic computing): This paradigm introduces a holistic distributed system abstraction enabling the deployment of lightweight microservices on resource-constrained platforms at the network edge, coupled with more complex microservices running on large-scale datacenters (Baresi et al. 2019). Applications can choose to execute parts of their logic on different infrastructures that constitute the continuum (e.g., mobile, edge, fog, and cloud computing), with the goal of minimizing latency and energy consumption while maximizing availability.
- *Optimizing vs Satisficing design strategy* (Simon 1956): Ecosystem engineering is concerned more with satisficing performance (a decision-making strategy aiming at finding a satisfying and sufficing solution), rather than optimization of system performance. In a complex environment (like the geospatial digital ecosystem one) the optimization strategy can be too expensive and time-consuming and the simple selection of the first “good enough” option can result more rational – e.g., for real-time decision-making.

Effective governance: The success of a digital ecosystem largely depends on the appropriate governance of the ecosystem as a whole. The enterprise systems constituting a geospatial digital ecosystem are existing and autonomous structures, managed by different organizations. The governance of a geospatial digital ecosystem must define and apply the set of rules and principles that will help steer the ecosystem evolution and effectiveness through the many changes occurring at the political, social, cultural, scientific, and

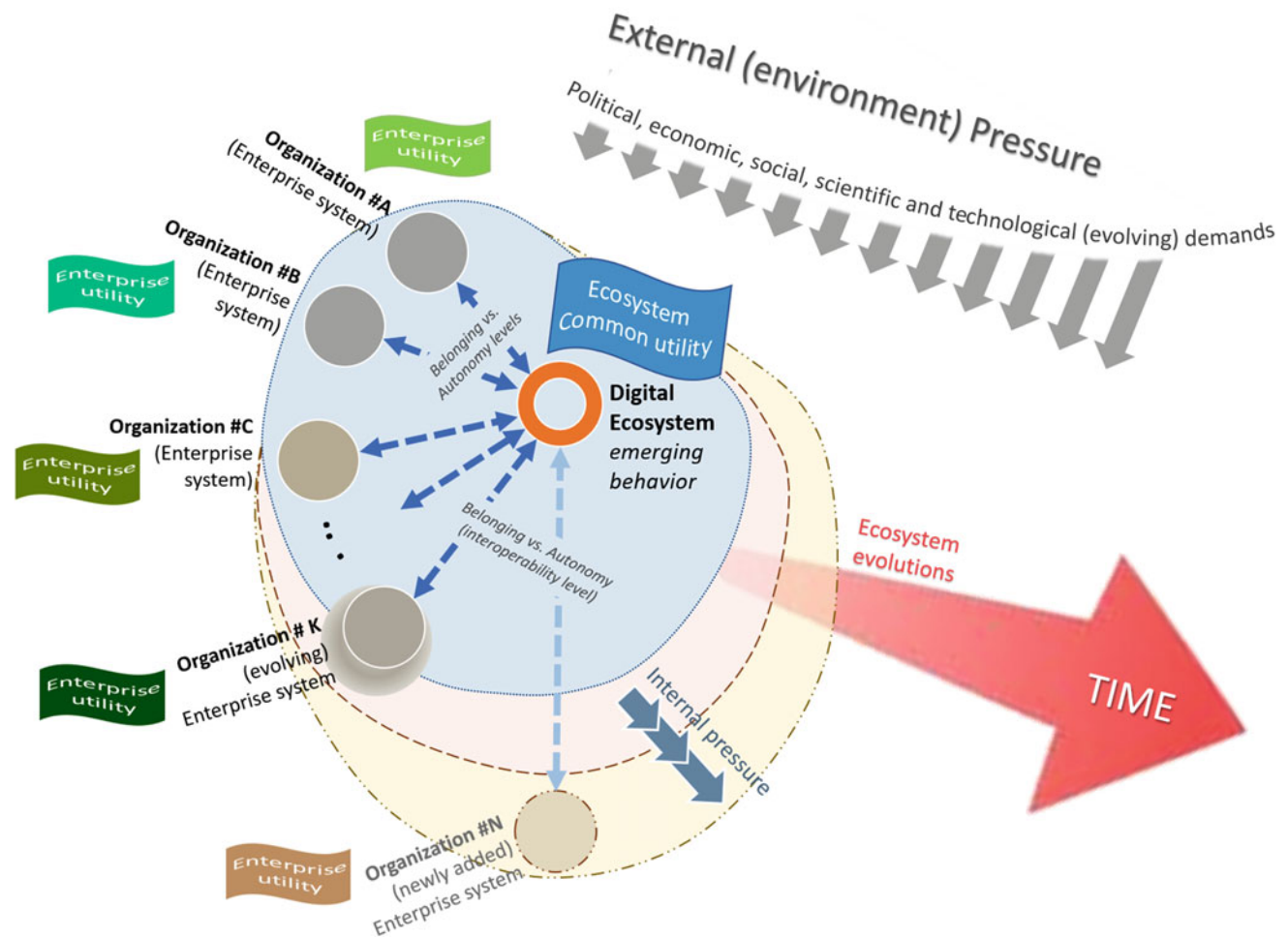
technological environment where it operates. Different governance styles are possible (Holt et al. 2015; Nativi et al. 2021):

- *Virtual governance*: There is no central management authority and no centrally agreed-upon purpose for the ecosystem.
- *Acknowledged governance*: There is a central management organization without coercive power to run the ecosystem, but enterprise systems interact more or less voluntarily to fulfill agreed-upon central purposes.
- *Collaborative governance*: Like in the following *Directed* governance style, there are recognized objectives, a designated ecosystem manager, and specific resources allocated for the ecosystem. However, like in the previous *Acknowledged* case, the normal operational mode of the enterprise systems is not subordinated to the central managed purpose and they retain their independent ownership, objectives, funding, and development and sustainment approaches.
- *Directed governance*: An integrated ecosystem is built to fulfill specific purposes and is centrally managed to continue to fulfill those purposes, as well as any new ones the ecosystem owners might wish to address.

Some Significant Research and Innovation Aspects

Evolutionary Aspect

Geospatial digital ecosystems are subject to a constant pressure caused by internal changes (e.g., enterprise system evolutions, new system addition) as well as external ones (e.g., political, socioeconomic, and technological changes). These inevitable modifications request the ecosystem to continuously evolve. To avoid controlling those changes would result in being very disruptive for the provision of the ecosystem common utility services and, hence, hampering the survival of the ecosystem itself. Therefore, a geospatial digital ecosystem must be designed, managed, and operated as a highly dynamic system-of-systems (Nativi et al. 2021) as depicted in Fig. 1, where, in addition to the external pressure, internal changes are represented by the organizations #K and #N that evolves over time and decides to join the ecosystem, respectively. To allow a geospatial digital ecosystem resilience and evolution, the governance must define and manage the acceptable behaviors of enterprise systems, the acceptable boundary of their time evolutions (i.e., belonging versus autonomy levels boundary), and the consequent communication and interoperability instruments, which must be supported by the digital ecosystem, to acknowledge and control such changes.



Geosciences Digital Ecosystems, Fig. 1 The dynamic nature of geospatial digital ecosystems (Nativi et al. 2021)

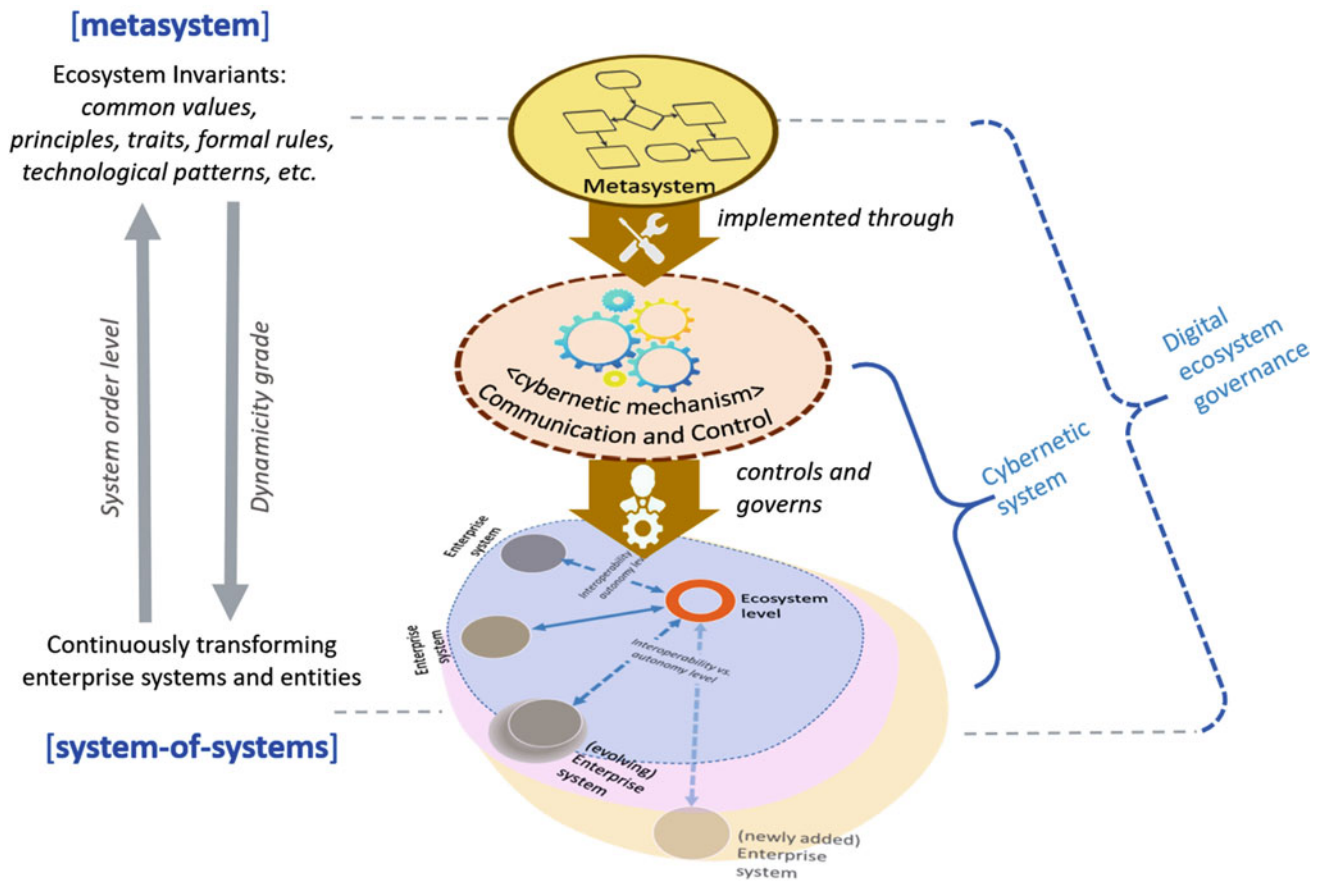
Metasystemic Aspect

Being a dynamic supersystem, a geospatial digital ecosystem must detect changes and respond to them. In other words, a geospatial digital ecosystem must be implemented as a viable system, which recognizes as invariant the process of change, rather than the component or services that change (Turchin 1995). A digital ecosystem becomes a collection of parts that are dynamically related and such dynamicity must be carefully formalized and governed, through the development of efficient communication and control functions (cybernetic mechanism). To survive, therefore, a geospatial digital ecosystem must realize a metasystem transition (Turchin 1995). As showed in Fig. 2, the metasystem level stands “above” and “beyond” the controlled system-of-systems: i.e., the dynamic set of connected and transforming enterprise systems. The metasystem communicates through a governance language (a metalanguage) that is based on a set of invariable elements – i.e., principles, interoperability rules, technology patterns, common values, etc. With reference to Fig. 2, the main elements and their relationships are (Nativi et al. 2021):

- Metasystem schema – including the geospatial digital ecosystem invariants.
- Governance style and Cybernetic mechanisms – to apply the metasystem rules.
- System-of-systems architecture topology – largely depending from, and governed by the governance style, the cybernetic mechanisms, and the agreed digital ecosystem common utility.

Summary and Conclusions

In a digital society, the interconnection between the physical and the digital world has become almost complete. To monitor, understand, and simulate our planet behavior, public and private stakeholders can now interact in the “cyber-physical” world, in a more effective and intense way. This represents an important paradigm shift (in respect to the traditional data exchange approach), which requires a cultural, organizational, economic, and legal evolution.



Geosciences Digital Ecosystems, Fig. 2 The main elements implementing a viable geospatial digital ecosystem (Nativi et al. 2021)

On the other hand, the increasing exchange of information, occurring in the cyber-physical domain among a dynamic set of autonomous systems, escalates the interaction complexity and introduces the need to preserve the effectiveness of the overall system for the benefit of all the stakeholders. Natural ecosystems with their capability to solve complex and dynamic problems through evolution and self-organization provide an interesting paradigm to address the complexity in an ever-changing cyber-physical environment.

Geoscience digital ecosystems leverage such organizational model to assure the generation and sharing of environmental knowledge for the benefit of Society through collaboration and competition of autonomous systems. As in the case of natural ecosystems, geoscience digital ones are subject to internal and external disruptive changes that threaten its effectiveness in providing the ecosystem services required by Society. For this reason, geospatial digital ecosystems need a governance assuring curation and protection. Such a governance should include a metasystem level implementing the cybernetic functionalities of communication and control that assure the desired invariance of an ecosystem that lives in a highly dynamic environment.

Then, the ecosystem is free to evolve, while keeping constant its values, principle, and traits for the benefit of the community.

Today, geospatial digital ecosystems provide those scalable analytical and interpretation services that are required to develop advanced simulation and projection models, notably the Digital Twins of the Earth (Nativi and Craglia 2021).

Cross-References

► Digital Twins of the Earth

Bibliography

- Baresi L, Mendonça DF, Garriga M, Guinea S, Quattrocchi G (2019) A unified model for the Mobile-edge-cloud continuum. *ACM Trans Internet Technol* 19:1–21
- Blew R (1996) On the definition of ecosystem. *Bull Ecol Soc Am* 77: 171–173
- Briscoe G, De Wilde P (2006) Digital ecosystems: evolving service-oriented architectures. In: *IEEE first international conference on bio inspired mOdelS of NETwork*. Information and Computing Systems

- (BIONETICS). Available: https://www.researchgate.net/publication/2213704_Digital_Ecosystems_Evolving_Service-Oriented_Architectures
- DiMario MJ, Boardman JT, Sauser BJ (2009) System of systems collaborative formation. *IEEE Syst J* 3:360–368
- European Commission (2015) A digital single market strategy for Europe-COM (2015) 192 final. European Commission, Luxembourg
- Holt J, Perry S, Payne R, Bryans J, Hallerstedte S, Hansen FO (2015) A model-based approach for requirements Engineering for systems of systems. *IEEE Syst J* 9:252–262
- Jansen S, Brinkkemper S, Cusumano M (2013) Software ecosystems: analyzing and managing business networks in the software industry. Edward Elgar Publishing Limited, Cheltenham
- Maier MW (1998) Architecting principles for system of systems. *Syst Eng* 1:267–284
- Moore JF (1996) The death of competition: leadership & strategy in the age of business ecosystems. Harper-Business, New York
- Nativi S, Craglia M (2021) Digital Twins of the Earth Encyclopedia of Mathematical Geosciences, Encyclopedia of Earth Sciences Series, Springer Nature Switzerland. https://doi.org/10.1007/978-3-030-26050-7_458-1
- Nativi S, Mazzetti P, Craglia M (2021) Digital ecosystems for developing digital twins of the earth: the destination earth case. *Remote Sens* 13:2119–2144
- Scott WR (2013) Institutions and organizations: ideas, interests, and identities. SAGE, Thousand Oaks
- Simon HA (1956) Rational choice and the structure of the environment. *Psychol Rev* 63:129–138
- Turchin VF (1995) A dialogue on metasystem transition. *World Futures* 45:5–57

Geostatistical Seismic Inversion

Leonardo Azevedo and Amílcar Soares
CERENA, DECivil, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

Definition

Subsurface rock properties models play a vital role in different geosciences fields such as hydrocarbon reservoir modeling and characterization, groundwater management, and mining resources and reserves evaluation. These models require the integration of all the available data, which often comprises direct and indirect measurements. These data sets comprise borehole and geophysical data. The main problem of the integration of these data in the typical forward models, joint estimation or simulation, is the large difference between their spatial support. Hence, integrating and predicting the spatial distribution of these properties from geophysical data requires solving a challenging inverse problem. In geophysical inverse problems, the model parameters – subsurface rock and elastic properties – are perturbed until the predicted geophysical data match the observed one. In geostatistical geophysical inversion methods, geostatistical simulation tools are used to perturb and update the model parameters

due to their ability to integrate data with different resolutions, to account for the geological patterns, and to assess spatial uncertainty of the predictions.

Geophysical Inversion

Geophysical survey methods are effective tools for imaging subsurface rock properties in different contexts, particularly in deeper and complex geological environments. In geosciences, seismic reflection surveys are widely used to model and characterize the subsurface properties due to the versatility of the method to image the subsurface at different depths with distinct levels of detail. Briefly, seismic data represent acoustic waves reflected at interfaces of geological media with different elastic properties. The observed geophysical data, $\mathbf{d}_{\text{obs}}$, are linked to the subsurface rock properties, $\mathbf{m}$ (e.g., porosity and fluid saturations) through a mathematical operator $\mathbf{F}$ (i.e., the forward model):

$$\mathbf{d}_{\text{obs}} = \mathbf{F}(\mathbf{m}) + \mathbf{e}, \quad (1)$$

where $\mathbf{e}$ represents measurement errors and approximations of the physical phenomena during data processing. In seismic inversion, $\mathbf{F}$ is commonly approximated with the convolutional model

$$\mathbf{A} = \mathbf{r} * \mathbf{w}, \quad (2)$$

where the recorded seismic amplitude ($\mathbf{A}$) represents the convolution of the reflection coefficients ($\mathbf{r}$) with a known wavelet ($\mathbf{w}$). The reflection coefficients depend on the subsurface elastic properties (i.e., rock density and P- and S-wave propagation velocities). The wavelet is a numerical representation of the seismic pulse generated at the source location.

In geophysical inversion, one knows the solution $\mathbf{d}_{\text{obs}}$ and aims to predict the model parameter ($\mathbf{m}$) that better matches $\mathbf{d}_{\text{obs}}$. This is a challenging inverse problem as $\mathbf{F}^{-1}$ is not well defined. Geophysical inverse problems are ill-posed and non-unique solution due to bandwidth and resolution limitations of the recorded data, physical assumptions, and measurement errors. Consequently, inverse solutions, retrieved by any of the many available inverse methodologies, always involve a certain degree of uncertainty that needs to be assessed during their interpretation (Tarantola 2005).

Seismic inversion may be solved under deterministic or stochastic frameworks (Bosch et al. 2010). Deterministic methods allow retrieving a single best-fit model that matches the observed data in the least squares sense. The predicted models are always a smooth representation of the subsurface, and the uncertainty assessment from deterministic inverse solutions is hardly suitable for highly nonlinear inverse problems. On the other hand, stochastic seismic inversion methods allow predicting multiple possible models of the subsurface,

allowing to assess the uncertainty of the predictions. The uncertainty is represented by the full posterior distribution of the model parameters, assuming a prior probability distribution and given a set of observed data. There are several alternative stochastic inversion methods, which might be divided in two main categories.

Bayesian linearized seismic inversion methods (e.g., Buland and Omre 2003, Grana and Della Rossa 2010) allow obtaining an analytical expression for the posterior distribution under Gaussian, or multi-Gaussian, assumptions about the prior distribution of the model properties and the error distribution present in the data and the linearization of the forward model. Due to these assumptions, the solution is mathematically tractable but lacks real exploration of the uncertainty space in lieu of exact prior information.

Alternatively, stochastic sampling and optimization methods, which require the spatial coupling of the subsurface property of interest, might be used to solve the seismic inversion problem. A comprehensive review about these seismic inversion methods can be found in Doyen (2007) and Bosch et al. (2010). Iterative geostatistical seismic inversion is a class of stochastic seismic inversion algorithms. Due to the ability to include direct observations and incorporate the spatial information about the expected continuity patterns, these inversion approaches use geostatistical methods, namely, stochastic sequential simulation and co-simulations. These are generally used to generate and update a set of models until a satisfactory match between observed and predicted seismic data is achieved (Azevedo and Soares 2017).

Depending on the observed data, the existing inversion methods might predict the spatial distribution of elastic properties such as acoustic and elastic impedances and density, P- and S-wave velocities or be inverted directly for rock properties like facies, porosity, fluid saturation, and mineral fraction. The former requires the integration of a rock physics models to link the rock with the elastic domains (Mavko et al. 2003).

Trace-by-Trace Geostatistical Seismic Inversion

Bortoli et al. (1993) introduced the pioneering work on geostatistical seismic inversion proposing a sequential trace-by-trace approach to invert fullstack seismic for acoustic impedance. Each seismic trace (i.e., a vertical column of grid cells within the inversion grid) is visited individually following the framework of a stochastic sequential simulation. At each step along a predefined random path, N_s realizations of acoustic impedance are simulated using sequential Gaussian simulation (Deutsch and Journel 1998) taking the well-log data and previously visited/simulated traces into account. For each simulated impedance trace, the corresponding reflection coefficient is computed and convolved with a wavelet. This step results in a set of N_s synthetic seismic traces. Each synthetic trace is compared against the

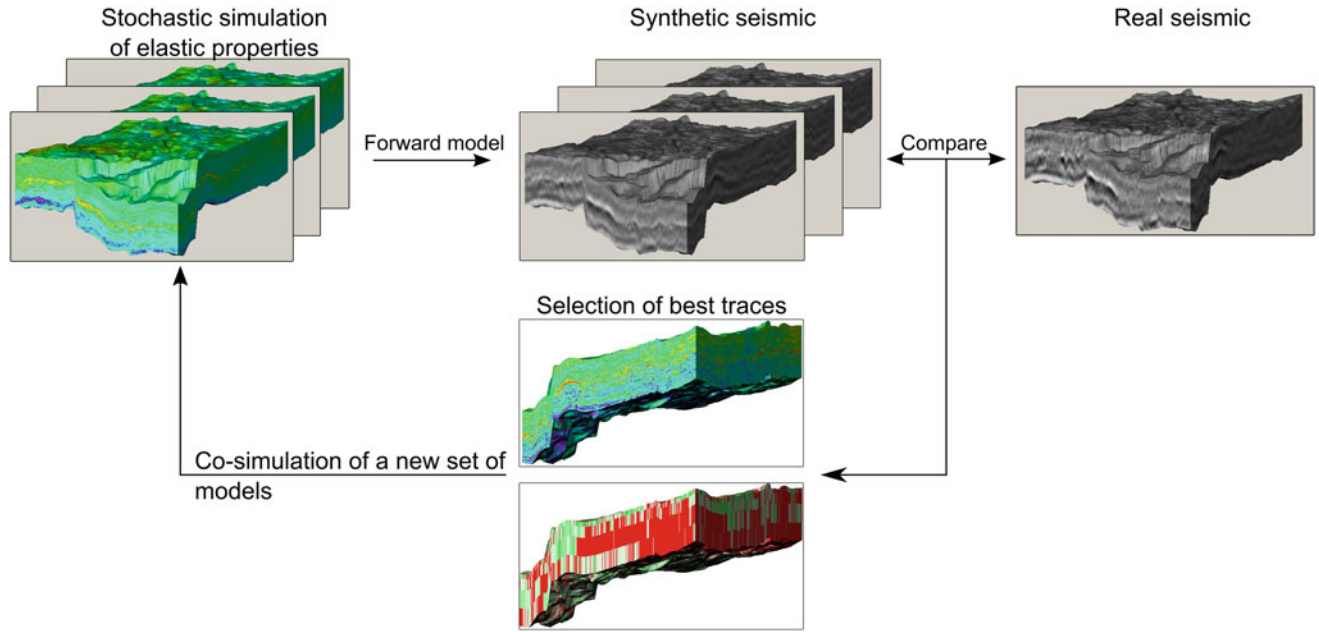
observed seismic trace for the same location given an objective function such as the Pearson's correlation coefficient. The acoustic impedance realization that produces the lowest mismatch between the real and the synthetic seismic traces is integrated in the inversion grid as conditioning data for the simulation of the next location following the predefined random path. Different runs produce variable inverted models as the random path changes and, consequently, modify the conditioning data at each trace location. The variability among a set of inverted models allows the assessment of the spatial uncertainty related with the property of interest.

One of the main drawbacks of trace-by-trace stochastic seismic inversion methodologies is related to those areas of low signal-to-noise ratio. In areas of poor seismic signal, the sequential trace-by-trace approaches impose inverted models that fit the observed noisy seismic reflection data. As the simulated trace is assumed as true data for the next steps, this can lead to the spread of the noisy values.

Global Geostatistical Seismic Inversion

Global iterative geostatistical seismic inversion methods differ from the trace-by-trace approaches since the inversion grid is simulated at once for all the locations (Soares et al. 2007). This family of methods is based on two main ideas common to the different available methods. Direct sequential simulation and co-simulation are the preferable methods to generate sets of models at each iteration. The convergence of the iterative procedure to a known solution is assured with an optimization method based on a genetic algorithm. The elastic traces that at a given iteration ensure synthetic seismic traces with the highest similarity against the observed one are selected as conditioning data (i.e., secondary information) in the generation of a new set of models in the subsequent iteration. Figure 1 summarizes the main steps of global geostatistical seismic inversion.

In geostatistical elastic inversion, the first iteration starts with the simulation of a set of N_s models of the elastic properties of interest given the existing well-log data and a given spatial pattern as revealed by a variogram or spatial covariance model. A variogram model is a mathematical tool that describes the expected spatial continuity pattern of the property in study. Each simulated realization, by reproducing the variogram model, ensures the same spatial continuity pattern of the main properties. These realizations are then forward modeled to generate synthetic data. Depending on the observed data, the forward model might comprise the calculation of the normal incidence or angle-dependent reflection coefficients, which are then convolved with a wavelet (Eq. 1). Synthetic and real seismic traces are compared taking into account the waveform and amplitude content of the signal. Various equations might be used. Here we show a single equation that is sensitive to both parameters:



Geostatistical Seismic Inversion, Fig. 1 Schematic representation of the general workflow used in global iterative geostatistical seismic inversion methods

$$S = \frac{2 \sum_{k=-n}^n (d_{t+k} d_{t+k}^*)}{\sum_{k=-n}^n (d_{t+k})^2 + \sum_{k=-n}^n (d_{t+k}^*)^2}, \quad (3)$$

where d_t and d_t^* are the observed and predicted seismic traces at sample t and n is the size of a moving window used to compute the similarity between both traces. As in a correlation coefficient S varies between -1 and 1 , for implementation purposes, the negative values are truncated at zero. The elastic traces that produce the highest S at a given iteration are stored along with the similarity value in two auxiliary volumes. These volumes will be used as secondary variables in the stochastic sequential co-simulation of a new set of models in the following iterations. Regions of the inversion grid where S is high will exhibit a similar spatial pattern as the auxiliary elastic volumes. On the contrary, in regions where S is small, the auxiliary volume will have low impact. In other words, S controls the degree of assimilation of the observed data in the generation of elastic models in the next iteration. The iterative procedure is over when a given number of predefined iteration is reached or when the global similarity between observed and predicted data is above a user-defined threshold.

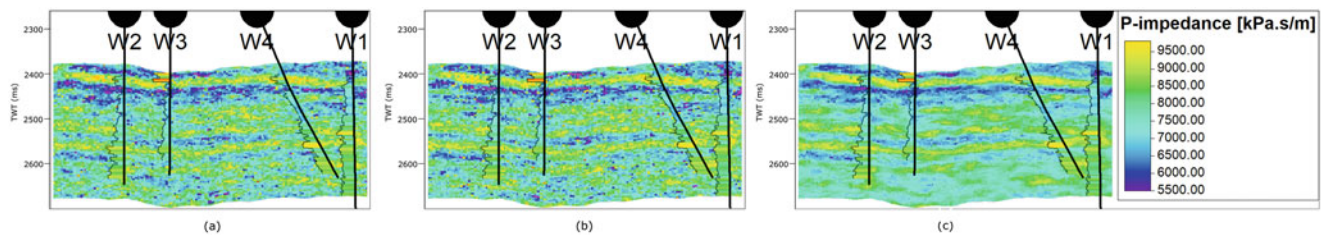
Once the convergence is reached, all the co-simulated models in the last iteration produce synthetic seismic highly similar to the observed one. Areas of poor signal-to-noise ratio in the observed data or areas where the wavelet is not representative of the recorded seismic tend to remain

unmatched during the entire iterative procedure and represent areas of high uncertainty. This ensemble of models might be used to assess the spatial uncertainty of the predictions by computing, for example, the pointwise variance or the inter-quantile distance.

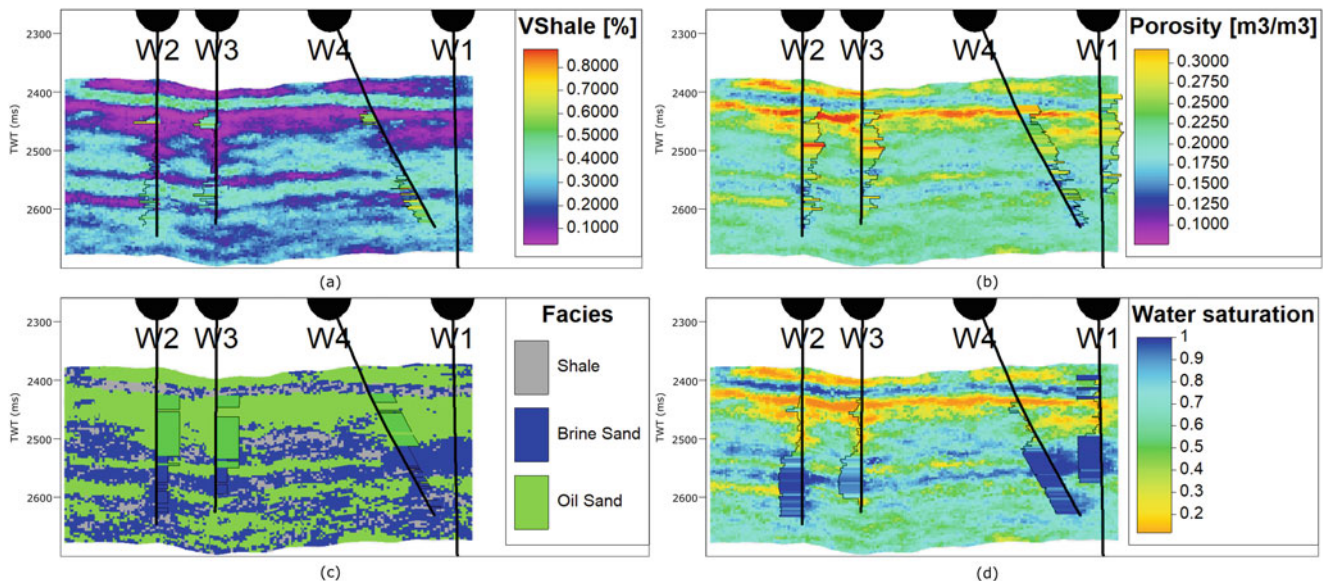
All the simulated and co-simulated models share some common characteristics. They reproduce exactly the well-log data at the well locations, the marginal and joint distribution of the properties of interest, and the spatial continuity patterns as revealed by the variogram models. Also, depending on the variogram model used during the model simulation, the inverted models show higher resolution than the recording seismic, allowing to characterize layer at and below the seismic resolution.

Geostatistical seismic inversion has been developed to invert fullstack seismic for acoustic impedance, partial angle stacks for acoustic (Fig. 2) and elastic impedances, and partial angle stacks and angle gathers for density P-wave and S-wave velocities. As the goal of a geoscientist is to model the subsurface rock properties, the inverted elastic models are often used as starting point for this step of the geo-modeling workflow. However, the derived rock properties are not directly linked to the observed geophysical data, which might result in non-plausible subsurface models.

Recently, geostatistical seismic inversion methods have been extended to predict rock properties such as porosity, fluid, mineral fractions, and facies directly from the seismic data (Azevedo et al. 2017). In these methods, the iterative procedure generates the rock properties through geostatistical



Geostatistical Seismic Inversion, Fig. 2 Vertical well sections extracted from: (left and middle) two realizations of I_p generated during the last iteration of the geostatistical seismic inversion and the pointwise average model of all the realizations generated during the last iteration



Geostatistical Seismic Inversion, Fig. 3 Vertical well sections extracted from the pointwise average model of all the realizations generated during the last iteration for volume of shale (Vshale), porosity, facies, and water saturation

simulation. The simulated rock models are then converted into elastic properties using a calibrated rock physics model (Mavko et al. 2003). A rock physics model is a set of mathematical equations that links the subsurface rock properties to their corresponding elastic response. Rock physics models differ in terms of assumptions and complexities of the physical and mathematical formulation. A simple rock physics model might include simple empirical relations inferred from laboratory experiments or be more complex and based on the Hertz-Mindlin contact theory or inclusion theory. A detailed description of rock physics models and workflows is available in Mavko et al. (2003). Contrary to geostatistical elastic inversion, in these methods, the predicted subsurface rock properties are directly linked to observed seismic data (Fig. 3).

Summary

Iterative geostatistical geophysical inversion is versatile methods to predict subsurface rock properties from

geophysical data. These techniques have had a significant evolution both in terms of applications and in new methodological approaches. New field of applications, like geotechnical characterization of the near-surface and for groundwater management, are using geostatistical geophysical inversion for better characterizing the spatial patterns of geomechanical and geotechnical properties of the subsurface.

As for the new methodological approaches, with the generalization of spatial data science methods, these can be integrated in the geostatistical seismic inversion workflow with two complementary objectives. For example, deep learning methods like generative adversarial networks are promising techniques to model categorical variables such as facies, and its inclusion in geophysical inversion might open the door to the integration of greater geological information content in the inversion procedure and be used to falsify unreliable geological scenarios.

In addition, a new generation of the self-learning methods has been developed to tune and optimize the inversion meta-parameters as defined by the local variogram parameters (i.e., the local variogram dip, azimuth, and ranges) and the

marginal and joint distributions of the inverted properties. Briefly, the data mismatch is used not only for the selection of the best elastic or rock property traces at the end of each iteration but also to update the meta-parameters using a self-learning approach.

Cross-References

- [Bayesian Inversion in Geoscience](#)
- [Inversion Theory](#)
- [Inversion Theory in Geoscience](#)

Bibliography

- Azevedo L, Soares A (2017) Geostatistical methods for reservoir geophysics. Springer, Berlin
- Bortoli LJ, Alabert F, Journel A (1993) Constraining stochastic images to seismic data. In: Geostatistics. Troia'92. Kluwer
- Bosch M, Mukerji T, Gonzalez EF (2010) Seismic inversion for reservoir properties combining statistical rock physics and geostatistics: a review. *Geophysics* 75(5):75A165–75A176
- Buland A, Omre H (2003) Bayesian linearized AVO inversion. *Geophysics* 68(1):185–198
- Deutsch CV, Journel AG (1998) GSLIB: geostatistical software library and user's guide, 2nd edn. Oxford University Press, Oxford
- Doyen P (2007) Seismic reservoir characterization. EAGE
- Grana D, Della Rossa E (2010) Probabilistic petrophysical-properties estimation integrating statistical rock physics with seismic inversion. *Geophysics* 75(3):O21–O37
- Mavko G, Mukerji T, Dvorkin J (2003) The rock physics handbook. Cambridge University Press
- Soares A, Diet JD, Guerreiro L (2007) Stochastic inversion with a global perturbation method. *Petroleum geostatistics*, EAGE, Cascais, Portugal (September 2007), pp 10–14
- Tarantola A (2005) Inverse problem theory and methods for model parameter estimation. Society for Industrial and Applied Mathematics, pp 1–54

Geostatistics

H. S. Pandalai¹ and A. Subramanyam²

¹Department of Earth Sciences, Indian Institute of Technology Bombay, Mumbai, India

²Department of Mathematics, Indian Institute of Technology Bombay, Mumbai, India

Definition

Geostatistics is the collection of techniques that deals with spatio-temporal phenomena using data assumed to be realizations of a collection of random variables generated via a random process. Typically, the data are such that the joint probability distributions of this random process cannot be

readily modeled. The geostatistical approach is characterized by modeling the random process by making appropriate hypotheses that permit estimation of second-order or higher moments of bivariate and multivariate distributions of the random variables. The two broad problems addressed by geostatistics are those of prediction and simulation of random variables on different supports, and their functions.

Introduction

Understanding natural phenomena often requires models to be built based on observations. A large part of comprehending and characterizing natural phenomena is achieved by making numerical observations on them. Mathematical models are made from such numerical observations. The subject of mathematical geosciences concerns the study of mathematical models developed for the geosciences. Often, numerical observations on natural phenomena have an element of randomness in them owing to the intrinsic nature of the phenomenon and due to random errors involved in making the observations. Statistical methods are used to make appropriate models that take into account such randomness. Statistical techniques such as inference, design of experiments, multiple regression, categorical data analysis, and other multivariate techniques are used in all branches of scientific enquiry, including the Earth Sciences. The term *Geostatistics* is, however, used for a collection of statistical techniques originally developed to address questions that arose in the mining industry, and later came to have much wider applications in several branches of scientific study. For example, a volume V (called a “panel”) which is a designated part of a mineral deposit is to be mined out. The volume V is generally mined out in smaller portions (called “blocks”) of volume v . The contents of such blocks are sent for extraction to a mill or to a waste dump depending upon whether their average grades exceed a threshold determined by economic parameters. Questions such as (i) what is the average grade of the panel, (ii) what proportion of blocks in a panel can be sent to the mill for extraction, or (iii) what is the average grade of material from panels that are sent to the mill, are of interest in the mining industry. Similar questions also arise in other fields of study such as in petroleum reservoir studies, agriculture, forestry, fisheries, and environmental studies, to name a few.

A feature that distinguishes geostatistics from many other statistical techniques is that it deals with the modeling of spatial phenomena with the aim of capturing the nature of their essential randomness as well as their inherent spatial structure. Thus, geostatistics is an aspect of the general theory of stochastic processes which deals with the study of collections of random variables. The general study of stochastic processes includes the study of time-series, Markov processes, diffusion processes, point processes, and others besides geostatistics.

Geostatistics had its origins in the observations of Daniel G. Krige in 1951 and 1952 that the grades of a panel estimated by the average grades of observed values in core-samples in and around the panel was often biased. In particular, panels estimated as having high grade often actually produced lower grade of ore when mined out. He sought to remedy this by assigning weights to the observations depending upon whether they were inside or outside the panels based on regression of the average grade of these groups of samples using historical mining data. Georges Matheron who had come across similar problems worked on a mathematical formulation for solving the problem and developed a more general method for assigning weights to observed data in the process of estimation. Matheron (1960, 1963) called this more generalized approach “kriging” (or “krigeage” in French) in honour of Daniel Krige.

The basis of the original technique developed by Matheron is a certain set of assumptions (called second-order stationarity) that will be discussed later in detail. Since the introduction of “kriging” in this form, it has been further generalized and modified to suit several other situations involving various other assumptions. These techniques involve the modeling of stochastic processes under various constraints.

The problem of “change of support” is a problem that was first identified in geostatistics and arose initially in applications concerning the mining industry. For example, when grades of core-samples are the observations available and it is desired to obtain estimates of functions of average grades defined over volumes different from that of the core-samples. Defining the relationship between the average grade of a panel of volume V with the help of core-samples of volume v constitutes a unique problem in the realm of stochastic processes. This problem is called the “problem of change of support” and is a contribution of geostatistics to the study and modeling of stochastic processes.

Another significant line of work in geostatistics has been that of developing techniques for generating multiple realizations of a stochastic process with a given characteristic – a technique known as simulation. With the increasing availability of superior computing power, great advances have been made in the simulation of stochastic processes, providing solutions to many problems addressed by geostatistics.

Statistical Formulation

The statistical approach to data analysis is to assume that observations are realizations of random variables (RV s). Let x be a point in D , the region of study, and let $s(x)$ be a small volume centered at x . Usually the volume $s(x)$ is constant for all x . The average value over $s(x)$ of the characteristic of interest is attributed to the point x and denoted by $z(x)$. The

approach to probabilistic analysis is to consider that $z(x)$ observed at each x is the realization of an RV , $Z(x)$. This leads to a collection of RV s, $\{Z(x), x \in D\}$. Such a collection of RV s is known variously in literature as *stochastic process*, *random process*, *random field*, *random function*, etc. In geostatistics the term random function is used more commonly. In the sequel these terms are used interchangeably. The collection $\{z(x), x \in D\}$ of realizations of RV s or a realization of the random function $\{Z(x), x \in D\}$ is called a *regionalization*. The region of study D can be multi-dimensional. In particular it could be a subset of three-dimensional space thereby bringing in spatial aspects into the analysis. In other examples D could also include a dimension of time, in which case the analysis would be spatio-temporal.

In many applications of geostatistics, the average value $z_V(x)$ of the characteristic of interest over a volume V centered at location x is desired. Mathematically, $z_V(x)$ is given by $\frac{1}{|V|} \int_{V(x)} z(u) du$. This value $z_V(x)$ is the realization of the random variable $Z_V(x)$. The problem in geostatistics is often that of predicting the value of $z_V(x)$ given $z(x_1), z(x_2), \dots, z(x_N)$ which are observations at N locations in and around V . This problem can be solved using standard statistical techniques provided the joint distribution of the RV s, $(Z_V(x), Z(x_1), Z(x_2), \dots, Z(x_N))$, is known. When this joint distribution is known, the conditional expectation, $E[Z_V(x) | Z(x_1), Z(x_2), \dots, Z(x_N)]$ is the function that minimizes the expected value of the squared error of prediction among all estimators of the type $g(Z(x_1), Z(x_2), \dots, Z(x_N))$, provided all the RV s have finite variance. This is the case, for example, if the random process is assumed to be Gaussian wherein the joint distributions can be shown to be multi-Gaussian.

In the sequel the random process (or random function) $\{Z(x), x \in D\}$ is abbreviated as Z . Probabilistically, the random function is completely specified only when all finite-dimensional joint distributions of collections of the type $\{Z(x_1), Z(x_2), \dots, Z(x_K)\}$ for $x_i \in D, i = 1, 2, \dots, K$ and for all $K = 1, 2, 3, \dots$ (termed, in geostatistics as the spatial law of the random function) are specified. This is the case when the random process is Gaussian. Unfortunately, it is most often not possible to specify the spatial law of random functions.

In practice, data is often only adequate to model the marginal distribution of $Z(x)$ under the assumption that all $\{Z(x), x \in D\}$ are identically distributed. To solve problems that arise in geostatistics, such as prediction problems, it is necessary to make reasonable assumptions on the random process Z that enable moments of higher order to be estimated from the data. In the study of stochastic processes, assumptions collectively termed “assumptions of stationarity” make probabilistic analysis mathematically tractable and form the basis of the mathematical model (Cressie 1991).

Stationarity

The assumptions of stationarity impose certain levels of homogeneity in the random process. A most stringent assumption on Z would be to assume that the joint distribution of $\{Z(x_1), Z(x_2), \dots, Z(x_N)\}$ is the same as that of $\{Z(x_1 + \vec{h}), Z(x_2 + \vec{h}), \dots, Z(x_N + \vec{h})\}$ for all $\vec{h}$ and all N such that $x_i, x_i + \vec{h} \in D$, $i = 1, 2, \dots, N$. This assumption on Z is termed *strict stationarity*. Essentially this means that the joint distribution of $\{Z(x_1), Z(x_2), \dots, Z(x_N)\}$ is invariant under translation. A weaker assumption would be to require that this assumption is true for $N = 2$ and every fixed $\vec{h}$ (i.e., *bivariate stationarity*). In other words, under this hypothesis it is assumed that all pairs of random variables $\{Z(x_1), Z(x_2)\}$ have the same distribution as $\{Z(x_1 + \vec{h}), Z(x_2 + \vec{h})\}$ and this is true for any pair x_1, x_2 such that $\{x_1, x_2\}$ and $\{x_1 + \vec{h}, x_2 + \vec{h}\} \in D$. Equivalently, pairs $\{Z(x), Z(x + \vec{h})\}$ have the same distribution for all x , $x + \vec{h} \in D$. A further weakening of bivariate stationarity is achieved by requiring that only the first and second moments of pairs $\{Z(x), Z(x + \vec{h})\}$ are the same for all x . This assumption is called the assumption of *second-order stationarity*. Under second-order stationarity, the following are true:

- (i) $E[Z(x)] = m \quad \forall x$
- (ii) $Var[Z(x)] = E[Z(x) - m] = \sigma^2 \quad \forall x$
- (iii) $Cov[Z(x), Z(x + \vec{h})] = E[(Z(x) - m)(Z(x + \vec{h}) - m)]$
 $= C(\vec{h}) \quad \forall x$

Second-order stationarity implies $\frac{1}{2}E[(Z(x) - Z(x + \vec{h}))^2] = \gamma(\vec{h}) \quad \forall x$ where $\gamma(\vec{h})$ is called the *semivariogram*. Clearly $C(0) = \sigma^2$ and $\gamma(\vec{h}) = C(0) - C(\vec{h})$ under second-order stationarity.

It is interesting to note that there exist processes in which $\gamma(\vec{h})$ exists but $C(\vec{h})$ does not. Obviously, such processes are not second-order stationary. Matheron (1973) termed such random functions as *intrinsic random functions of order 0*. In such processes, mean and the variance of differences exist (i.e., $E[Z(x)] = m \quad \forall x$, and $Var(Z(x) - Z(x + \vec{h})) = E[(Z(x) - Z(x + \vec{h}))^2] = 2\gamma(\vec{h}) \quad \forall x$).

The assumptions of stationarity cannot be expected to be valid over a large region. Hence, in practice, often sliding neighborhoods in which stationarity is assumed are considered. This assumption is termed *quasi-stationarity* (Journel and Huijbregts 1978). When there is only a single realization of the random function, the assumptions of stationarity cannot be verified and it remains a mathematical model (Myers 1989).

Properties of the Covariance and Variogram Functions

By considering the nonnegative quantity variance of a linear combination of *RVs*, $Var\left(\sum_{i=1}^n \lambda_i Z(x_i)\right) = \sum_{i=1}^n \lambda_i^2 \sigma^2 + 2 \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j C(x_i - x_j) \geq 0$, it follows that the $n \times n$ matrix whose ij^{th} element is given by $C_{ij}(x_i - x_j)$ is nonnegative definite. Therefore, the covariance function $C(\vec{h})$ of every second-order stationary random function is a positive definite function. In the case of second-order stationary random functions, it can be seen that $\sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j C(x_i - x_j)$ can be written as $\left(C(0) \sum_{i=1}^n \lambda_i \sum_{j=1}^n \lambda_j - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(x_i - x_j)\right) \geq 0$. When $\sum_{i=1}^n \lambda_i = 0$, $-\sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(x_i - x_j) \geq 0$. Such a function $\gamma(h)$ is called a *conditionally negative definite function*. The variogram of every second-order stationary random function should therefore be a conditionally definite function.

Other important properties of the covariance of a second-order stationary random function are (i) $C(0) \geq C(\vec{h})$, (ii) $C(\vec{h}) = -C(-\vec{h})$. For the semivariogram function, (i) $\gamma(0) = 0$ and (ii) $\gamma(\vec{h}) = \gamma(-\vec{h})$. If $C(\vec{h}) = 0$, or equivalently $\gamma(\vec{h}) = \text{constant}$ for $\|\vec{h}\| > r$, then the smallest value of r for which this occurs is called the *range* of the semivariogram or covariance function and the value of the constant is called the *sill* of the semivariogram. When this occurs the random function Z is said to be *transitive*.

Nonstationary Random Functions and the Generalized Covariance Function

It is not always possible to assume that $m(x)$, the mean of $Z(x)$, is a constant at all locations x . In many natural situations it is more realistic and appropriate to assume that $Z(x) = Y(x) + m(x)$, where $Y(x)$ has zero mean and $m(x)$ is a deterministic (non-random) function of x that is commonly called “*drift*.” It may be postulated that $m(x) = \sum_{l=0}^L a^l f^l(x)$, where $f^0, f^1, \dots, f^L$ are some known basis functions. Since the function $m(x)$ is unknown, the coefficients $a^0, a^1, a^2, \dots, a^L$ are also unknown. It is, however, possible to solve equations of the type $\sum_{\alpha=0}^n w_\alpha f^l(x_\alpha) = 0, \quad l = 0, 1, \dots, L$ under the condition that $w_0 = -1$ (i.e., $\sum_{\alpha=1}^n w_\alpha f^l(x_\alpha) = f^l(x_0), \quad l = 0, 1, \dots, L$). The weights w_α are said to filter the basis functions. For such weights, $w_\alpha, \alpha = 0, 1, \dots, n$, $\sum_{\alpha=0}^n w_\alpha Z(x_\alpha)$ has zero mean even though the a^l s are unknown. If it is assumed that $m(x)$ is a polynomial, then the f^l s can be taken as appropriate

monomials. For example in $\mathfrak{R}^2$, $x \equiv (x_1, x_2)$ and $f^0(x) = 1$, $f^1(x) = x_1$, $f^2(x) = x_2$, $f^3(x) = x_1^2$, $f^4(x) = x_2^2$ and $f^5(x) = x_1 x_2$.

If the process $Z_W(s) = \left\{ \sum_{\alpha=0}^n w_\alpha Z(x_\alpha + s), \forall s \in D \right\}$ is a second-order stationary process for every set of weights w_α that satisfy $\sum_{\alpha=0}^n w_\alpha f^l(x_\alpha) = 0$, $l = 0, 1, \dots, L$, then the process $Z = \{Z(x), x \in D\}$ is said to be an *intrinsic random function* and the linear combination $\sum_{\alpha=0}^n w_\alpha Z(x_\alpha)$ is said to be an *authorized linear combination*. The choice of basis functions is restricted by the requirement that $\sum_{\alpha=0}^n w_\alpha Z(x_\alpha + s)$ is an authorized linear combination for all s . In other words, the weights w_α should ensure that $\sum_{\alpha=0}^n w_\alpha f^l(x_\alpha + s) = 0$, $\forall s$ and $l = 0, 1, \dots, L$ thereby filtering the basis functions evaluated at $x_\alpha + s$, $\alpha = 0, 1, \dots, n$, $\forall s$. Owing to convenience it is customary to use polynomials, although other functions such as exponential and trigonometric functions meet the requirement. An intrinsic random function $Z = \{Z(x), x \in D\}$ with a polynomial drift of order k is said to be an *intrinsic random function of order k , (IRF - k)*.

Consider a process $Z = \{Z(x), x \in D\}$ with $Z(x) = Y(x) + m(x)$, where $E[Y(x)] = 0$ and $m(x)$ is a polynomial of order k . Assume that $m(x) = \sum_{l=0}^L a^l f^l(x)$. If $m(x) = 0 \quad \forall x$, (or is a known constant) and $Y = \{Y(x), x \in D\}$ is a second-order stationary process. Then the process Z is stationary and is said to be an *intrinsic random function of order (-1)* (i.e., an *IRF(-1)*). An *IRF(-1)* process is identical to a second-order stationary random process. If $m(x)$ is an unknown constant $\forall x$, then $m(x) = a^0 f^0(x)$ where $f^0(x) = 1$. So $\sum_{\alpha=0}^n w_\alpha f^0(x_\alpha) = 0$ implies $\sum_{\alpha=1}^n w_\alpha = 1$. Such a process Z is said to be an *IRF - 0* and is identical to a process that has the property of intrinsic stationarity. For example, when $n = 1$, $w_0 = -1$ and $w_1 = 1$ and an authorized linear combination is the difference $(Z(x_\alpha) - Z(x_0))$. Such differences are also called “increments.”

Consider now an *IRF - k* process Z and an authorized linear combination $Z_W(s) = \sum_{\alpha=0}^n w_\alpha Z(x_\alpha + s)$. Clearly, $E[Z_W(s)] = 0 \quad \forall s$ and $Var[Z_W(s)] = V_W$, say, $\forall s$. It can be shown that the covariance of the *IRF - k* process Z can be written as $Cov[Z(x), Z(y)] = C_Z(x, y) = K(x - y) + \sum_{l=0}^L c^l(x) f^l(y) + \sum_{l=0}^L c^l(y) f^l(x)$ where $K(\vec{h})$ is a function, $c^l s$ are appropriate constants, and $f^l s$ are monomials (Delfiner 1982). The variance of an authorized linear combination can be written as $E[Z_W(s)]^2 = \sum_{\alpha=0}^n \sum_{\beta=0}^n w_\alpha w_\beta K(x_\alpha - x_\beta) = V_W \geq 0$ for a function K . This function $K(\vec{h})$ is called the

generalized covariance function of the random process Z (Chiles and Delfiner 2012). A property of this function is that it is symmetric and conditionally positive definite (i.e., positive definite conditional to $Z_W(s)$ being an authorized linear combination). For example, for an *IRF - 0* process, the function $K(\vec{h}) = -\gamma(\vec{h})$ which is conditionally positive definite since $\gamma(\vec{h})$ is known to be conditionally negative definite.

When the random process Z is written as $Z(x) = Y(x) + m(x)$ with $E[Y(x)] = 0$, the process $Y(x)$ may or may not be second-order stationary. It can be seen that

$$Cov[Z(x), Z(y)] = C_Z(x, y) = E[(Z(x) - m(x))(Z(y) - m(y))] = Cov[Y(x), Y(y)] = C_Y(x, y), \text{ say} \quad \text{and} \\ \frac{1}{2} Var[Z(x) - Z(y)] = \gamma_Z(x, y) = \frac{1}{2} E[(Z(x) - Z(y)) - (m(x) - m(y))]^2 = \frac{1}{2} Var[Y(x) - Y(y)] = \gamma_Y(x, y), \text{ say}.$$

In general $C_Y(x, y)$ and $\gamma_Y(x, y)$ (and therefore $C_Z(x, y)$ and $\gamma_Z(x, y)$) are inaccessible, but in some situations where appropriate models can be inferred, these covariance functions and semivariograms can be used for predictions of Z under appropriate unbiasedness conditions. In working with an *IRF - k* process Z , one may use its covariance function $C_Z(x, y)$ if it is known or can be modeled. In all other situations one has to resort to working with authorized linear combinations $Z_W(s) = \sum_{\alpha=0}^n w_\alpha Z(x_\alpha + s)$ and the generalized covariance of Z . In general, to determine the appropriateness of an *IRF - k* model in a given situation and the estimation of the generalized covariance are difficult problems. Some iterative methods for evaluation of the order k of the drift and estimation of the generalized covariance are given in Delfiner (1982) and Chiles and Delfiner (2012).

Coregionalization

In many situations, at any given location, more than a single characteristic of a phenomenon may be manifested. For example, a core-sample of mineralized rock could contain a set of elements of interest. In such cases, at each location the observations consist of a vector $\mathbf{z}(x) = (z_1(x), z_2(x), \dots, z_p(x))^T$. The vector $\mathbf{z}(x)$ is considered as the realization of a random vector $\mathbf{Z}(x)$. Consequently, the expected value $E[\mathbf{Z}(x)] = \boldsymbol{\mu}(x)$ is a p -dimensional vector. The dispersion of the random vector is described by its covariance matrix $\boldsymbol{\Sigma}(x)$ which is a $p \times p$ matrix whose ij^{th} element is given by $Cov(Z_i(x), Z_j(x)) = E[(Z_i(x) - \mu_i(x))(Z_j(x) - \mu_j(x))]$. The random process $\mathbf{Z} = \{\mathbf{Z}(x), x \in D\}$ is similarly defined as in the case where $p = 1$. In particular, second-order stationarity would imply:

- (i) $E[\mathbf{Z}(x)] = \boldsymbol{\mu} \quad \forall x$
- (ii) $Var[\mathbf{Z}(x)] = \boldsymbol{\Sigma} \quad \forall x$
- (iii) $Cov[\mathbf{Z}(x), \mathbf{Z}(x + \vec{h})] = \boldsymbol{\Sigma}(\vec{h}) \quad \forall x$

where, $\Sigma(h)$ is a $p \times p$ matrix whose ij^{th} element is given by $Cov(Z_i(x), Z_j(x+h)) = C_{ij}(h)$. The term auto-covariance is used for the function $C_{ii}(h)$ and the term cross-covariance for $C_{ij}(h)$ when $i \neq j$. It is important to observe here that the matrix $\Sigma(h)$ is not symmetric, that is, it is not necessary that $C_{ij}(h) = C_{ji}(h)$.

The semivariogram matrix $\Gamma(\vec{h})$ is defined here as $\frac{1}{2}E[(\mathbf{Z}(x) - \mathbf{Z}(x+\vec{h}))(\mathbf{Z}(x) - \mathbf{Z}(x+\vec{h}))^T]$. The ij^{th} entry in $\Gamma(\vec{h})$ is given by $\frac{1}{2}E[(Z_i(x) - Z_i(x+\vec{h}))(Z_j(x) - Z_j(x+\vec{h}))]$. The matrix $\Gamma(\vec{h})$ is symmetric and may exist even when $\Sigma(\vec{h})$ may not.

It is difficult to characterize the covariance matrix $\Sigma(\vec{h})$, although it can be done in terms of spectral distributions and spectral densities as a generalization of Bochner's Theorem (see, e.g., Chiles and Delfiner 2012, and related references therein). However, for practical situations, the following models are useful: (i) $\Sigma(\vec{h}) = \mathbf{B} C(\vec{h})$ where $C(\vec{h})$ is the covariance function of a univariate random function (i.e., a positive definite function) and $\mathbf{B}$ is a positive definite matrix, and (ii) If $C_1(\vec{h}), C_2(\vec{h}), \dots, C_k(\vec{h})$ are positive definite functions and $\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_k$ are positive definite matrices, then $\sum_{i=1}^k \mathbf{B}_i C_i(\vec{h})$ is a valid matrix covariance function (Wackernagel 2003). Similarly, valid matrix semivariogram functions $\Gamma(\vec{h}) = \mathbf{B} \gamma(\vec{h})$ and $\Gamma(\vec{h}) = \sum_{i=1}^k \mathbf{B}_i \gamma_i(\vec{h})$ can be obtained where $\gamma(\vec{h})$ and $\gamma_i(\vec{h})$ are valid semivariograms and $\mathbf{B}$ and $\mathbf{B}_i$ are positive definite matrices. When the covariance matrix $\Sigma(\vec{h})$ is modeled as in (i) above, it is postulated that random functions $Z_1, Z_2, \dots, Z_p$ that are elements of the random vector $\mathbf{Z}$ are fundamentally related to a single underlying generating process whose covariance and semivariogram are characterized by $C(\vec{h})$ and $\gamma(\vec{h})$, respectively, and $\mathbf{Z}$ is said to be *intrinsically coregionalized*. If on the other hand, $\Sigma(\vec{h})$ is modeled as in (ii) above, the random vector $\mathbf{Z}$ is considered as the weighted sum of k independent random vectors $\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_k$, with covariance matrices $\Gamma_1(\vec{h}), \Gamma_2(\vec{h}), \dots, \Gamma_k(\vec{h})$, respectively, wherein each random vector $\mathbf{W}_i$, $i = 1, 2, \dots, k$ is modeled as in (i) above. In such a case, the random vector $\mathbf{Z}$ is said to admit the *linear model of coregionalization (LMC)*.

In applications, it is often of interest to study functions $\mathbf{Y}(x) = g(\mathbf{Z}(x))$. In such situations, a new process $\mathbf{Y} = \{\mathbf{Y}(x), x \in D\}$ is considered. The matrix covariance and semivariogram functions are denoted by $\Sigma_Y(\vec{h})$ and $\Gamma_Y(\vec{h})$, respectively. Often the function g of interest is the indicator function $\mathbf{Y}(x) = \mathbf{I}_{\mathbf{Z}(x) \in A}(x)$ where A is a set in $\mathfrak{R}^p$. When $p = 1$, $Y(x) = I_{Z(x) \in A}(x)$ where A is some set in $\mathfrak{R}$ and correspondingly the covariance and semivariogram functions

of $Y(x)$ are termed *indicator covariance* and *indicator semivariogram*, respectively. In the general case the ij^{th} elements of $\Sigma_Y(\vec{h})$ and $\Gamma_Y(\vec{h})$ are given by $C_{Y_i Y_j}(\vec{h})$ and $\gamma_{Y_i Y_j}(\vec{h})$ which are the cross-indicator covariance and cross-indicator semivariogram, respectively.

Estimating Covariance and Variogram Functions

When using models of $\mathbf{Z}$ that are specified in terms of $\Sigma(h)$ or $\Gamma(h)$, these matrix functions are to be estimated from observed data. This involves the estimation of auto- and cross-covariance and semivariogram functions, that is, $C_{ij}(h)$ and $\gamma_{ij}(h)$, respectively.

Data Configurations

In applications of geostatistics, the p -dimensional random process $\mathbf{Z} = \{\mathbf{Z}(x), x \in D\}$, may be observed at locations, x_α , $\alpha = 1, 2, \dots, N$. In many applications, all p elements of the random vector $\mathbf{Z}(x)$ may not be observed at all locations. The most general data configuration is that at each x_α at least one of the components of the random vector $\mathbf{Z}(x)$ is observed. Data configurations can broadly be classified into *isotopic* and *heterotopic*. A data configuration is said to be isotopic if all the p -components of $\mathbf{Z}(x)$ are observed at all N data locations; otherwise, the data configuration is said to be heterotopic. In the case where $p = 1$, the data configuration would be isotopic. When the data configuration is heterotopic, it is said to be *completely heterotopic* (or *dislocated*) if no two components of the random vector $\mathbf{Z}(x)$ are observed at the same location. Heterotopic data in which more than one component of $\mathbf{Z}(x)$ is observed at some locations is said to be *partially heterotopic*.

In applications, it may be of interest to predict some of the components of $\mathbf{Z}(x)$ at location x_0 where these components are unobserved. In such situations, the component that is to be estimated is said to be the *primary variable* while the other components are said to be secondary or *auxiliary variables*. Auxiliary variables may or may not be observed at the location x_0 where the primary variable is to be predicted. When auxiliary variables are available at the location where prediction is required, the data configuration is said to be *collocated*. Further, if the auxiliary variables are observed only at the location x_0 where prediction is desired and nowhere else (or if auxiliary variables elsewhere except at x_0 are ignored for the purpose of prediction), the data configuration is said to be *strictly collocated*. The term *multicollocated* is used to describe the situation where auxiliary variables are observed at all locations where the primary variable is observed and additionally at the location x_0 where prediction of the primary variable is to be made.

Modeling Experimental Covariances and Semivariograms

The estimation of auto-covariances and semivariograms is considered first and for this purpose, without loss of generality, p is taken to be equal to 1. The observed data $z(x_1), z(x_2), \dots, z(x_N)$ are the realizations of $Z(x_1), Z(x_2), \dots, Z(x_N)$. Under second-order stationarity (or intrinsic stationarity) $E[Z(x)]$ is estimated by $\frac{1}{N} \sum_{i=1}^N z(x_i) = m$, say. Clearly all pairs $(z(x_i), z(x_j))$ with $(x_i - x_j) = \vec{h}$ can be used to estimate $C(\vec{h})$ and $\gamma(\vec{h})$. The estimators $\hat{C}(\vec{h})$ and $\hat{\gamma}(\vec{h})$ called the *experimental covariance* and *experimental semivariogram*, respectively, are given by $\hat{C}(\vec{h}) = \frac{1}{n} \left(\sum_{i=1}^n z(x_i) z(x_i + \vec{h}) - m^2 \right)$ and $\hat{\gamma}(\vec{h}) = \frac{1}{2n} \sum_{i=1}^n \left(z(x_i) - z(x_i + \vec{h}) \right)^2$, where n is the number of pairs available.

If γ or C depend on $\vec{h}$ only through its length $\|\vec{h}\|$, then Z is said to be *isotropic*, that is, $\gamma(\vec{h}) = \gamma(h)$ where $h = \|\vec{h}\|$. All pairs $(z(x_i), z(x_j))$ with $\|x_i - x_j\| = h$ can then be used to estimate $\gamma(h)$. In the isotropic case, the contour $\{\vec{h} : \gamma(\vec{h}) = k\}$ is a sphere. When this is not the case, Z is said to be *anisotropic*. Often anisotropy can be corrected by a transformation of coordinates. When this is the case, the random function Z is said to have *geometric anisotropy*. Correction of anisotropy by transformation of coordinates is possible when the contour $\{\vec{h} : \gamma(\vec{h}) = k\}$ can be approximated by an ellipse. In many applications $D \subset \mathbb{R}^3$ and in such cases, the contour $\{\vec{h} : \gamma(\vec{h}) = k\}$ is approximated by $\{\vec{h} : \frac{h_1^2}{a_1^2} + \frac{h_2^2}{a_2^2} + \frac{h_3^2}{a_3^2} = k\}$ where $\vec{h} = (h_1, h_2, h_3)^T$. By change of coordinates the vector $\vec{h} = (h_1, h_2, h_3)^T$ is transformed to $\vec{h}' = (h'_1, h'_2, h'_3)^T$ with $h'_1 = h_1, h'_2 = h_2 \cdot a_1/a_2, h'_3 = h_3 \cdot a_1/a_3$. A function $\tilde{\gamma}(\vec{h}') = \gamma(\vec{h}) \forall \vec{h}$ is defined such that the contour $\{\vec{h}' : \tilde{\gamma}(\vec{h}') = k'\}$ is given by $\{\vec{h}' : \frac{h'^2_1}{a_1^2} + \frac{h'^2_2}{a_1^2} + \frac{h'^2_3}{a_1^2} = k'\}$ that is, a sphere. To see this, consider the set $\{\vec{h}' : \tilde{\gamma}(\vec{h}') = k'\} = \{\vec{h}' : \gamma(\vec{h}) = k'\} = \{\vec{h}' : \frac{h'^2_1}{a_1^2} + \frac{h'^2_2}{a_2^2} + \frac{h'^2_3}{a_3^2} = c\}$ for an appropriate c because the contours of γ are ellipses. Multiplying by a_1^2 , one has $\{\vec{h}' : h'^2_1 + h'^2_2 a_1^2/a_2^2 + h'^2_3 a_1^2/a_3^2 = a_1^2 c\} = \{\vec{h}' : \frac{h'^2_1}{a_1^2} + \frac{h'^2_2}{a_1^2} + \frac{h'^2_3}{a_1^2} = c\}$ showing that the contours of $\tilde{\gamma}$ are spheres.

Another kind of anisotropy is called *zonal anisotropy* (Journel and Huijbregts 1978). Zonal anisotropy can occur when the random process Z is the sum of two or more independent random processes $Z_1, Z_2, \dots$, each with different geometric anisotropies. The semivariogram (or covariance) of Z is then the sum of the semivariograms (or covariances), respectively, of each of the independent random processes, that is, $\gamma(\vec{h}) = \gamma_1(\vec{h}) + \gamma_2(\vec{h}) + \dots$ where each $\gamma_i(\vec{h})$ may

have its own geometric anisotropy. In particular, one or more components of the random process Z may have highly extended ellipses of geometric anisotropy in one or more directions.

In data analysis, the experimental semivariograms $\hat{\gamma}(\vec{h})$ of Z are computed in several directions and the principal directions of variability are identified and can often be related to anisotropies in the underlying physical processes. When the random function Z is transitive, geometric anisotropy is indicated when the range of the variogram (or covariance) varies with direction. If geometric anisotropy is present, then the principal directions of variability are taken as the principal axes of the ellipse of geometric anisotropy. In the case of zonal anisotropy, the variogram of Z is modeled as the sum of more than one variograms, each with its own geometric anisotropy. *Semivariogram modeling* consists of fitting valid (conditional negative definite) functions to $\hat{\gamma}(\vec{h})$. Commonly parametric models such as the spherical, exponential, Gaussian, and nugget models are used for random functions that show transitive phenomena (Journel and Huijbregts 1978). Other valid models can also be used according to the nature of the random function Z . When the random function Z is second-order stationary, models of the covariance are immediately obtained from models of the semivariogram since $C(\vec{h}) = C(0) - \gamma(\vec{h})$.

When a random vector $\mathbf{Z}$ is studied, modeling of its semivariogram $\Gamma(\vec{h})$ and covariance $\Sigma(\vec{h})$ is done by obtaining the ij^{th} experimental cross-semivariograms and cross-covariances, $\hat{\gamma}_{ij}(\vec{h})$ and $\hat{C}_{ij}(\vec{h})$, respectively, where $\hat{\gamma}_{ij}(\vec{h}) = \frac{1}{2n} \sum_{\alpha=1}^n \left(z_i(x_\alpha) - z_i(x_\alpha + \vec{h}) \right) \left(z_j(x_\alpha) - z_j(x_\alpha + \vec{h}) \right)$ and $\hat{C}_{ij}(\vec{h}) = \frac{1}{n} \sum_{\alpha=1}^n z_i(x_\alpha) z_j(x_\alpha + \vec{h}) - m_i m_j$ with n being the number of available pairs, $m_i = \frac{1}{N} \sum_{\alpha=1}^N z_i(x_\alpha)$, and $m_j = \frac{1}{N} \sum_{\alpha=1}^N z_j(x_\alpha)$. The intrinsic model of coregionalization or the linear model of coregionalization may be used to guarantee consistency when such models are appropriate.

Kriging

One of the most important problems of geostatistics, in its simplest form, is the following: A stochastic process Z is observed at locations $x_\alpha, \alpha = 1, 2, \dots, N$, that is, realizations $z(x_1), z(x_2), \dots, z(x_N)$ of the random variables $Z(x_1), Z(x_2), \dots, Z(x_N)$ are available. The random variable $Z(x_0)$ corresponding to the location x_0 has not been observed and is to be predicted (or estimated) using the data at $x_1, x_2, \dots, x_N$.

Towards a solution, assume all random variables have finite variance and consider the class ζ of all random variables of the form $g(Z(x_1), Z(x_2), \dots, Z(x_N))$ with finite variance.

A reasonable predictor of $Z(x_0)$ is a member $Z^*(x_0)$ of the class ζ which minimizes $E[(Z(x_0) - Z^*(x_0))^2]$, the least squares estimator (Journal and Huijbregts 1978). The least squares estimator $Z^*(x_0)$ that is obtained turns out to be the conditional expectation $E[Z(x_0) | Z(x_1), Z(x_2), \dots, Z(x_N)]$. Evaluation of this conditional expectation, however, needs the $N + 1$ - variate joint distribution of $Z(x_0), Z(x_1), Z(x_2), \dots, Z(x_N)$. Unfortunately, most applications in geostatistics do not lend themselves to such modeling. This indicates that the class ζ is "too big." Choosing a smaller class leads to implementable solutions.

Linear Kriging

The simplest linear kriging solution considers the class ζ_0 of all functions of the type $\lambda_0 + \sum_{\alpha=1}^N \lambda_\alpha Z(x_\alpha)$ where, $\lambda_0, \lambda_1, \dots, \lambda_N$ are real numbers. The object is to choose the $\lambda_0, \lambda_1, \dots, \lambda_N$ so as to minimize the quantity $E[(Z(x_0) - Z^*(x_0))^2] =$

$E\left[Z(x_0) - \left(\lambda_0 + \sum_{\alpha=1}^N \lambda_\alpha Z(x_\alpha)\right)\right]^2$, termed the *estimation variance*. By differentiating the expression for variance of estimation with respect to λ_α , $\alpha = 1, 2, \dots, N$ and λ_0 the following equations are obtained for the optimal choice of λ_0 ,

$$\lambda_1, \dots, \lambda_N: E\left[\left(Z(x_0) - \left(\lambda_0 + \sum_{\beta=1}^N \lambda_\beta Z(x_\beta)\right)\right)Z(x_\alpha)\right] = 0, \\ \alpha = 1, 2, \dots, N, \text{ and } E\left[Z(x_0) - \left(\lambda_0 + \sum_{\alpha=1}^N \lambda_\alpha Z(x_\alpha)\right)\right] = 0.$$

These equations can be written as $\sum_{\beta=1}^N \lambda_\beta C(x_\beta, x_\alpha) = C(x_0, x_\alpha)$, $\alpha = 1, 2, \dots, N$ and $\lambda_0 = m(x_0) - \sum_{\alpha=1}^N \lambda_\alpha m(x_\alpha)$

where $C(x_\beta, x_\alpha) = \text{Cov}[Z(x_\beta), Z(x_\alpha)]$, $C(x_0, x_\alpha) = \text{Cov}[Z(x_0), Z(x_\alpha)]$, $m(x_\alpha) = E[Z(x_\alpha)]$ and $m(x_0) = E[Z(x_0)]$. Equivalently, using the relation $\gamma(\vec{h}) = C(0) - C(\vec{h})$, the preceding equations can be written in terms of the semi-

variogram as $\sum_{\beta=1}^N \lambda_\beta \gamma(x_\beta, x_\alpha) = \gamma(x_0, x_\alpha)$, $\alpha = 1, 2, \dots, N$ and $\lambda_0 = m(x_0) - \sum_{\alpha=1}^N \lambda_\alpha m(x_\alpha)$. The variance of estimation for the optimal solution, that is, $\text{Var}[Z(x_0) - Z^*(x_0)] = E[(Z(x_0) - Z^*(x_0))^2]$ is obtained as $C(0) - \sum_{\alpha=1}^N \lambda_\alpha C(x_0, x_\alpha)$ or as

$\sum_{\alpha=1}^N \lambda_\alpha \gamma(x_0, x_\alpha)$, (noting that $\gamma(0) = 0$). The variance of the optimal estimator, that is, $\text{Var}[Z^*(x_0)]$ is given by $\sum_{\alpha=1}^N \sum_{\beta=1}^N \lambda_\alpha \lambda_\beta C(x_\alpha, x_\beta) = \sum_{\alpha=1}^N \lambda_\alpha C(x_0, x_\alpha)$.

Geometrically, the optimal solution $Z^*(x_0)$ is the projection of $Z(x_0)$ on to ζ_0 and the preceding equations are precisely

the ones satisfied by the projection, that is, the vector $Z(x_0) - Z^*(x_0)$ is perpendicular to each member of ζ_0 . Geometrically, the variance of estimation is the norm of the vector $(Z(x_0) - Z^*(x_0))^2$, that is, $\|Z(x_0) - Z^*(x_0)\|^2$. The process of projecting $Z(x_0)$ onto ζ_0 is called *simple kriging* (Journal and Huijbregts 1978) and this optimal estimator is commonly denoted by $Z_{SK}^*(x_0)$. The variance of estimation is denoted by $\sigma_{SK}^2(x_0)$. A property of $Z_{SK}^*(x_0)$ is that it is unbiased, that is, $E[Z_{SK}^*(x_0) - Z(x_0)] = 0$ or, $E[Z_{SK}^*(x_0)] = E[Z(x_0)] = m(x_0)$. Unbiasedness is guaranteed because $E[Z_{SK}^*(x_0) - Z(x_0)] = 0$ gives $\lambda_0 = m(x_0) - \sum_{\alpha=1}^N m(x_\alpha)$, one of the equations of simple kriging.

If it is assumed that the stochastic process Z is second-order stationary, then $C(x_\alpha, x_\beta) = C(x_\alpha - x_\beta) \quad \forall \alpha, \beta$ and $\gamma(x_\alpha, x_\beta) = \gamma(x_\alpha - x_\beta) \quad \forall \alpha, \beta$ and the values of these quantities can be obtained from the covariance and semivariogram models of Z . Also, in this case, $m(x) = m \quad \forall x$ giving $\lambda_0 = \left(1 - \sum_{\alpha=1}^N \lambda_\alpha\right)m$.

When $m(x)$ is not constant, but is known everywhere, $\lambda_0 = m(x_0) - \sum_{\alpha=1}^N \lambda_\alpha m(x_\alpha)$ and it can be seen that the equations for obtaining the optimal estimator remain the same. This is expected since when $m(x)$ is known everywhere, by considering $Y(x) = Z(x) - m(x)$, the estimation procedure reduces to one where m is constant and known to be zero.

When m is unknown, but assumed to be constant, the subspace of $\zeta_1 \subset \zeta_0$ of all random variables of the type $\sum_{\alpha=1}^N \lambda_\alpha Z(x_\alpha)$, is considered, that is, λ_0 is taken to be zero.

This imposes the condition $\sum_{\alpha=1}^N \lambda_\alpha = 1$ and hence the projection of $Z(x_0)$ is done onto a linear manifold $\xi_1 \subset \zeta_1$. Since $\sum_{\alpha=1}^N \lambda_\alpha = 1$, $Z^*(x_0)$ is unbiased, that is, $E\left[\sum_{\alpha=1}^N \lambda_\alpha Z(x_\alpha) - Z(x_0)\right] =$

$\left(\sum_{\alpha=1}^N \lambda_\alpha - 1\right)m = 0$, irrespective of m . Thus, it is not required to estimate m .

The process of projecting $Z(x_0)$ onto ξ_1 is called *ordinary kriging* and the optimal estimator so obtained is called the ordinary kriging estimator, $Z_{OK}^*(x_0)$, with a corresponding variance of estimation $\|(Z(x_0) - Z^*(x_0))^2\| = \sigma_{OK}^2(x_0)$. Minimization of $E[(Z(x_0) - Z^*(x_0))^2]$ is done under the condition that $\sum_{\alpha=1}^N \lambda_\alpha = 1$. This is done using the method of Lagrange multipliers giving $\sum_{\beta=1}^N \lambda_\beta C(x_\beta, x_\alpha) - \mu = C(x_0, x_\alpha)$, $\alpha = 1, 2, \dots, N$ and $\sum_{\alpha=1}^N \lambda_\alpha = 1$, where μ is the Lagrange multiplier. The variance of ordinary kriging,

$\sigma_{OK}^2(x_0) = C(0) - \sum_{\alpha=1}^N \lambda_{\alpha} C(x_0, x_{\alpha}) + \mu$. Equivalently, in terms of the semivariogram the $N + 1$ equations of ordinary kriging are given by $\sum_{\beta=1}^N \lambda_{\beta} \gamma(x_{\beta}, x_{\alpha}) + \mu = \gamma(x_0, x_{\alpha})$, $\alpha = 1, 2, \dots, N$ and $\sum_{\alpha=1}^N \lambda_{\alpha} = 1$, with $\sigma_{OK}^2(x_0) = \sum_{\alpha=1}^N \lambda_{\alpha} \gamma(x_0, x_{\alpha}) + \mu$ (noting again that $\gamma(0) = 0$). When Z is intrinsic, the ordinary kriging equations in terms of the semivariogram must be used.

It can be shown that the ordinary kriging estimator is the simple kriging estimator with the unknown mean m replaced by m_{OK}^* , where $m_{OK}^* = \sum_{\alpha=1}^N \lambda_{\alpha}^m Z(x_{\alpha})$ is the ordinary kriging estimator of the mean $M = \frac{1}{N} \sum_{\alpha=1}^N Z(x_{\alpha})$. In other words, $Z_{OK}^*(x_0) = m_{OK}^* + \sum_{\alpha=1}^N \lambda_{\alpha} (Z(x_{\alpha}) - m_{OK}^*)$ where λ_{α} , $\alpha = 1, 2, \dots, N$ are solutions to the simple kriging system with $\lambda_0 = \left(1 - \sum_{\alpha=1}^N \lambda_{\alpha}\right) m_{OK}^*$.

In many situations, $m(x)$ varies with x and is unknown. If it is assumed that $m(x) = \sum_{l=0}^L a^l f^l(x)$, for some basis functions $f^0, f^1, \dots, f^L$, where the a^l 's are unknown, then it is possible to obtain an unbiased optimal predictor of $Z(x_0)$ without estimating the a^l 's. One choice for the basis functions $f^0, f^1, \dots, f^L$ is a set of monomials in x with $f^0(x) = 1$. The required estimator of the type $Z^*(x_0) = \sum_{\alpha=1}^N w_{\alpha} Z(x_{\alpha})$ is one where the w_{α} 's satisfy the unbiasedness of the estimator $Z^*(x_0)$. Since $E[Z^*(x_0) - Z(x_0)] = \sum_{\alpha=1}^N w_{\alpha} \sum_{l=0}^L f^l(x_{\alpha}) - \sum_{l=0}^L f^l(x_0) = 0$ this gives the required unbiasedness conditions $\sum_{l=0}^L w_{\alpha} f^l(x_{\alpha}) = f^l(x_0)$, $l = 0, 1, \dots, L$. Geometrically, since $m(x)$ is unknown, and $L + 1$ unbiasedness conditions are introduced, a subspace $\xi_L \subset C_1$ with L linear constraints is considered to project $Z(x_0)$. This projection is called *universal kriging*. The optimal estimator $Z^*(x_0)$, termed the universal kriging estimator is a member of ξ_L and is given by $\sum_{\beta=1}^N \lambda_{\beta} C(x_{\beta}, x_{\alpha}) - \sum_{l=0}^L \mu_l f^l(x_{\alpha})$, $\alpha = 1, 2, \dots, N$, and $\sum_{\alpha=1}^N \lambda_{\alpha} f^l(x_{\alpha}) = f^l(x_0)$, $l = 0, 1, \dots, L$ with variance of estimation given by $\sigma_{UK}^2 = C(0) - \sum_{\alpha=1}^N \lambda_{\alpha} C(x_0, x_{\alpha}) + \sum_{l=0}^L \mu_l f^l(x_0)$. The μ_l 's in the universal kriging equations are Lagrange multipliers. The equations of universal kriging can similarly also be obtained in terms of the semivariogram as in previous cases. As discussed earlier, $C(x_{\alpha}, x_{\beta})$ and $\gamma(x_{\alpha}, x_{\beta})$ cannot be estimated directly from data since $m(x)$ is unknown. Universal kriging is

possible only when approximate models of the covariance or semivariogram of the random function can be constructed.

Another approach is to model Z as an intrinsic random function of order k ($IRF - k$) with $m(x) = \sum_{l=0}^L a^l f^l(x)$ and use authorized models which give a second-order stationary random function $Z_W(s) = \sum_{\alpha=0}^L w_{\alpha} Z(x_{\alpha} + s) \quad \forall s$ with w_{α} , $\alpha = 1, 2, \dots, n$ satisfying $\sum_{\alpha=1}^n w_{\alpha} f^l(x_{\alpha} + s) = f^l(x_0 + s) \quad \forall s$, $l = 0, 1, \dots, L$. The random function Z has generalized covariance $K(\vec{h})$. The set of optimal equations obtained is identical to that of universal kriging except that the generalized covariance $K(x_{\alpha}, x_{\beta}) = K(x_{\alpha} - x_{\beta})$ is used instead of $C(x_{\alpha}, x_{\beta})$. The μ_l 's in the $IRF - k$ kriging equations are Lagrange multipliers that ensure that the λ 's lead to an authorized linear combination. As mentioned earlier, in the case of $IRF - 0$, $K(\vec{h}) = -\gamma(\vec{h})$. Geometrically, the $IRF - k$ estimator $Z_{IRF-k}^*(x_0)$ is the projection of $Z(x_0)$ onto ξ_L .

Cokriging

Let the random process $\mathbf{Z} = \{\mathbf{Z}(x), x \in D\}$ be observed at x_{α} , $\alpha = 1, 2, \dots, N$ and let $\mathbf{Z}(x_{\alpha}) = (Z_1(x_{\alpha}), Z_2(x_{\alpha}), \dots, Z_p(x_{\alpha}))^T$ be the p -dimensional vector observed at x_{α} . The random vector $\mathbf{Z}(x_0) = (Z_1(x_0), Z_2(x_0), \dots, Z_p(x_0))^T$ is unobserved. For brevity of notation, denote by Z_j^z the j^{th} component of $\mathbf{Z}(x_{\alpha})$, $\alpha = 0, 1, 2, \dots, N$ and let $Z_i^{0*} = \lambda_0^i + \sum_{\alpha=1}^N \sum_{j=1}^p \lambda_{j\alpha}^i Z_j^z$ be the estimator of Z_i^0 where $\lambda_{j\alpha}^i$ are coefficients of Z_j^z , $j = 1, 2, \dots, p$, and $\alpha = 1, 2, \dots, N$ in the estimation of Z_i^0 , and λ_0^i is a constant. Since the unbiasedness condition for the estimator of Z_i^0 is given by $E[Z_i^0 - Z_i^{0*}] = E\left[Z_i^0 - \lambda_0^i - \sum_{\alpha=1}^N \sum_{j=1}^p \lambda_{j\alpha}^i Z_j^z\right] = 0$, $\lambda_0^i = m_i^0 - \sum_{\alpha=1}^N \sum_{j=1}^p \lambda_{j\alpha}^i m_j^z$ where $m_i^0 = E[Z_i^0]$ and $m_j^z = E[Z_j^z]$, $j = 1, 2, \dots, p$; $\alpha = 1, 2, \dots, N$. The optimal values of the $\lambda_{j\alpha}^i$ that minimize $E[Z_i^0 - Z_i^{0*}]^2 = E\left[Z_i^0 - \lambda_0^i - \sum_{\alpha=1}^N \sum_{j=1}^p \lambda_{j\alpha}^i Z_j^z\right]^2$ are then given by $\sum_{\beta=1}^N \sum_{k=1}^p \lambda_{k\beta}^i \text{Cov}(Z_k^z, Z_j^z) = \text{Cov}(Z_i^0, Z_j^z)$, $j = 1, 2, \dots, p$; $\alpha = 1, 2, \dots, N$. The optimal estimator Z_i^{0*} so obtained is called the *simple cokriging* estimator. The variance of estimation of this estimator is given by $\sigma_{COK,i}^2 = E[Z_i^0 - Z_i^{0*}]^2 = C_i(0) - \sum_{\alpha=1}^N \sum_{j=1}^p \lambda_{j\alpha}^i \text{Cov}(Z_i^0, Z_j^z)$. If $\mathbf{Z}$ is a second-order stationary random process, $m_j = E[Z_j^z]$, $j = 1, 2, \dots, p$; $\alpha = 0, 1, \dots, N$, $\text{Cov}(Z_k^z, Z_j^z) = C_{kj}(x_{\beta} - x_{\alpha})$ and $\text{Cov}(Z_i^0, Z_j^z) = C_{ij}(x_0 - x_{\alpha})$.

When $m_j = E[Z_j^z]$, $j = 1, \dots, p$; $\alpha = 0, 1, \dots, N$ are unknown, but constant for each component of $\mathbf{Z}$, as in the ordinary kriging case, the unbiasedness condition for the estimator Z_i^{0*} of Z_i^0 is obtained as $E[Z_i^0 - Z_i^{0*}] = E[Z_i^0 - \lambda_0^i - \sum_{\alpha=1}^N \sum_{j=1}^p \lambda_{j\alpha}^i Z_j^z] = \left(m_i - \lambda_0^i - \sum_{\alpha=1}^N \sum_{j=1}^p \lambda_{j\alpha}^i m_j \right) = \left(m_i - \lambda_0^i - \sum_{\alpha=1}^N \lambda_{i\alpha}^i m_i - \sum_{j \neq i} \sum_{\alpha=1}^N \lambda_{j\alpha}^i m_j \right) = 0$. The preceding

equation is satisfied for any choice of m_i, m_j only when $\lambda_0^i = 0$, with $\sum_{\alpha=1}^N \lambda_{i\alpha}^i = 1$ and $\sum_{\alpha=1}^N \lambda_{j\alpha}^i = 0, \forall j \neq i$. The optimal equations for minimization of $E[Z_i^0 - Z_i^{0*}]^2$ subject to the preceding constraints are obtained as

$$\left\{ \begin{array}{l} \sum_{\beta=1}^N \sum_{k=1}^p \lambda_{k\beta}^i \text{Cov}(Z_k^\beta, Z_j^z) - \mu_k = \text{Cov}(Z_i^0, Z_j^z), \quad j = 1, 2, \dots, p; \quad \alpha = 1, 2, \dots, N \\ \sum_{\alpha=1}^N \lambda_{k\alpha}^i = 0, \quad k \neq i \\ \sum_{\alpha=1}^N \lambda_{i\alpha}^i = 1 \end{array} \right.$$

where $\mu_k, k = 1, 2, \dots, p$ are Lagrange multipliers.

The estimator Z_i^{0*} obtained using the optimal weights obtained from the preceding equations is called the *ordinary cokriging estimator*.

If all variables are not available at every location, one may have $n_j \leq N$ observations on the variable j . Let $\alpha_j = 1, 2, \dots, n_j$ denote the locations where the process Z_j has been observed. The above ordinary cokriging equations are then given by

$$\left\{ \begin{array}{l} \sum_{k=1}^p \sum_{\beta_k=1}^{n_k} \lambda_{k\beta_k}^i \text{Cov}(Z_k^{\beta_k}, Z_j^{\alpha_j}) - \mu_k = \text{Cov}(Z_i^0, Z_j^{\alpha_j}), \quad j = 1, 2, \dots, p; \quad \alpha_j = 1, 2, \dots, n_j \\ \sum_{\alpha_k=1}^{n_k} \lambda_{k\alpha_k}^i = 0, \quad k \neq i \\ \sum_{\alpha_i=1}^{n_i} \lambda_{i\alpha_i}^i = 1 \end{array} \right.$$

These equations also result if in the previous set of equations $\lambda_{j\alpha}^i$ is taken to be zero if Z_j is not observed at location x_α .

When the random process $\mathbf{Z}$ is second-order stationary, the cokriging equations can be written as

$$\text{Cov}(Z_k^\beta, Z_j^\alpha) = C_{kj}(x_\alpha - x_\beta), \quad \forall k, j = 0, 1, \dots, p; \quad \alpha, \beta = 0, 1, \dots, N \text{ with} \\ \text{Var}(Z_j^z) = C_j(0), \quad j = 1, 2, \dots, p; \quad \alpha = 0, 1, \dots, N$$

The variance of the ordinary cokriging estimator is given by $\sigma_{\text{COK},i}^2 = E[Z_i^0 - Z_i^{0*}]^2 = C_i(0) - \sum_{\alpha=1}^N \sum_{j=1}^p \lambda_{j\alpha}^i C_{ij}(x_0 - x_\alpha) + \mu_i$, or $\sigma_{\text{COK},i}^2 = E[Z_i^0 - Z_i^{0*}]^2 = C_i(0) - \sum_{j=1}^p \sum_{\alpha_k=1}^{n_j} \lambda_{j\alpha_k}^i C_{ij}(x_0 - x_{\alpha_k}) + \mu_i$, depending on whether the data is isotopic or heterotopic.

In matrix form, considering simple cokriging in the second-order stationary isotopic case where all p variables are observed at all data locations, let $\mathbf{\Lambda}$ be a $(Np \times p)$ dimensional matrix of $N \times p$ weights for the simple cokriging of the vector $(Z_1^0, Z_2^0, \dots, Z_p^0)^T$. Let $\mathbf{Z}_D$ be the $(Np \times 1)$ vector of observations ordered by locations with p random variables at each of the N locations (i.e., the first p elements of $\mathbf{Z}_D$ are the p random variables observed at location x_1 , the next set being the p random variables observed at location x_2 , and so on). The $(Np \times Np)$ covariance matrix of $\mathbf{Z}_D$ is given by $\mathbf{\Sigma}_D$, say, where $\mathbf{\Sigma}_D = [C_{jk}(\alpha, \beta)]$, with $C_{jk}(\alpha, \beta) = \text{Cov}(Z_j^\alpha, Z_k^\beta), j, k = 1, 2, \dots, p; \quad \alpha, \beta = 1, 2, \dots, N$. If $\mathbf{Z}^0$ denotes the $(p \times 1)$ vector at location x_0 that is to be predicted, the simple cokriging vector $\mathbf{Z}^{0*}$ is given by $\mathbf{Z}^{0*} = \mathbf{\Lambda}^T \mathbf{Z}_D$, where $\mathbf{\Lambda} = \mathbf{\Sigma}_D^{-1} \mathbf{C}_0$ with $\mathbf{C}_0$ being the $(Np \times 1)$ vector whose elements

are given by $C_{ij}(0, \alpha) = \text{Cov}(Z_i^0, Z_j^\alpha)$, $\alpha = 1, 2, \dots, N$; $j = 1, 2, \dots, p$. When data locations and locations where predictions are to be made are grouped suitably into "blocks," block-matrix operations can be used effectively to obtain cokriged estimated at unobserved data locations.

An alternate way of ordering, the $(Np \times p)$ vector $\mathbf{Z}_D$ of observations is to do it random variable-wise with the first N elements of $\mathbf{Z}_D$ being $(Z_1^1, Z_1^2, Z_1^3, \dots, Z_1^N)^T$ followed by $(Z_2^1, Z_2^2, Z_2^3, \dots, Z_2^N)^T$ and so on. The $(Np \times Np)$ covariance matrix of $\mathbf{Z}_D$ is given by Σ_{D^*} , say, where $\Sigma_{D^*} =$

$$[C_{jk}(\alpha, \beta)] = \begin{bmatrix} C_{11}(\alpha, \beta) & C_{12}(\alpha, \beta) & \dots & C_{1p}(\alpha, \beta) \\ C_{21}(\alpha, \beta) & C_{22}(\alpha, \beta) & \dots & C_{2p}(\alpha, \beta) \\ \vdots & \vdots & \ddots & \vdots \\ C_{p1}(\alpha, \beta) & C_{p2}(\alpha, \beta) & \dots & C_{pp}(\alpha, \beta) \end{bmatrix}.$$

The matrices $C_{jk}(\alpha, \beta)$ give the covariance between vectors $(Z_j^1, Z_j^2, Z_j^3, \dots, Z_j^N)^T$ and $(Z_k^1, Z_k^2, Z_k^3, \dots, Z_k^N)^T$. The simple cokriging equations remain the same and the estimator is given by $\mathbf{Z}^{0*} = \Lambda^T \mathbf{Z}_D$ where $\Lambda = \Sigma_{D^*}^{-1} \mathbf{C}_0$ with $\mathbf{C}_0$ being the $(Np \times 1)$ vector whose elements are given by $C_{ij}(0, \alpha) = \text{Cov}(Z_i^0, Z_j^\alpha)$, $j = 1, 2, \dots, p$; $\alpha = 1, 2, \dots, N$. The p -dimensional vector of simple cokriging variances is given by $\sigma_{\text{COK}}^2 = \mathbf{C}(0) - \text{diag}[\Lambda^T \mathbf{C}_0]$ where $\mathbf{C}(0)$ is a p -dimensional vector of variances of Z_i^0 , $i = 1, 2, \dots, p$.

Kriging of $Z_V(x_0)$ and Functions of $Z(x_0)$

If it is required to predict the random function W at x_0 (i.e., $W(x_0)$ denoted by W^0) using observations on the random function Z at x_α , $\alpha = 1, 2, \dots, N$, denoted here as Z^α , $\alpha = 1, 2, \dots, N$, it is clear from the simple and ordinary kriging equations that minimization of $E[W^0 - W^{0*}]^2$ where $W^{0*} = \lambda_0 + \sum_{\alpha=1}^N \lambda_\alpha Z^\alpha$ can only be done when $\text{Cov}(W^0, Z^\alpha)$, $\alpha = 1, 2, \dots, N$ is known. As an example, let $W^0 = Z_V^0$ with $Z_V^0 \equiv Z_V(x_0) = \frac{1}{|V|} \int_V Z(s) ds$ where V is a volume centered at x_0 . Here, $\text{Cov}(Z_V^0, Z^\alpha) = \frac{1}{|V|} \int_V \text{Cov}(x_s, x_\alpha) ds$ (or if $Z_V^0 = \frac{1}{N} \sum_{i=1}^N Z^{s_i}$, $\text{Cov}(Z_V^0, Z^\alpha) = \frac{1}{N} \sum_{i=1}^N \text{Cov}(x_{s_i}, x_\alpha)$) where x_{s_i} are chosen points located in the volume V centered at x_0 . When Z is a second-order stationary random function, and $C(\vec{h})$ or $\gamma(\vec{h})$ have been modeled, $\text{Cov}(Z_V^0, Z^\alpha) = \frac{1}{|V|} \int_V C(x_s - x_\alpha) ds$, (or $\text{Cov}(Z_V^0, Z^\alpha) = \frac{1}{N} \sum_{i=1}^N C(x_{s_i} - x_\alpha)$). The solution to the simple or ordinary kriging system is obtained as before. The variance of estimation by simple and ordinary kriging are given by $\sigma_{\text{SK}}^2 = \bar{C}(V, V) - \sum_{\alpha=1}^N \lambda_\alpha \bar{C}(V, \alpha)$, or $\sigma_{\text{OK}}^2 = \bar{C}(V, V) - \sum_{\alpha=1}^N \lambda_\alpha \bar{C}(V, \alpha) + \mu$, respectively, where $\bar{C}(V, V) = \text{Var}(Z_V(x_0))$

$$= \frac{1}{|V|^2} \int_V \int_V \text{Cov}(x_s, x_{s'}) ds ds' \quad (\text{or when } Z_V^0 = \frac{1}{N} \sum_{i=1}^N Z^{s_i}, \bar{C}(V, V) = \text{Var}(Z_V(x_0)) = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \text{Cov}(x_{s_i}, x_{s_j})).$$

When Z is a second-order stationary random function, these can be written as $\bar{C}(V, V) = \frac{1}{|V|^2} \int_V \int_V C(x_s - x_{s'}) ds ds'$ or $\bar{C}(V, V) = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N C(x_{s_i} - x_{s_j})$.

If a function $W(x_0) = g(Z(x_0))$ is of interest, the estimation of $W(x_0)$ by $W^{0*} = \lambda_0 + \sum_{\alpha=1}^N \lambda_\alpha W^\alpha$, where $W^\alpha = g(Z(x_\alpha))$, is considered and the optimal solution is obtained by minimizing $E[W^0 - W^{0*}]^2$. The required covariances, $\text{Cov}(W(x_\alpha), W(x_\beta))$, $\alpha = 0, 1, \dots, N$; $\beta = 0, 1, \dots, N$, are obtained by modeling the random function $W = g(Z)$. If second-order stationarity of W is assumed, the covariance of W can be modeled and the kriging of $W(x_0) = g(Z(x_0))$ reduces to the simple and ordinary kriging cases discussed earlier.

A function g that is often used in geostatistics is the indicator function $I_{Z(x) \leq z}(x) = \begin{cases} 1 & \text{if } Z(x) \leq z \\ 0 & \text{otherwise} \end{cases}$. To estimate $W = \{W(x) \equiv I_{Z(x) \leq z}(x), x \in D\}$ at unobserved locations, its covariance or semivariogram (i.e., the indicator-covariance or indicator-semivariogram) are required. Depending on the stationarity assumptions made on the random function W , simple or ordinary kriging is used to obtain the estimator and the corresponding procedures are called *simple kriging of indicators* or *ordinary kriging of indicator* respectively.

If the p -dimensional indicator random process $\mathbf{W} = \{\mathbf{W}(x) \equiv (I_{Z(x) \leq z_i}(x), i = 1, 2, \dots, p)^T, x \in D\}$ is considered, one option to estimate the random vector $\mathbf{W}(x_0) = (I_{Z(x_0) \leq z_i}(x_0), i = 1, 2, \dots, p)^T$ at location x_0 using sets of indicator random variables $\mathbf{W}(x_\alpha) = (I_{Z(x_\alpha) \leq z_i}(x_\alpha), i = 1, 2, \dots, p)^T$; $\alpha = 1, 2, \dots, N$ at locations x_α where Z has been observed, is to use simple or ordinary cokriging depending on stationarity assumptions on $\mathbf{W}$. To obtain the optimal estimator $\mathbf{W}^*(x_0) = (I_{Z(x_0) \leq z_i}^*(x_0), i = 1, 2, \dots, p)^T$ simple or ordinary cokriging are used. The terms *simple cokriging of indicators* and *ordinary cokriging of indicators*, respectively, are used for these procedures. To perform the cokriging, auto- and cross-indicator-covariances or indicator-semivariograms need to be modeled.

It is relevant to note here that $E[I_{Z(x) \leq z}(x)] = F_{Z(x)}(z)$ where $F_{Z(x)}(z)$ is the frequency distribution function of $Z(x)$ evaluated at z and that $E[I_{Z(x) \leq z}(x) I_{Z(x+\vec{h}) \leq z'}(x+\vec{h})] =$

$F^{\vec{h}}(z, z')$, say is the joint distribution of the random variables $(Z(x), Z(x + \vec{h}))$ evaluated at z and z' . The cokriging of indicators gives an estimate $F_{Z(x)}^*(z) | N$ for $E[I_{Z(x) \leq z}(x) | Z(x_\alpha) = z_\alpha, \alpha = 1, 2, \dots, N] = F_{Z(x)}(z) | N$, say, for a discrete set of values of z , where $F_{Z(x)}(z) | N$ is the conditional distribution of $Z(x) | Z(x_\alpha) = z_\alpha, \alpha = 1, 2, \dots, N$. For details, one may see Journel (1983, 1984).

Non-Parametric Estimation: Multiple Indicator Kriging

Let $z(x_\alpha), \alpha = 1, 2, \dots, N$ be observations on Z at $x_1, x_2, \dots, x_N$.

Consider the function $I_{Z(x) \leq z}(x) = \begin{cases} 1 & \text{if } Z(x) \leq z \\ 0 & \text{otherwise} \end{cases} = I(x; z)$,

say, and let $I^*(x; z | N)$ be the (ordinary indicator kriging) estimate of $I(x; z)$ at locations x where Z has not been observed.

Define a partition of the range of Z with appropriately chosen nodes $z_1 < z_2 < \dots < z_K$ and define $F^*(x; z | N) =$

$$\begin{cases} 0, & z < z_1 \\ I^*(x; z_i | N), & z_i \leq z < z_{i+1}, i < K \\ 1, & z \geq z_K \end{cases}$$

is an estimate of the conditional distribution function at x given the N data. An estimate $\phi^*(Z(x))$ of $\phi(Z(x))$, a function of $Z(x)$, is obtained as $E^*[\phi(Z(x))]$ where E^* denotes expectation computed under $F^*(x; z | N)$, that is,

$$E^*[\phi(Z(x))] = \sum_{i=1}^K \phi(z'_i) (F^*(x, z_{i+1} | N) - F^*(x, z_i | N)), \text{ where,}$$

z'_i is a chosen value in the interval $[z_i, z_{i+1}]$. This estimation procedure is called *multiple indicator kriging (MIK)*.

The MIK procedure is widely practiced because of ease of use and its ability to handle outliers. Modifications of the procedure have been used to incorporate information from “soft” categorical data. In MIK, estimates $I^*(x; z | N)$ are obtained by ordinary kriging (and not by cokriging) thereby losing some information contained in the cross-indicator covariances. Post-kriging correction of $I^*(-x; z | N)$ may be required to ensure monotonicity of $F^*(-x; z | N)$. This may occur at locations where the indicator variograms or covariances may not accurately capture local spatial structure of the regionalization. One other problem encountered in MIK is the problem of de-structuring of the variograms at high cut-offs due to sparse data of high values (Matheron 1982). A practical solution to this problem is to use the structure of the indicator-variogram at the median (median-indicator) for all indicators. For details of various techniques used in the indicator-based estimation procedures, one may see and Goovaerts (1997). Another way of addressing the loss of information in indicator kriging is the technique of probability kriging where the cross-covariances between the indicators and Z are modeled and utilized for cokriging the indicator functions.

Projection onto Larger Spaces

Consider the space ζ of all random variables of the form $g(Z(x_1), Z(x_2), \dots, Z(x_N))$ with finite variances. Since random variables can be added and multiplied by real numbers a vector-space structure may be imposed on ζ in a natural way. Since random variables can be multiplied, an inner product may be defined as follows: For X, Y in ζ , $\langle X, Y \rangle = E[XY]$, the uncentered covariance between X and Y . If $Z(x_0) \equiv Z^0$, say lies outside ζ , and Z_ζ^{0*} is its projection on ζ , Z_ζ^{0*} is characterized by the residual $Z^0 - Z_\zeta^{0*}$ being orthogonal to every member of ζ , that is, $\langle Z^0 - Z_\zeta^{0*}, X \rangle = 0$ for all $X \in \zeta$. Similarly, if ζ_D is a subspace of ζ and Z is to be projected into ζ_D , the projection is $Z_{\zeta_D}^{0*} \in \zeta_D$ and satisfies $\langle Z^0 - Z_{\zeta_D}^{0*}, H \rangle = 0, \forall H \in \zeta_D$. The variance of estimation with respect to an estimator Z^{0*} is given by $\|(Z^0 - Z^{0*})^2\|$.

A property of the projection is that $\|(Z^0 - Z_\zeta^{0*})^2\| \leq$

$$\|(Z^0 - Z_\xi^{0*})^2\| \text{ for } \xi \subset \zeta \text{ where } Z_\xi^{0*} \text{ is the projection of } Z^0$$

onto ξ . Linear kriging estimators discussed previously are projections onto the space ζ_0 for linear functions of $Z(x_1), Z(x_2), \dots, Z(x_N)$ or subsets of ζ_0 , where $\zeta_0 \subset \zeta_D \subset \zeta$. For the purpose of reducing variance of estimation, it is desirable to project onto the largest space possible. Projection of Z^0 onto ζ requires knowledge of the $N + 1$ -variate joint distribution of $Z(x_0), Z(x_1), Z(x_2), \dots, Z(x_N)$ which is generally unavailable. However, if the random process Z is modeled as a Gaussian process, this joint distribution becomes available as part of the assumption and therefore projection onto ζ is possible.

As mentioned earlier, if Z is a Gaussian process, then all the necessary joint distributions are available and Z^0 may be predicted by the conditional expectation $E[Z^0 | z^1, z^2, \dots, z^N]$ given by $\int_{-\infty}^{\infty} z g(z | z^1, z^2, \dots, z^N) dz$ where $g(z | z^1, z^2, \dots, z^N)$

is the Gaussian density function of the random variable $Z^0 | z^1, z^2, \dots, z^N$ with mean given by z_{SK}^* , the simple kriging of Z^0 using $z^1, z^2, \dots, z^N$ and variance given by σ_{SK}^2 , the variance of estimation of simple kriging. The conditional expectation $E[Z^0 | Z^1, Z^2, \dots, Z^N] = Z_{SK}^*$ and the conditional variance is σ_{SK}^2 are optimal in the bigger class ζ .

If the random process Z cannot be modeled as a Gaussian process, there may exist a function ϕ such that $Y = \phi(Z)$ is a Gaussian process. This is examined in section “Disjunctive Kriging in the Bivariate Gaussian Model” and “Disjunctive Kriging” deals with the projection of Z^0 onto a space ζ_D that is larger than ζ_0 .

Disjunctive Kriging

In disjunctive kriging, the subspace ζ_D , the set of all random variables of the form $\sum_{\alpha=1}^N h_\alpha(Z^\alpha)$ is considered, where $h_1, h_2, \dots, h_N$ are functions of one variable and $Z^\alpha \equiv Z(x_\alpha), \alpha = 0, 1, \dots, N$.

To project $g(Z^0)$ onto ζ_D the projection $g(Z^0)^* = \sum_{\alpha=1}^N \Lambda_\alpha(Z^\alpha)$ satisfies

$\left\langle \sum_{\alpha=1}^N \Lambda_\alpha(Z^\alpha) - g(Z^0), H \right\rangle = 0$ for all $H \in \zeta_D$. Equivalently, $\left\langle \sum_{\alpha=1}^N \Lambda_\alpha(Z^\alpha), H \right\rangle = \langle g(Z^0), H \rangle$ for all $H \in \zeta_D$.

This holds for all $H \in \zeta_D$, $\left\langle \sum_{\alpha=1}^N \Lambda_\alpha(Z^\alpha), h_\beta(Z^\beta) \right\rangle = \langle g(Z^0), h_\beta(Z^\beta) \rangle$, $\beta = 1, 2, \dots, N$ and for all functions h_β of Z^β . The projection of $g(Z^0)^* = \sum_{\alpha=1}^N \Lambda_\alpha(Z^\alpha)$ onto ζ_β and the projection of $g(Z^0)$ onto ζ_β , where ζ_β is the vector space of all random variables $h_\beta(Z^\beta)$ of Z^β coincide for each $\beta = 1, 2, \dots, N$. In other words, $E[g(Z^0)^* | Z^\beta] = E[g(Z^0) | Z^\beta]$, $\beta = 1, 2, \dots, N$. This set of equations are the equations for disjunctive kriging and are called the *DK-equations* (Matheron 1976a).

The DK-equations involve bivariate distributions of pairs of random variables (Z^0, Z^β) and (Z^α, Z^β) , $\alpha, \beta = 1, 2, \dots, N$ and hence these sets of bivariate distributions need to be modeled appropriately.

Bivariate Distributions and Isofactorial Models

To facilitate evaluation of conditional expectation of bivariate distributions, it is necessary to model the bivariate distributions of each pair of random variables involved in the estimation procedure. Some properties of bivariate distributions are useful for this purpose and are considered in the sequel.

Let X and Y be random variables with joint probability function given by $f(x, y)$. Let $f_1(x)$ and $f_2(y)$ denote the marginals of the random variables. Let $L_2(X)$ and $L_2(Y)$ be the vector-spaces of all functions of X and Y , respectively, that are square-integrable, that is, with finite variance. Let $\{\chi_i^{(1)}, i = 0, 1, 2, \dots\}$ and $\{\chi_j^{(2)}, j = 0, 1, 2, \dots\}$ be orthogonal functions such that any member $g(X)$ and $h(Y)$ of $L_2(X)$ and $L_2(Y)$, respectively, can be written as $\sum_{n=0}^{\infty} a_n \chi_n^{(1)}(X)$ and $\sum_{n=0}^{\infty} b_n \chi_n^{(2)}(Y)$, where $a_n = E[g(X) \chi_n^{(1)}(X)]$, and $b_n = E[h(Y) \chi_n^{(2)}(Y)]$. If X and Y are absolutely continuous random variables: $a_n = \int_{-\infty}^{\infty} g(x) \chi_n^{(1)}(x) f_1(x) dx$ and $b_n = \int_{-\infty}^{\infty} h(y) \chi_n^{(2)}(y) f_2(y) dy$. The functions $\{\chi_i^{(1)}, i = 0, 1, 2, \dots\}$ and $\{\chi_j^{(2)}, j = 0, 1, 2, \dots\}$ form an orthonormal basis of $L_2(X)$ and $L_2(Y)$, respectively. Taking $\chi_0^{(1)} = \chi_0^{(2)} = 1$, the orthonormality of these basis-functions implies the following: (i) $E[\chi_n^{(1)}(X)] = E[\chi_n^{(2)}(Y)] = 0$, $n = 1, 2, \dots$, (ii) $E[\chi_n^{(1)}(X)]^2 = E[\chi_n^{(2)}(Y)]^2 = 1$, $n = 1, 2, \dots$, (iii) $E[\chi_n^{(1)}(X) \chi_m^{(1)}(X)] = E[\chi_n^{(2)}(Y) \chi_m^{(2)}(Y)] = 0$ $n \neq m$.

If $w(x, y) = \frac{f(x, y)}{f_1(x)f_2(y)}$ is such that $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} w(x, y)^2 f_1(x) f_2(y) dx dy = E[w(x, y)] < \infty$, then the joint probability function

can be written as $f(x, y) = f_1(x) f_2(y) \left[1 + \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} T_{ij}^{XY} \chi_i^{(1)}(x) \chi_j^{(2)}(y) \right]$ where, $T_{ij}^{XY} = E[\chi_i^{(1)}(X) \chi_j^{(2)}(Y)]$. The bivariate distribution is said to be *isofactorial* if $E[\chi_i^{(1)}(X) \chi_j^{(2)}(Y)] = 0$ $i \neq j$, in which case $f(x, y) = f_1(x) f_2(y) \left[1 + \sum_{i=1}^{\infty} T_{ii}^{XY} \chi_i^{(1)}(x) \chi_i^{(2)}(y) \right]$.

Consider the second-order stationary random function Z with marginal $f_1 = f_2 = f$. The bivariate distribution $f^{\alpha\beta}(x, y)$ of $Z(x_\alpha)$ and $Z(x_\beta)$ is given by $f^{\alpha\beta}(x, y) = f(x) f(y) \left[1 + \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} T_{ij}^{\alpha\beta} \chi_i(x) \chi_j(y) \right]$ and the conditional distribution of $Z(x_\alpha) | Z(x_\beta) = y$ by $f^{\alpha|\beta}(x|y) = f(x) \times \left[1 + \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} T_{ij}^{\alpha\beta} \chi_i(x) \chi_j(y) \right]$.

The DK Equations

Let x_0 be a location where the second-order random process Z is to be predicted. Let $g(Z(x_0)) \equiv g(Z^0)$, say, a function with finite variance, be a function to be estimated. Clearly, $g(Z^0) \in L_2(Z^0)$ and hence this function can be written as $\sum_{n=0}^{\infty} a_n \chi_n(Z^0)$, where $a_n = E[g(Z^0) \chi_n(Z^0)]$. The projection of a linear combination of functions onto a vector-space is the linear combination of the projection of each function onto that vector-space. Therefore, to project $g(Z^0) = \sum_{n=0}^{\infty} a_n \chi_n(Z^0)$ onto the space ζ_D generated by functions of the type $\sum_{\alpha=1}^N h_\alpha \left(Z(x_\alpha) \equiv \sum_{\alpha=1}^N h_\alpha(Z^\alpha) \right)$, say, it is enough to project each $\chi_n(Z^0)$, $n = 1, 2, \dots$ onto ζ_D . If $\chi_n^*(Z^0)$, $n = 1, 2, \dots$, is the projection of $\chi_n(Z^0)$ onto ζ_D , then the projection $g(Z^0)^*$ of $g(Z^0)$ onto ζ_D is given by $g(Z^0)^* = \sum_{n=0}^{\infty} a_n \chi_n^*(Z^0)$. In practice $g(Z^0)$ is approximated by $g(Z^0) = \sum_{n=0}^M a_n \chi_n(Z^0)$ and its estimator by $g(Z^0)^* = \sum_{n=0}^M a_n \chi_n^*(Z^0)$ where M is chosen so that $Var[g(Z^0)] = \sum_{n=1}^{\infty} a_n^2$ can be approximated to the desired level by $\sum_{n=1}^M a_n^2$.

The projection $\chi_n^*(Z^0) \in \zeta_D$ and hence $\chi_n^*(Z^0) = \sum_{\alpha=1}^N \Lambda_\alpha^{(n)}(Z^\alpha)$, where $\Lambda_\alpha^{(n)}(Z^\alpha) = \sum_{j=0}^{\infty} \lambda_{j\alpha}^{(n)} \chi_j(Z^\alpha)$. This gives $\chi_n^*(Z^0) = \sum_{\alpha=1}^N \sum_{j=0}^{\infty} \lambda_{j\alpha}^{(n)} \chi_j(Z^\alpha)$. The projection $\chi_n^*(Z^0)$ is characterized by $E[\chi_n^*(Z^0) | Z^\beta] = E[\chi_n(Z^0) | Z^\beta]$, $\beta = 1, 2, \dots, N$. Since $E[\chi_n^*(Z^0) | Z^\beta]$ and $E[\chi_n(Z^0) | Z^\beta] \in \zeta_{Z^\beta}$, (where ζ_{Z^β} is the space of all single-variable functions of Z^β (i.e., $L_2(Z^\beta)$), the expectations can be computed using

conditional distributions of the type $f^{z|\beta}(x|y) = f(x)[1 + \sum_{i=1}^{\infty} T_{ij}^{\alpha\beta} \chi_i(x) \chi_j(y)]$, $\alpha = 0, 1, \dots, N$; $\beta = 0, 1, 2, \dots, N$. The

conditional expectation, $E[\chi_l(Z^\alpha) | Z^\beta]$ is given by $E[\chi_l(Z^\alpha) | Z^\beta = y] = \int_{-\infty}^{\infty} \chi_l(x) f^{z|\beta}(x|y) dx = \int_{-\infty}^{\infty} \chi_l(x) f(x) \left[1 + \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} T_{ij}^{\alpha\beta} \chi_i(x) \chi_j(y) \right] dx = \sum_{j=1}^{\infty} T_{lj}^{\alpha\beta} \chi_j(Z^\beta)$.

The DK equations for $\chi_n^*(Z^0)$ are given by $E\left[\sum_{\alpha=1}^N \sum_{l=0}^{\infty} \lambda_{l\alpha}^{(n)} \chi_l(Z^\alpha) | Z^\beta\right] = E[\chi_n(Z^0) | Z^\beta]$, $\beta = 1, 2, \dots, N$ giving $\sum_{\alpha=1}^N \sum_{l=0}^{\infty} \lambda_{l\alpha}^{(n)} \sum_{j=1}^{\infty} T_{lj}^{\alpha\beta} \chi_j(Z^\beta) = \sum_{j=1}^{\infty} T_{nj}^{0\beta} \chi_j(Z^\beta)$, $\beta = 1, 2, \dots, N$.

Comparing coefficients of $\chi_j(Z^\beta)$, the DK equations reduce to $\sum_{\alpha=1}^N \sum_{l=0}^{\infty} \lambda_{l\alpha}^{(n)} T_{lj}^{\alpha\beta} = T_{nj}^{0\beta}$, $j = 1, 2, \dots$; $\beta = 1, 2, \dots, N$. In practice the summation over l is restricted to M terms. On examining these DK equations, it is clear that in the general case each $\chi_n(Z^0)$ is cokriged using $\chi_l(Z^\beta)$, $\beta = 1, 2, \dots, N$; $l = 0, 1, \dots, M$. The resulting cokriging matrix is the same as the simple cokriging matrix discussed in section “Cokriging” with $T_{lj}^{\alpha\beta} = \text{Cov}(\chi_l(Z^\alpha), \chi_j(Z^\beta))$. In the isofactorial case, due to the orthogonality of $(\chi_n(Z^\alpha), \chi_m(Z^\beta))$, $n \neq m$, $\alpha = 0, 1, \dots, N$; $\beta = 0, 1, \dots, N$, $T_{nm}^{\alpha\beta} = 0$, $n \neq m \quad \forall \alpha, \beta = 0, 1, \dots, N$. Thus, the simple cokriging matrix reduces to a block diagonal matrix with block diagonal matrices given by $\sum_{\alpha=1}^N \lambda_{n\alpha}^{(n)} T_{nj}^{\alpha\beta} = T_{nn}^{0\beta}$, $\beta = 1, 2, \dots, N$, for the n^{th} diagonal block. This means that the projection of $\chi_n(Z^0)$ onto ζ_D reduces to its simple kriging using only the $\chi_l(Z^\beta)$, $l = 0, 1, \dots, M$. It is therefore advantageous to use bivariate distributions that have an isofactorial form. More details on disjunctive kriging can be found in Matheron (1976a) and Armstrong and Matheron (1986).

In many applications an estimate of the “local conditional distribution,” $F(Z^0 | Z^1, Z^2, \dots, Z^N) \equiv F(Z^0 | N)$, say is obtained by taking $g(Z^0)$ to be the indicator function $I_{Z^0 \leq z} = \begin{cases} 1 & \text{if } Z^0 < z \\ 0, & \text{otherwise} \end{cases}$ and writing it as $I_{Z^0 \leq z} = \sum_{n=0}^M a_n \chi_n(Z^0)$.

Some of the important bivariate distributions that have isofactorial representations with polynomial factors are the bivariate Gaussian, bivariate Gamma, and bivariate negative binomial. More details of these and other models can be found in Matheron (1989) and Chiles and Delfiner (2012). To use these models the bivariate distribution is specified a priori and transformation of data such that at each location the random variable has the required marginal needed along with the assumption that each bivariate pair has the specified bivariate distribution. In the discrete case, the discrete diffusion

isofactorial model (Matheron 1984c; Lajaunie and Lantuejoul 1989) and the orthogonal indicator residual model (Rivoirard 1989, 1994) provide methods to develop isofactorial models based on factors that are obtained from the raw data. Broadly, the isofactorial models can also be classified into diffusion-type (e.g., bivariate Gaussian and bivariate gamma) called “models with edge effect” and other models that are called “models without edge effect” (Rivoirard 1994). In diffusion-type models (i.e., models without edge effect), between any two values in the regionalization there must exist all other intermediate values. On the other hand, in models without edge effect (such as the mosaic model and the orthogonal indicator residual model), the random function remains constant or fairly constant in small regions and varies abruptly across these. It is conceptualized that the random function can be realized over an underlying tessellation of tiles (or mosaics) wherein across tiles random variables do not show correlation whereas within tiles they may be constant, as in the mosaic model, or vary with a specified covariance structure from the centre to the edge, as in the orthogonal indicator residual model. The mosaic model is one in which corregionalized random variables show intrinsic correlation whereby cokriging reduces to kriging. Hence, for example, random functions that are modeled using the mosaic model justify replacing indicator-cokriging with indicator kriging (Matheron 1982).

Reducing the cokriging of components of a random vector $\mathbf{Z}(x) = (Z_1(x), Z_2(x), \dots, Z_p(x))^T$ to kriging of each of its components $Z_i(x)$, $i = 1, 2, \dots, p$ requires spatial orthogonality of the process $\mathbf{Z}$, that is, $\text{Cov}(Z_i(x), Z_j(x + \vec{h})) = 0$, $\forall x, \vec{h}, i \neq j$. Under the assumption of stationarity, methods such as principal components give transformations for orthogonality at $\|\vec{h}\| = 0$. The same transformation will not generally result in spatial orthogonality. Methods such as indicator principal component kriging assume that correlation between principal components is small for small values of $\|\vec{h}\|$. The problem is to obtain a transform $\mathbf{Y} = \mathbf{AZ}$ such that $\mathbf{AC}(\vec{h})\mathbf{A}^T = \mathbf{D}$ where $\mathbf{D}$ is a diagonal matrix for all $\vec{h}$. This is the problem of simultaneous diagonalization of the family of matrices $\mathbf{C}(\vec{h})$. Subramanyam and Pandalai (2001, 2004) give conditions under which this is possible. For models such as the linear model of corregionalization where $\mathbf{C}(\vec{h}) = \sum_{l=1}^K \mathbf{B}_l c^l(\vec{h})$, it is necessary that the matrices $\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_k$ form a commuting family of matrices so that simultaneous diagonalization is possible. When $k = 1$, that is, the intrinsic model, this condition is always met. Switzer and Green (1984) and Green et al. (1988) have shown that a two-stage rotation procedure always achieves spatial orthogonality. The factors obtained by this procedure are

called MIN/MAF autocorrelation factors. Details of the procedure as applied in geostatistics can be found in Vargas-Guzmán and Dimitrakopoulos (2003). The technique therefore allows building models where transforms of indicator functions with covariance models such as LMC where $k = 2$ are spatially orthogonal and hence give a general model for constructing isofactorial models (Rivoirard et al. 2014).

There are situations where certain covariance structures such as the Markov models (Journel 1999) along with specific data configurations result in reduction of cokriging systems to kriging Subramanyam and Pandalai (2008).

Disjunctive Kriging in the Bivariate Gaussian Model

The bivariate Gaussian distribution of random variables X and Y given by $g(x, y)$ can be written as $g(x, y) =$

$g(x)g(y) \left[1 + \sum_{n=1}^{\infty} \frac{\rho^n}{n!} H_n(x)H_n(y) \right]$, where $g(x)$, $g(y)$ are $N(0, 1)$ densities, $\rho = \text{Cov}(X, Y)$, $\rho^n/n! = \text{Cov}(H_n(X), H_n(Y))$ with $H_n(x)$ and $H_n(y)$ being the n^{th} Hermite polynomials given by $H_n(x) = \frac{1}{g(x)} \frac{\partial^n g(x)}{\partial x^n}$ and $H_n(y) = \frac{1}{g(y)} \frac{\partial^n g(y)}{\partial y^n}$. The Hermite polynomials are orthogonal polynomials with $H_0(X) = 1$, $E[H_n(X)] = 0$, and $\text{Var}(H_n(X)) = n!$, $n = 1, 2, \dots$. Thus $\frac{H_n(X)}{\sqrt{n!}} = \eta_n(X)$, say $n = 0, 1, \dots$, are orthonormal polynomials.

If $Z(x) = \varphi(Y(x))$, where $Y(x)$ is $N(0, 1)$ and if it is assumed that $(Y(x), Y(x + \vec{h}))$ are standard bivariate normal, that is, $N(0, 0, 1, 1, \rho_h)$, then the process Y is a bivariate Gaussian process. The bivariate density of all pairs of random variables $(Y(x), Y(x + \vec{h}))$ is isofactorial with $g(y(x), y(x + \vec{h})) = g(y(x))g(y(x + \vec{h})) \left[1 + \sum_{n=1}^{\infty} \frac{\rho_h^n}{n!} H_n(y(x))H_n(y(x + \vec{h})) \right]$. The DK-equations for H_n^* are given by $\sum_{\alpha=1}^N \lambda_{n\alpha} \rho_{\alpha\beta}^n = \rho_{0\beta}^n$, $\beta = 1, 2, \dots, N$.

From the observed data the transformation to $y(x)$ is obtained using the normal-score transform, that is, $y(x) = G^{-1}\hat{F}(z(x))$, where $\hat{F}(z)$ is the empirical distribution function of the observations on Z and G is the distribution function of the standard Gaussian random variable. The monotone function ϕ that gives $Z(x) = \phi(Y(x))$ where $\phi(Y(x)) = \sum_{n=0}^{\infty} \frac{a_n}{n!} H_n(Y(x))$ with $a_n = \int_{-\infty}^{\infty} \phi(y) H_n(y) g(y) dy$ is known in geostatistics as the *anamorphosis function*. The covariance function is modeled using the experimental covariance function of the transformed data or computed from the modeled covariance function of Z since $\text{Cov}(Z(x), Z(x + \vec{h})) = \text{Cov}(\phi(Y(x)), \phi(Y(x + \vec{h}))) = \sum_{n=1}^{\infty} \frac{a_n^2}{n!} \rho_h^n$.

Change of Support

Observations $z(x)$ on the random process Z at the location x are generally averages over small volumes centered at x . Averages $Z_v(x) = \frac{1}{|v|} \int_{v(x)} Z(x) dx$ of Z over volumes v centered at x are often required to be estimated (or simulated). The volume v over which the random function Z is averaged to get $Z_v(x)$ is called the *support* of the random variable $Z_v(x)$. If $Z(x)$ is taken at the point x only, it is called a *point-support* random variable. The probability distribution of $Z_v = \{Z_v(x), x \in D\}$ cannot be modeled as there are no observations on Z_v . Calculation of the probability distribution of Z_v requires all the finite-dimensional distributions of Z , which are usually not available. Hence, models for the probability distribution of Z_v need to be developed. This is referred to as the *global change of support problem*. In many estimation (and simulation) procedures, the conditional distribution of $Z_v(x_0) | Z(x_\alpha)$, $\alpha = 1, 2, \dots, N$ also needs to be specified. This is called the *local change of support problem*. Global and local change of support problems are important problems in geostatistics.

An early solution to the global change of support problem was based on the hypothesis of the *law of permanence of distributions*. Here it is assumed that the standardized probability distribution of Z_v is the same as that of Z , that is, $\frac{Z_v - m_v}{\sigma_v} \sim \frac{Z - m}{\sigma}$, where $E[Z_v] = m_v$, $E[Z] = m$, $\text{Var}(Z_v) = \sigma_v^2$ and $\text{Var}(Z) = \sigma^2$. When Z is a second-order stationary random function, $m_v = m$ and $\sigma_v^2 = \frac{1}{|v|^2} \int_v \int_v C(x - y) dx dy = \overline{C}(vv)$, say, which can be estimated using $C(\vec{h})$ representing the modeled covariance of Z which is available. Since $\frac{Z_v - m_v}{\sigma_v}$ and $\frac{Z - m}{\sigma}$ have identical distribution F , say, an estimate F^* of the distribution function of Z leads to an estimate F_v^* , the distribution function of Z_v , that is, $F_v^*(z) = F^*\left(m + \frac{\sigma}{\sigma_v}(z - m)\right)$. The global estimate $\psi^*(Z_v)$ of any function $\psi(Z_v)$ is then given by $\int_{-\infty}^{\infty} \psi(u) F_v^*(u) du$.

The problem of estimating the local conditional distribution $F_v(x; z | N) = \text{Prob}[Z_v(x) \leq z | (Z(x_\alpha) = z(x_\alpha), \alpha = 1, 2, \dots, N)]$ is achieved by post-kriging modification of $F^*(x; z | N)$.

The *affine change of support* model relies upon the “hypothesis of permanence” and the first two moments of the marginal distributions of $Z_v(x)$ and $Z(x)$ and hence a more general model is desirable for better estimation of random variables with a different support than that of the observed data. One of the important models that is used for addressing the change of support problem is known as Cartier’s relation after Matheron (1984a, b) who introduced this model for (additive) random variables that are averages of point-support random variables. This model uses the notion of a randomly chosen point $\tilde{x}$ within a volume v . Let $Z(\tilde{x})$ be the random variable associated with the randomly chosen

point $\tilde{x}$. The expected value of $Z(\tilde{x})$ is given by $E[Z(\tilde{x})] = \frac{1}{|v|} \int_{v(x)} Z(u) du = Z_v(x)$ where $Z_v(x)$ is the random variable associated with the volume v centered at x . It is now assumed that $E[Z(\tilde{x}) | Z_v(x)] = Z_v(x)$. Due to the assumption of stationarity, $E[Z(\tilde{x}) | Z_v] = Z_v$. This assumption is called *Cartier's relation*. A consequence of this is that $Var(Z_v) = \sigma_v^2 < Var(Z) = \sigma^2$. A similar relation, in respect of the grade $Z_v(\tilde{x})$ of a random block of volume v within a panel of volume V gives $E[Z_v(\tilde{x}) | Z_V(x)] = Z_V(x)$.

In mining geology, the problem of estimation of recovery functions, such as recoverable tonnage in blocks of volume v that are above specified cutoff grade, and the quantity of metal contained in such blocks has received much attention. In addition to the “*effect of support*” on the recovery functions, there is also an “*effect of information*” since selection of blocks is made on the basis of the estimated grades of these blocks and not on their true grades. This aspect has been addressed in local and global recovery estimation problems (Matheron 1976b; Marechal 1984) and led to the development of the selectivity of distributions (Matheron 1984b) which was emphasized by Matheron as the “second principle of geostatistics.”

As discussed earlier in section “*Disjunctive Kriging in the Bivariate Gaussian Model*” for the Gaussian case, it is convenient to assume $Z(x) = \phi(Y(x))$ and to choose a suitable model for the process Y which is assumed to be bivariate-stationary. This anamorphosis (transformation) function $\phi(Y(x))$ is written as before as $\phi(Y(x)) = \sum_{n=0}^{\infty} \psi_n \chi_n(Y(x))$ and in particular $Z(\tilde{x}) = \phi(Y(\tilde{x})) = \sum_{n=0}^{\infty} \psi_n \chi_n(Y(\tilde{x}))$. The Cartier's relation then gives $E[\phi(Y(\tilde{x})) | Y_v] = \phi_v(Y_v)$, where $Z_v = \phi_v(Y_v)$ with ϕ_v being the anamorphosis function for Z_v . The random variable $Z_v(x) = \frac{1}{|v|} \int_{v(x)} Z(x) dx$ can be written as $Z_v(x) = \phi_v(Y_v(x)) = \sum_{n=0}^{\infty} \theta_n \chi_n^v(Y_v(x))$ where χ_n^v , $n = 0, 1, 2, \dots$ form a basis for $L_2(Y_v(x))$ and $\theta_n = E[\phi_v(Y_v(x)) \chi_n^v(Y_v(x))]$.

To illustrate the method, assume that $Y(\tilde{x})$ and $Y_v(x)$ are standard bivariate Gaussian random variables of which the density can be written as $g(y(\tilde{x}), y_v(x)) = g(y(\tilde{x}))g(y_v(x)) \left[1 + \sum_{n=1}^{\infty} \frac{r_v^n}{n!} H_n(y(\tilde{x}))H_n(y_v(x)) \right]$ where $r_v^n = Cov(H_n(Y(\tilde{x})), H_n(Y_v(x)))$ and $\tilde{x}$ is a randomly chosen point in $v(x)$, the support of $Y_v(x)$. (Bivariate density functions such as the one above that describe the joint behavior of random variables on point and block support are called *point-block models*. In the case of bivariate Gaussian density, this model is known as the *discrete Gaussian model*.) This gives ϕ_v

$$(Y_v(x)) = \sum_{n=0}^{\infty} \frac{b_n}{n!} H_n(Y_v(x)) = \sum_{n=0}^{\infty} \frac{a_n}{n!} H_n(Y(\tilde{x})) | Y_v(x) = \sum_{n=0}^{\infty}$$

$\frac{a_n}{n!} r_v^n H_n(Y_v(x))$, where $a_n = E[\phi(Y(x)) H_n(Y(x))]$, and $b_n = E[\phi_v(Y_v(x)) H_n(Y_v(x))] = a_n r_v^n$. Under stationarity, $\phi_v(Y_v) = \sum_{n=0}^{\infty} \frac{a_n}{n!} r_v^n H_n(Y_v)$. Using the relation $Var(Z_v(x)) = Var(\phi_v(Y_v(x))) = \sum_{n=0}^{\infty} \frac{a_n^2}{n!} r_v^{2n}$, the value of r_v can be estimated from data since $Var(Z_v(x)) = \frac{1}{|v|} \int_{v(x)} \int_{v(x)} C(y-y') dy dy' = \overline{C}(vv)$ which can be computed from the modeled variogram of Z . For details of change of support in isofactorial models one may see Matheron (1984a).

Since ϕ_v is identified, $P[Z_v \leq z]$ is given by $P[\phi_v(Y_v) \leq z] = P[Y_v \leq \phi^{-1}(z)]$, which resolves the problem of global change of support since the distribution function of Z_v is specified.

For local estimation, the conditional distribution of $Z_v(x_0) | (Z(x_\alpha), \alpha = 1, 2, \dots, N)$ is required. In general, estimates of $g(Z_v(x_0)) | (Z(x_\alpha), \alpha = 1, 2, \dots, N)$ may be required, where $g(Z_v(x_0)) = h\left(Y_v(x_0) = \sum_{n=0}^{\infty} \frac{b_n}{n!} H_n(Y_v(x_0))\right)$ with $b_n = E[h(Y_v(x_0)) H_n(Y_v(x_0))]$. In disjunctive kriging, the projection $g(Z_v(x_0))^*$ of functions $g(Z_v(x_0))$ of $Z_v(x_0)$ onto ζ_D (the vector-space generated by all single variable functions of $Z(x_\alpha), \alpha = 1, 2, \dots, N$) is considered. This is characterized by $E[g(Z_v(x_0))^* | Z(x_\alpha)] = E[g(Z_v(x_0)) | Z(x_\alpha)], \alpha = 1, 2, \dots, N$, where $g(Z_v(x_0))^* = \sum_{\alpha=1}^N \Lambda_\alpha(Z(x_\alpha))$. As before the estimate

$g(Z_v(x_0))^*$ is given by $g(Z_v(x_0))^* = \sum_{n=0}^{\infty} \frac{b_n}{n!} H_n^*(Y_v(x_0))$ where $H_n^*(Y_v(x_0))$ is obtained as the solution of DK equations characterized by $E[H_n^*(Y_v(x_0)) | Y(x_\alpha)] = E[H_n(Y_v(x_0)) | Y(x_\alpha)], \alpha = 1, 2, \dots, N$. This gives the DK equations $\sum_{\beta=1}^N \lambda_{n\beta}^{(n)} \rho_{\alpha\beta}^n = \rho_{v\alpha}^n, \alpha = 1, 2, \dots, N$, where $\rho_{v\alpha} = Cov(Y_v(x_0), Y(x_\alpha))$ can be estimated from data as $Cov(Z_v(x_0), Z(x_\alpha)) = \frac{1}{|v|} \int_{v(x)} C(y, \alpha) dy = \overline{C}(v, \alpha) = \sum_{n=0}^{\infty} \frac{a_n^2}{n!} r_v^n \rho_{v\alpha}^n$.

Since $H_n(Y_v(x))$ is estimated, the estimate of any function $g(Z_v(x_0)) = h\left(Y_v(x_0) = \sum_{n=0}^{\infty} \frac{b_n}{n!} H_n(Y_v(x_0))\right)$ is obtained as $g(Z_v(x_0))^* = \sum_{n=0}^{\infty} \frac{b_n}{n!} H_n(Y_v(x_0))^*$ (approximated in practice by $g(Z_v(x_0))^* = \sum_{n=0}^M \frac{b_n}{n!} H_n(Y_v(x_0))^*$ where M is chosen suitably). In particular, the conditional expectations, $E[I_{Z_v(x_0) \leq z}(x_0) | (Z(x_\alpha), \alpha = 1, 2, \dots, N)]$ and $E[Z_v(x_0) I_{Z_v(x_0) \leq z}(x_0) | (Z(x_\alpha), \alpha = 1, 2, \dots, N)]$ are important in some applications in mining geology.

In many mining problems, larger panels of volume V may be divided into N smaller blocks of volume v . As before, writing $Z_v(x) = \phi_v(Y_v(x))$ and developing in terms of Hermite polynomials gives $\phi_v(Y_v(x)) = \sum_{n=0}^{\infty} \frac{a_n}{n!} r_v^n H_n(Y_v(x))$ where the

a_n 's are coefficients as given previously and $r_v^n n! = \text{Cov}(H_n(Y(\tilde{x})), H_n(Y_v(x)))$. It may be required to estimate recovery functions $\frac{1}{N} \sum_{i=1}^N \psi(Z_v(x_i))$ of N small blocks of volume v centered at x_i within larger panels of volume V centered at x . Writing $\psi(Z_v(x)) = \sum_{n=0}^{\infty} \frac{b_n}{n!} H_n(Y_v(x))$ where $b_n = E[\psi(Z_v(x)) H_n(Y_v(x))]$, the function required to be estimated is $\frac{1}{N} \sum_{i=1}^N \sum_{n=0}^{\infty} \frac{b_n}{n!} H_n(Y_v(x_i)) = \sum_{n=0}^{\infty} \frac{b_n}{n!} \cdot \left(\frac{1}{N} \sum_{i=1}^N H_n(Y_v(x_i)) \right)$. This problem of local estimation can be tackled by DK by writing $E \left[\frac{1}{N} \sum_{i=1}^N H_n(Y_v(x_i)) | Y(x_\alpha) \right] = \frac{1}{N} \sum_{i=1}^N \rho^n(v_i, \alpha)$ where $\rho(v_i, \alpha) = \text{Cov}(Y_i(x_i), Y(x_\alpha))$. The DK estimate $\left(\frac{1}{N} \sum_{i=1}^N H_n(Y_v(x_i)) \right)^{DK} = \left(\frac{1}{N} \sum_{i=1}^N H_n(Y_v(x_i)) \right)$ is obtained by solving the DK equations $\sum_{\alpha=1}^N \lambda_{n\alpha} \rho_{n\alpha}^n = \frac{1}{N} \sum_{i=1}^N \rho^n(v_i, \beta)$, $\beta = 1, 2, \dots, N$.

Another problem is to estimate $E \left[\frac{1}{N} \sum_{i=1}^N \psi(Z_v(x_i)) | Z_V(x) \right]$, where $Z_V(x)$ is the grade of the panel of volume V centered at x written as $Z_V(x) = \phi_V(Y_V(x)) = \sum_{n=0}^{\infty} \frac{a_n}{n!} r_V^n H_n(Y_V(x))$ where r_V is obtained from $\text{Var}(Z_V(x))$ as explained previously and $Z_v(x) = \sum_{n=0}^{\infty} \frac{a_n}{n!} r_v^n H_n(Y_v(x))$. Under the discrete Gaussian model, the quantity to be estimated can be written as the conditional expectation, $E[\psi(Z_v(\tilde{x}) | Z_V(x))]$, where $Z_v(\tilde{x})$ is a random block of volume v within the panel of volume $V(x)$. The expected value $E[\psi(Z_v(\tilde{x}) | Z_V(x))]$ can be written as $E[\psi'(Y_v(\tilde{x}) | Y_V(x))]$, where $\psi'(Y_v(\tilde{x})) = \sum_{n=0}^{\infty} \frac{b_n}{n!} H_n(Y_v(\tilde{x}))$ with $b_n = E[\psi(Z_v(\tilde{x})) H_n(Y_v(\tilde{x}))]$. Since $E[Z_v(\tilde{x}) | Z_V(x)] = Z_v(x)$ this gives $E[H_n(Y_v(\tilde{x}) | Y_V(x))] = \left(\frac{r_v}{r_V} \right)^n$ and thus, $E[\psi'(Y_v(\tilde{x}) | Y_V(x))] = E \left[\sum_{n=0}^{\infty} \frac{b_n}{n!} H_n(Y_v(\tilde{x})) | Y_V(x) \right] = \sum_{n=0}^{\infty} \frac{b_n}{n!} \left(\frac{r_v}{r_V} \right)^n H_n(Y_V(x))$. Since $Z_V(x)$ is unknown, one of the ways in which this problem is often tackled is by making the approximation that $Z_V^*(x)$ representing the estimate of $Z_V(x)$ can be written as $Z_V^*(x) = \sum_{n=0}^{\infty} \frac{a_n}{n!} r_V^n H_n(Y_V^*(x))$. This is an approximation since $Y_V^*(x) = \phi_V^{-1}(Z_V^*(x))$ is not $N(0, 1)$. The approximation gives $E[\psi'(Y_v(\tilde{x}) | Y_V(x))] \approx \sum_{n=0}^{\infty} \frac{b_n}{n!} \left(\frac{r_v}{r_V} \right)^n H_n(Y_V^*(x))$. This method of local recovery estimation is called the method of *uniform conditioning*. Here the estimate $Z_V^*(x)$ can be obtained by any procedure such as ordinary kriging and hence is useful in situations where local stationarity can be assumed (see, e.g., Guibal and Remacre 1984). Other more exact methods of uniform conditioning to

estimate $E \left[\frac{1}{N} \sum_{i=1}^N \psi(Z_v(x_i)) | Z_V^*(x) \right]$ can be developed with the estimate $Y_V^*(x) = \sum_{\alpha=1}^K \lambda_\alpha Y(x_\alpha)$, where $Y(x_\alpha)$, $\alpha = 1, 2, \dots, K$, are the Gaussian transforms at data locations, but these methods rely on stronger assumptions of stationarity.

Other bivariate distributions with isofactorial representations (such as the bi-gamma, bivariate negative binomial, and other models, such as the orthogonal residual model and the discrete isofactorial models) also admit valid change of support models (Matheron 1989). Other change of support models, such as the mosaic change of support model (different from the mosaic model for a random function), have also been studied for specific applications (Matheron 1984d).

Another approach is the multi-Gaussian approach (see, e.g., Verly 1983; Rivoirard 1994). Here, the random variables $Y_v(x_0), Y(x_1), Y(x_2), \dots, Y(x_N)$ are multivariate Gaussian with the conditional distribution $g(Y_v(x_0) | (Y(x_\alpha) = y(x_\alpha), \alpha = 1, 2, \dots, N)) \sim N(y_v^{SK}(x_0), \sigma_{SK}^2(x_0))$, where $y_v^{SK}(x_0)$ and $\sigma_{SK}^2(x_0)$ are the simple kriging and variance of simple kriging, respectively, of $Y_v(x_0)$. Estimation or change of support problems involving functions $\psi(Z_v(x_0)) = h(Y_v(x_0) = \sum_{n=0}^{\infty} \frac{b_n}{n!} H_n(Y_v(x_0)))$ reduce to that of specifying $g(Y_v(x_0) | (Y(x_\alpha), \alpha = 1, 2, \dots, N))$ where $Y(x_\alpha) = \phi^{-1}(Z(x_\alpha))$ with ϕ being the anamorphosis function for the point-support random variables. This gives $\psi(Z_v(x_0))^* = E[h(Y_v(x_0) | Y(x_\alpha), \alpha = 1, 2, \dots, N)] = \int_{-\infty}^{\infty} h(u) g(u | (y(x_\alpha), \alpha = 1, 2, \dots, N)) d(u)$. Another equivalent way to obtain the estimate $\psi(Z_v(x_0))^* = E[h(Y_v(x_0) | Y(x_\alpha), \alpha = 1, 2, \dots, N)]$ is to evaluate it by using the development of $h(Y_v(x_0))$ in terms of Hermite polynomials. Writing $Y_v(x_0) | (Y(x_\alpha), \alpha = 1, 2, \dots, N) = (y_v^{SK}(x_0) + \sigma_{SK}(x_0)U)$, where U is uniform $[0, 1]$, the estimate is given by $\psi(Z_v(x_0))^* = E \left[\sum_{n=0}^{\infty} \frac{b_n}{n!} H_n(y_v^{SK}(x_0) + \sigma_{SK}(x_0)U) \right] = E \left[\sum_{n=0}^{\infty} \frac{b_n}{n!} H_n(Y_v(x_0)) | \frac{Y_v^{SK}(x_0)}{S} \right] = \sum_{n=0}^{\infty} \frac{b_n}{n!} E \left[H_n(Y_v(x_0)) | \frac{Y_v^{SK}(x_0)}{S} \right] = \sum_{n=0}^{\infty} \frac{b_n}{n!} S^n H_n \left(\frac{Y_v^{SK}(x_0)}{S} \right)$, where $S^2 = \text{Var}(Y_v^{SK}(x_0))$. When $Y(x_\alpha) = y(x_\alpha)$, $\alpha = 1, 2, \dots, N$, $E[h(Y_v(x_0) | Y(x_\alpha) = y(x_\alpha), \alpha = 1, 2, \dots, N)] = \sum_{n=0}^{\infty} \frac{b_n}{n!} S^n H_n \left(\frac{Y_v^{SK}(x_0)}{s} \right)$ with $s = \sum_{\alpha=1}^N \lambda_\alpha^{SK} C_Y(x_\alpha, v)$, where $C_Y(\frac{-}{h})$ is the model of the covariance function of Y .

This section should be regarded as an overview of the principle of change of support. Specific applications may require appropriate adaptation of the preceding principles.

Simulation

Natural phenomena are often complex and details of complexity vary with the scale of observation. Complex patterns are formed due to the interaction of several processes that operate independently, often at different scales. Smooth patterns that are sometimes observed at one scale become rough at other smaller scales of observation. Discerning and describing such patterns depends on the density of data and the scale at which description is desired. Information that may be available may include measurements and categorizations of observations, called *hard-data*, or other empirical information such as bounds on possible values (such as bounds on rainfall data in a region), expert opinions, order-relations (such as probable order of successions of strata in a terrain), and other subjective information that are collectively termed *soft-data*. It is of interest to make informed inferences on patterns that can be discerned from hard and soft data available on natural phenomena to make use of them to produce possible maps with desired density of detail that display patterns akin to those observed in reality.

Discerning patterns in natural phenomena on the basis of hard data can be done using deterministic or probabilistic methods. Due to lack of complete information or understanding of natural phenomena, and due to inherent randomness in many natural processes, deterministic methods often have limited success in modeling the true map of natural processes. Probabilistic methods attempt to overcome some of the lacunae by using models that try to capture the nature and form of randomness in natural processes. In the probabilistic approach, natural phenomena are considered as random processes with data being observations on them. The description of the process is defined by modeling the joint distribution of the collection of random variables that constitute the random process by parametrizing the model and estimating parameters from observed data. In practice, various assumptions of homogeneity (stationarity) are invoked to reduce the modeling to second-order moments involving covariances which are defined through pairs of random variables (and hence termed *two-point statistics*). This approach is useful when the phenomena under study exhibit a degree of homogeneity at the desired scale. An example is the likely homogeneity of reflectances in pixels in specific bands of the electromagnetic spectrum over the canopy of a featureless forested terrain. On the other hand, if the terrain has topographic features and includes forested terrains plus braided and meandering streams with fluvial sand bars and wash-over flats, modeling the observed satellite or areal image over any band of the spectrum becomes complicated. Another example of a complicated reality could be the problem of modeling average grades of ore in volumes of rock where the orebody consists of a randomly branching system of veins. Probabilistic methods based on two-point statistics may become incapable

of capturing the true nature of a natural process at some desired scale as the process becomes more and more complex. The performance of probabilistic methods may be enhanced by considering *multi-point statistics* such as data-driven models of joint distributions. For example, in porous strata, the characteristic of interest could be the connectedness of pores of a minimum size. Such problems could be addressed by examining expected values of the product of indicators at multiple locations separated by a vector $\vec{h}$.

In the estimation problem dealt with in previous sections, the objective is to obtain optimal values of the random process Z at unobserved locations $s_1, s_2, \dots, s_M$, say, based on observations on Z at locations $x_1, x_2, \dots, x_N$ wherein optimization is performed by minimizing $\text{Var}(Z(s_i) - Z^*(s_i))$ at location s_i , where $Z^*(s_i)$ is a suitable estimator of the unknown $Z(s_i)$. The optimization results in *smoothing*, that is, the predicted values are less dispersed than the actual values. A consequence of this is that the inherent variability of the random process Z is not reproduced in the set of predicted values, and moreover, the patterns or features that are characteristic of Z may not be manifested in the map of estimated values.

An alternative is to remove the smoothing effect on the predicted map of Z by perturbing or roughening it while trying to retain essential features that characterize the random phenomenon. This process is called *simulation*. There are many techniques that have been explored for performing simulations of random processes. If T denotes hard and soft data on the random process Z , simulation can be considered as a mapping $M : T \rightarrow S$ where S is the set of possible maps. Simulation which uses only hard data is discussed first in the sequel and is followed by methods that use both hard data and soft information. Object-based simulation techniques fall at the opposite end of the spectrum and place emphasis on reproduction of geometries of geological objects along with their lateral and vertical successions. In the sequel some simulation techniques are discussed in detail and a summary of some of the other techniques is provided at the end.

Simulation of a Second-Order Stationary Process

Let $Z = \{Z(x) \in D\}$ be a random process characterized by the collection of all k -variate distribution functions $F^{(k)}$ for $k = 1, 2, \dots$ and all possible choice of support points $x \in D$. Let $s_1, s_2, \dots, s_M$ be locations of interest at which Z is not observed and let $F^{(M)} = F_{Z(s_1), Z(s_2), \dots, Z(s_M)}$ be the M -variate joint distribution of $Z(s_1), Z(s_2), \dots, Z(s_M)$. A realization $z^s(s_1), z^s(s_2), \dots, z^s(s_M)$ of Z at $s_1, s_2, \dots, s_M$ generated from $F^{(M)}$ is said to be a *non-conditional simulation* $Z^{(NS)}$ of Z at $s_1, s_2, \dots, s_M$. If $F^{(N+M)} = F_{Z(s_1), Z(s_2), \dots, Z(s_M), Z(x_1), Z(x_2), \dots, Z(x_N)}$ is the joint distribution of $Z(s_1), Z(s_2), \dots, Z(s_M), Z(x_1), Z(x_2), \dots, Z(x_N)$ and $z(x_1), z(x_2), \dots, z(x_N)$ are observations on the random variables

$Z(x_1), Z(x_2), \dots, Z(x_N)$, then a realization $z^s(s_1), z^s(s_2), \dots, z^s(s_M), z(x_1), z(x_2), \dots, z(x_N)$ generated from the joint distribution $F^{(M+N)}$ is said to be a *conditional simulation* $Z^{(CS)}$ of Z at $s_1, s_2, \dots, s_M$. At any location $s \in D$, the random variable $Z(s)$ can be written as $Z(s) = Z^*(s) + (Z(s) - Z^*(s))$ where $Z^*(s)$ is an estimate of $Z(s)$ and $Z(s) - Z^*(s)$ is called error. If the error is replaced by an error $(Z^{(NS)}(s) - Z^{(NS)*}(s))$ called *isomorphous error* using a non-conditional simulation $Z^{(NS)}$ at $s_1, s_2, \dots, s_M, x_1, x_2, \dots, x_N$ from the joint distribution $F^{(M+N)}$, independent of the data at $x_1, x_2, \dots, x_N$, where $Z^{(NS)*}(s)$ is the estimate of $Z^{(NS)}(s)$ using $Z^{(NS)}(x_1), Z^{(NS)}(x_2), \dots, Z^{(NS)}(x_N)$, then $Z^{(CS)}(s) = Z^*(s) + (Z^{(NS)}(s) - Z^{(NS)*}(s))$ with $Z^{(CS)}(x)$ being the conditional simulation of Z at s . Many simulation methods build isomorphous error through various techniques specific to them.

If $\mathbf{Z} = \{Z(x), x \in D\}$, where $\mathbf{Z}(x)$ is a p -dimensional random vector, then the conditional simulation $\mathbf{Z}^{(CS)}(s)$ or non-conditional simulation $\mathbf{Z}^{(NS)}(s)$, at location s , can be generated by simultaneously simulating components of $\mathbf{Z}(s)$. Such simulation is called *co-simulation*. Co-simulation may often use cokriging as an intermediate step in the procedure.

Similarly, if $\mathbf{Y} = \Psi(\mathbf{Z})$, then conditional or non-conditional simulations $\mathbf{Y}^S$ of $\mathbf{Y}$ at $s_1, s_2, \dots, s_M$ can be built from appropriate joint distributions. If Ψ is the indicator function $(I_{Z(x) < z_j}, j = 1, 2, \dots, p)^T$, then co-simulation of the p -variate random function $\mathbf{Y}$ is called *indicator co-simulation*. If each component of $\mathbf{Y}$ is simulated independently using only information on that component, then the simulation is called *indicator simulation*.

Let $\mathbf{C}_{11}$ be the covariance matrix of $\mathbf{Z}^D = (Z(x_1), Z(x_2), \dots, Z(x_N))^T$, $\mathbf{C}_{22}$ that of $\mathbf{Z}^G = (Z(s_1), Z(s_2), \dots, Z(s_M))^T$ and $\mathbf{C}_{12}$ the matrix of cross-covariances and let $\mathbf{C} =$

$$\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} \text{ be the covariance matrix of } \mathbf{Z} = [\mathbf{Z}^D \ \mathbf{Z}^G]^T.$$

Let $\mathbf{L}$ be such that $\mathbf{LL}^T = \mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}$. Let $\mathbf{Z}^S = (Z^s(s_1), Z^s(s_2), \dots, Z^s(s_M))^T$ and let $\mathbf{Z}^S = \mathbf{Z}_s^* + \mathbf{LW}$ where $\mathbf{W} = (W_1, W_2, \dots, W_M)^T$ is a vector of independent random variables with mean $\mathbf{0}$ and variance $\mathbf{1}$ and $\mathbf{Z}_s^* = (Z^*(s_1), Z^*(s_2), \dots, Z^*(s_M))^T$ is the vector of kriged estimates of $\mathbf{Z}^G = (Z(s_1), Z(s_2), \dots, Z(s_M))^T$. If Z is a second-order stationary process, $(z(x_1), z(x_2), \dots, z(x_N), z^s(s_1), z^s(s_2), \dots, z^s(s_M))^T$ is a realization of a second-order stationary process with the covariance structure $\mathbf{C}$. This process of obtaining $\mathbf{Z}^S$ is a method of simulating the random process Z at locations $s_1, s_2, \dots, s_M$. If Z is assumed to be a Gaussian process and $W_1, W_2, \dots, W_M$ are independent $N(0, 1)$ random variables, the realization $(z(x_1), z(x_2), \dots, z(x_N), z^s(s_1), z^s(s_2), \dots, z^s(s_M))^T$ is a realization from the joint distribution of $\mathbf{Z} = [\mathbf{Z}^D \ \mathbf{Z}^G]^T$ which is multi-Gaussian. In this case $\mathbf{LL}^T$ is the covariance matrix of

$\mathbf{Z}^G | \mathbf{Z}^D$ and $\mathbf{z}^S$ is a realization from this conditional distribution. For this reason, the matrix $\mathbf{LL}^T$ is called the *conditional covariance matrix*.

Sequential Gaussian Simulation

A node-by-node implementation of the procedure detailed in the previous section results in a simulation technique known as the *Sequential Gaussian Simulation*. As before, let the random process Z be observed at $x_1, x_2, \dots, x_N$ and let $Z(x) = \phi(Y(x))$, $x \in D$ and let $Y(x) \sim N(0, 1)$, $x \in D$, where ϕ is a suitable anamorphosis function as discussed previously. If in addition, it is assumed that Y is a Gaussian process, $Y = \phi^{-1}(Z)$ has covariance matrix

$$\mathbf{C}_Y = \begin{bmatrix} \mathbf{C}_{11}^Y & \mathbf{C}_{12}^Y \\ \mathbf{C}_{21}^Y & \mathbf{C}_{22}^Y \end{bmatrix}. \text{ Sequential Gaussian simulation proceeds}$$

by defining a random path through locations that are to be simulated. Now let $s_1, s_2, \dots, s_M$ be that ordered path of locations to be simulated. Let s_i be the node to be simulated at the i^{th} step of the procedure. The simulated value is $z^S(s_i) = \phi(y^S(s_i))$ where $y^S(s_i) = y_{SK}^*(s_i) + \sigma_{SK}u$ with $y_{SK}^*(s_i)$ being the simple kriging of $y(s_i)$. The kriging is done with $y(x_1), y(x_2), \dots, y(x_N), y^s(s_1), y^s(s_2), \dots, y^s(s_{i-1})$, the N data $(y(x_1), y(x_2), \dots, y(x_N))^T = \mathbf{y}^D$, say, and the $i-1$ previously simulated values. The corresponding variance of estimation by simple kriging is given by $\sigma_{SK}^2(s_i)$ and u is independently drawn from $N(0, 1)$. The simulated value $y^S(s_i)$ at the i^{th} step is a realization of the random variable $Y^S(s_i) = y_{SK}^*(s_i) + \sigma_{SK}U$ where $y_{SK}^*(s_i) = E[Y(s_i) | \mathbf{y}^D, y^s(s_1), y^s(s_2), \dots, y^s(s_{i-1})]$ and $\sigma_{SK}^2(s_i)$ is the variance of $Y(s_i) | \mathbf{y}^D, y^s(s_1), y^s(s_2), \dots, y^s(s_{i-1})$. It can be shown that $y(x_1), y(x_2), \dots, y(x_N), y^s(s_1), y^s(s_2), \dots, y^s(s_M)$ is a realization from the process Y . This gives $\mathbf{Z}^S = \phi(\mathbf{Y}^S)$. At the i^{th} step of sequential Gaussian simulation, $y(s_i)$ is the realization of $Y(s_i)$ with conditional distribution $f(y(s_i) | \mathbf{y}^D, y(s_1), y(s_2), \dots, y(s_{i-1}))$. On completion of simulation, $y(x_1), y(x_2), \dots, y(x_N), y(s_1), y(s_2), \dots, y(s_M)$ is a realization from the required joint probability function $f(y(x_1), y(x_2), \dots, y(x_N), y(s_1), y(s_2), \dots, y(s_M))$ which is obtained as the product $f(\mathbf{y}^D) \cdot f(y(s_1) | \mathbf{y}^D) \cdot f(y(s_2) | \mathbf{y}^D, y(s_1)) \cdot \dots \cdot f(y(s_M) | \mathbf{y}^D, y(s_1), \dots, y(s_{M-1}))$ since at each step, u is generated from independent $N(0, 1)$ distributions.

Sequential Indicator Simulation

A non-parametric method of simulation that has found wide application is that of *Sequential Indicator Simulation* (Journel and Alabert 1988). Here, at each location s where it is required to simulate the random field Z , an estimate $\hat{F}_{Z(s)}(z|N)$ given by $\hat{F}_{Z(s)}(z|N) = E[I_{Z(s) \leq z} | N]$ (i.e., the kriged value of $I_{Z(s) \leq z}$ at location s as discussed previously in section “Non-Parametric Estimation: Multiple Indicator Kriging”). A value $z^s(s)$ is drawn from $\hat{F}_{Z(s)}(z|N)$. To simulate the next point along a random path through the set $\{s\}$ of all points

where simulation is desired, the process is repeated with the newly generated $z^s(s)$ included as part of data.

Simulation by Matrix Decomposition

When M and N are large, finding $\mathbf{L}$ such that $\mathbf{LL}^T = \mathbf{C}$ poses computational challenges. A method of finding $\mathbf{L}$ is the Cholesky decomposition by partitioning described below. To implement the simulation $\mathbf{Z}^S = \mathbf{Z}_s^* + \mathbf{LW}$ as described in earlier, let $\mathbf{G}$, the set of nodes at which it is desired to simulate the random function $\mathbf{Z}$, be partitioned into $\mathbf{G}_1, \mathbf{G}_2, \dots, \mathbf{G}_p$ where $\mathbf{G}_i$, $i = 1, 2, \dots, p$ contain r_i elements, $i = 1, 2, \dots, p$ and let $\mathbf{Z}^{G_i}$, $i = 1, 2, \dots, p$ be the corresponding partition of $\mathbf{Z}^G$. The covariance matrix of $\mathbf{Z}^G | \mathbf{Z}^D$, viz. $\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}$, then can be partitioned as

$$\begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \cdots & \mathbf{V}_{1p} \\ \mathbf{V}_{21} & \mathbf{V}_{22} & \cdots & \mathbf{V}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{V}_{p1} & \mathbf{V}_{p2} & \cdots & \mathbf{V}_{pp} \end{bmatrix}$$

where $\mathbf{V}_{ij}$ is an $r_i \times r_j$ matrix where $i, j = 1, 2, \dots, p$ and $r_1 + r_2 + \dots + r_p = M$. Let

$$\mathbf{L} = \begin{bmatrix} \mathbf{L}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{L}_{21} & \mathbf{L}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{L}_{p1} & \mathbf{L}_{p2} & \cdots & \mathbf{L}_{pp} \end{bmatrix}$$

be the corresponding partition of the Cholesky factor of $\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}$. By equating

$$\mathbf{LL}^T = \mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \cdots & \mathbf{V}_{1p} \\ \mathbf{V}_{21} & \mathbf{V}_{22} & \cdots & \mathbf{V}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{V}_{p1} & \mathbf{V}_{p2} & \cdots & \mathbf{V}_{pp} \end{bmatrix} = \mathbf{V},$$

clearly, $\mathbf{L}_{11}\mathbf{L}_{11}^T = \mathbf{V}_{11}$ and, $\mathbf{L}_{i1}\mathbf{L}_{11}^T = \mathbf{V}_{i1}$, for $i = 1, 2, \dots, p$. For $p \geq i \geq j \geq 2$, it can be seen from the partitioning of $\mathbf{V}$ that,

$$\mathbf{V}_{ij} = \sum_{k=1}^j \mathbf{L}_{ik}\mathbf{L}_{jk}^T \quad \text{and hence,} \quad \mathbf{L}_{ij}\mathbf{L}_{jj}^T = \mathbf{V}_{ij} - \sum_{k=1}^{j-1} \mathbf{L}_{ik}\mathbf{L}_{jk}^T.$$

Denoting $\mathbf{S}_{ij} = \mathbf{V}_{ij} - \sum_{k=1}^{j-1} \mathbf{L}_{ik}\mathbf{L}_{jk}^T$ for $p \geq i \geq j \geq 2$, $\mathbf{L}_{jj}$ is the Cholesky factor of $\mathbf{S}_{jj}$ and $\mathbf{L}_{ij}$ satisfies $\mathbf{L}_{ij}\mathbf{L}_{jj}^T = \mathbf{S}_{ij}$. Using these equations, the Cholesky decomposition of large matrices can be found from the decomposition of much smaller matrices.

Finally,

$$\begin{aligned} \mathbf{Z}^S &= \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{Z}^D + \mathbf{LW} = \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{Z}^D + \begin{bmatrix} \mathbf{L}_{11} & \mathbf{0} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{L}_{21} & \mathbf{L}_{22} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{L}_{31} & \mathbf{L}_{32} & \mathbf{L}_{33} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \ddots & \mathbf{0} & \mathbf{0} \\ \mathbf{L}_{p1} & \mathbf{L}_{p2} & \mathbf{L}_{p3} & \cdots & \cdots & \mathbf{L}_{p,p-1} & \mathbf{L}_{pp} \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \mathbf{w}_3 \\ \vdots \\ \mathbf{w}_p \end{bmatrix} \\ &= \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{Z}^D + \begin{bmatrix} \mathbf{L}_{11}\mathbf{w}_1 \\ \mathbf{L}_{21}\mathbf{w}_1 + \mathbf{L}_{22}\mathbf{w}_2 \\ \mathbf{L}_{31}\mathbf{w}_1 + \mathbf{L}_{32}\mathbf{w}_2 + \mathbf{L}_{33}\mathbf{w}_3 \\ \vdots \\ \mathbf{L}_{p1}\mathbf{w}_1 + \mathbf{L}_{p2}\mathbf{w}_2 + \cdots + \mathbf{L}_{pp}\mathbf{w}_p \end{bmatrix} = \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{Z}^D + \sum_{i=1}^p \mathbf{L}^{(i)}\mathbf{w}_i \end{aligned}$$

where $\mathbf{L}^{(i)}$ is the i^{th} column of $\mathbf{L}$. If the matrix $\mathbf{C}_{21}$ =

$(\mathbf{C}_{21}^{(1)}, \mathbf{C}_{21}^{(2)}, \dots, \mathbf{C}_{21}^{(p)})^T$ is partitioned into p blocks,

compatible with the partition of $\mathbf{L}$, one may rewrite the above equations as

$$\mathbf{Z}^S = \begin{bmatrix} \mathbf{C}_{21}^{(1)} \mathbf{C}_{11}^{-1} \mathbf{Z}^D \\ \mathbf{C}_{21}^{(2)} \mathbf{C}_{11}^{-1} \mathbf{Z}^D \\ \mathbf{C}_{21}^{(3)} \mathbf{C}_{11}^{-1} \mathbf{Z}^D \\ \vdots \\ \mathbf{C}_{21}^{(p)} \mathbf{C}_{11}^{-1} \mathbf{Z}^D \end{bmatrix} + \begin{bmatrix} \mathbf{L}_{11} \mathbf{w}_1 \\ \mathbf{L}_{21} \mathbf{w}_1 + \mathbf{L}_{22} \mathbf{w}_2 \\ \mathbf{L}_{31} \mathbf{w}_1 + \mathbf{L}_{32} \mathbf{w}_2 + \mathbf{L}_{33} \mathbf{w}_3 \\ \vdots \\ \mathbf{L}_{p1} \mathbf{w}_1 + \mathbf{L}_{p2} \mathbf{w}_2 + \cdots + \mathbf{L}_{pp} \mathbf{w}_p \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{Z}_{(1)}^S \\ \mathbf{Z}_{(2)}^S \\ \mathbf{Z}_{(3)}^S \\ \vdots \\ \mathbf{Z}_{(p)}^S \end{bmatrix}, \text{ say.}$$

This equation shows that if the calculation of all $\mathbf{L}_{aj}, j \leq \alpha$, $\alpha \leq i$ has been completed, then, $\mathbf{Z}^{(1)S}, \mathbf{Z}^{(2)S}, \mathbf{Z}^{(3)S}, \dots, \mathbf{Z}^{(i)S}$, the simulations up to the i^{th} block can be obtained. If the calculations proceed row-wise and if $\mathbf{Z}$ and $\mathbf{W}_i$'s are appropriately multi-Gaussian, then the simulation is the Sequential Gaussian Simulation.

Conditional Simulation Using Successive Residuals

Let $\mathbf{D}$ be the set of locations where data is available and let $\mathbf{D}$ be partitioned into $\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_l$. The grouping of data-locations into the partitions may be in any order. Thus $\mathbf{Z} = (\mathbf{Z}^D, \mathbf{Z}^G)^T$ is now partitioned into $(\mathbf{Z}^{D_1}, \mathbf{Z}^{D_2}, \dots, \mathbf{Z}^{D_l}, \mathbf{Z}^{G_1}, \mathbf{Z}^{G_2}, \dots, \mathbf{Z}^{G_p})^T$. The covariance matrix $\mathbf{C}$ of $\mathbf{Z}$ is correspondingly partitioned into $\mathbf{C} = [\mathbf{V}_{ij}]$ where $\mathbf{V}_{ij}$ is the ij^{th} block of the partition. The algorithm for computation of the lower-triangular matrix $\mathbf{L} = [\mathbf{L}_{ij}]$ where $\mathbf{L}\mathbf{L}^T = \mathbf{C}$, with $\mathbf{L}_{ij}$ corresponding to the ij^{th} block can be obtained by the algorithm described in the previous section with newly defined $[\mathbf{V}_{ij}]$. If $\mathbf{L}$ is rewritten as $\mathbf{L} = \begin{bmatrix} \mathbf{L}^{11} & \mathbf{0} \\ \mathbf{L}^{21} & \mathbf{L}^{22} \end{bmatrix}$ corresponding to $\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}$, the required Cholesky decomposition of $\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}$ is given by $\mathbf{L}^{22}$. This avoids the explicit prior computation of the Schur complement $\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}$ of $\mathbf{C}_{11}$ as required in the previous algorithm. The algorithm for Cholesky decomposition can be performed in multiple ways where $\mathbf{L}_{ij} = \left[\mathbf{V}_{ij} - \sum_{k=1}^{j-1} \mathbf{L}_{ik} \mathbf{L}_{jk}^T \right] \mathbf{L}_{jj}^{T-1}$ for $j \leq i$. This computation can be seen as a series of updating operations on $\mathbf{V}_{ij}$ that can be written as $\mathbf{V}_{ij}^{(k)} = \mathbf{V}_{ij}^{(k-1)} - \mathbf{L}_{ik} \mathbf{L}_{jk}^T$ where, the superscript in brackets denotes the stage of updating of $\mathbf{V}_{ij}$, with $\mathbf{V}_{ij}^{(0)} = \mathbf{V}_{ij}$.

One may write, $\mathbf{S}_{ij} = \mathbf{V}_{ij} - \sum_{k=1}^{j-1} \mathbf{L}_{ik} \mathbf{L}_{jk}^T = \xi_{ij}^{j-1}$ where, in the notation of Vargas-Guzmán and Dimitrakopoulos (2002), ξ_{ij}^{j-1} is the $j-1^{th}$ successive residual conditional covariance matrix corresponding to the i^{th} and j^{th} blocks.

Truncated Gaussian Simulation and Plurigaussian Simulation
The random function $\mathbf{Z} = \{\mathbf{Z}(x), x \in D\}$ may be such that some of the components of $\mathbf{Z}$ may be categorical. For example in an iron-ore deposit, at each location x , there may be different types of ore such as "hard-ore," "laminated ore," "soft-ore," and "blue-dust," based on physical parameters of the ore. Another variable of interest could be the value of $\text{Fe}_2\text{O}_3\%$ of the sample. A third variable could be a classification of the sample into an ordered list A,B,C,D based on the quality of ore vis-à-vis its industrial utility with A grade better than B and so on. Here, the random vector $\mathbf{Z}(x) = (Z_1(x), Z_2(x), Z_3(x))^T$ where $Z_1(x)$ is a nominal random variable and corresponds to the type of ore, $Z_2(x)$ is a numerical random variable and corresponds to $\text{Fe}_2\text{O}_3\%$ and $Z_3(x)$ is an ordinal random variable that corresponds to the quality of ore. It may be required to obtain simulations of all three types of random variables. In the sequel, the case where all components of $\mathbf{Z}$ are categorical is considered first.

For the purpose of discussion, consider $\mathbf{Z}$ to be a one-dimensional categorical random process. Let $F = \{F(x), x \in D\}$ be a random process taking values $1, 2, \dots, n$. Let $\mathbf{Y} = \{\mathbf{Y}(x), x \in D\}$ be a k -dimensional standard Gaussian random process where the components $Y_i, i = 1, 2, \dots, k$ are an independent standard Gaussian process coregionalized with F . Let $(A_1, A_2, \dots, A_n)$ partition $\mathcal{R}^k$ and define $F(x) = i \Leftrightarrow \mathbf{Y}(x) \in A_i, i = 1, 2, \dots, n$. Since F is defined through the process $\mathbf{Y}$, the process $\mathbf{Y}$ is considered to be latent with respect to F . Let $C_{ij}(\vec{h}) = E[I_{F(x)=i}(x)I_{F(x+h)=j}(x+\vec{h})] = P[F(x)=i, F(x+\vec{h})=j]$. It can be seen that $C_{ij}(h) = E[I_{A_i}(\mathbf{Y}(x))I_{A_j}(\mathbf{Y}(x+\vec{h}))] = \int \int g_{\mathbf{p}}(\vec{h})(\mathbf{u}, \mathbf{v}) d\mathbf{u} d\mathbf{v}$ where $g_{\mathbf{p}}(\vec{h})$ is the probability density function of $N_{2k}(\mathbf{0}, \mathbf{0}, \mathbf{1}, \mathbf{1}, \mathbf{p}(\vec{h}))$ where $\mathbf{p}(\vec{h}) = [\rho_{ij}(\vec{h})]$ with $\rho_{ij}(\vec{h}) = \text{Cov}(Y_i(x), Y_j(x+\vec{h})) = \begin{cases} \rho_i(\vec{h}) & \text{if } i=j \\ 0, & \text{otherwise} \end{cases}$, and the A_i are chosen such that $E[I_{A_i}(\mathbf{Y}(x))] = P[F(x)=i] = p_i(x)$, the proportion of the i^{th} category at x . The indicator functions $I_{A_i}(Y_i(x)) = \begin{cases} 1 & \text{if } Y_i(x) \in A_i \\ 0 & \text{otherwise} \end{cases} = \phi_i(Y_i(x), i = 1, 2, \dots, k$ are developed in terms of Hermite polynomials to give for the i^{th} categorical variable: $C_{Y_i}(\vec{h}) = E[I_{A_i}(Y_i(x))I_{A_i}(Y_i(x+\vec{h}))] = \left(1 - G(y_i(x))\right) \left(1 - G(y_i(x+\vec{h}))\right) + \sum_{n=0}^{\infty} \frac{H_{n-1}(y_i(x))H_{n-1}(y_i(x+\vec{h}))}{n!} \rho_i^n(\vec{h}) g(y_i(x))g(y_i(x+\vec{h}))$ where $C_{Y_i}(\vec{h})$ is the uncentered covariance of the indicator for the i^{th} category. $\hat{C}_{Y_i}(\vec{h})$ is estimated from data, from which an estimate $\hat{\rho}_i(\vec{h})$ of the function $\rho_i(\vec{h})$ is computed.

The process $\mathbf{Y}(x)$ that is simulated and the relation $F(x) = i \Leftrightarrow \mathbf{Y}(x) \in A_i$, $i = 1, 2, \dots, n$ are used to simulate F . This method of simulation of F is called *Truncated Gaussian Simulation* if $k = 1$, otherwise the simulation is called *Plurigaussian Simulation*. In the simulation procedure $\left[\rho_i(\vec{h})\right]$ has to be chosen so that $C_{ij}^S(\vec{h})$, the covariance of the simulated data, is the same as that of the model $C_{ij}(\vec{h})$ and $E[I_{A_i}(\mathbf{Y}^S(x))] = p_i(x)$ where $\mathbf{Y}^S(x)$ denotes the simulated random vector at x . If F represents a stationary process, $p_i(x) = p_i \quad \forall x$.

A limitation of the Truncated Gaussian Simulation technique is that when $\rho(\vec{h})$ is continuous, $P[Y(x) \in A_i, Y(x + \vec{h}) \in A_j] > P[Y(x) \in A_i, Y(x + \vec{h}) \in A_l] \quad \forall l : |l - i| > |j - i|$ for $\vec{h} : \rho(\vec{h}) > 0$, that is, if $F(x) = i$ then $F(x + \vec{h})$ is more likely to be the category j than the category l . This results in the simulation having an ordering and has to be borne in mind while choosing the assignment of categories to partitions of $\mathfrak{R}$. If such orderings are not observed in the data, and hence not desirable, this property is a limitation of the Truncated Gaussian Simulation technique (Galli et al. 1994). An early example of the use of Truncated Gaussian Simulation for generating the geometry of fluvio-glacial reservoirs can be seen in Matheron et al. (1987).

On the other hand, when $k > 1$, the A_i 's are partitions of $\mathfrak{R}^k$ and therefore afford more flexibility in assigning categories to partitions so that more complicated relations among categories can be incorporated in the assignment of the partitions. The relations between categories would hence be reflected in plurigaussian simulations (Armstrong et al. 2011). In practice the partitions A_i are chosen as k -dimensional rectangles in $\mathfrak{R}^k$.

In applications, $F(x)$ may be observed at a finite set of points $x_1, x_2, \dots, x_N$ and the task is to simulate $F(x)$ at locations $s_1, s_2, \dots, s_M$. The values of the latent random variables at locations $x_1, x_2, \dots, x_N$ are unknown except for constraints that if $F(x) = i$, then $\mathbf{Y}(x) \in A_i$. The simulation of $F(x)$ at $s_1, s_2, \dots, s_M$ is obtained by simulating $\mathbf{Y}$ at these locations subject to constraints on $\mathbf{Y}$ at $x_1, x_2, \dots, x_N$.

In practice, the first step consists of simulating $\mathbf{Y}(x_1), \mathbf{Y}(x_2), \dots, \mathbf{Y}(x_N)$ subject to $\mathbf{Y}(x_l) \in A_i$ if $F(x_l) = i$, $l = 1, 2, \dots, N$. To do this, it can be seen that the joint distribution of $\mathbf{Y}(x_1), \mathbf{Y}(x_2), \dots, \mathbf{Y}(x_N)$ is given by $h_N(\mathbf{Y}) = \frac{1}{k} g(\mathbf{Y}) \cdot \prod_{l=1}^N \left(I_{A_{F(x_l)}}(\mathbf{Y}(x_l)) \right)$, where $k = \int_{A_{F(x_1)}} \int_{A_{F(x_2)}} \dots \int_{A_{F(x_N)}} g(\mathbf{Y}) d\mathbf{Y}$, that is, the normalizing factor, and $\mathbf{Y} = (\mathbf{Y}(x_1)^T, \mathbf{Y}(x_2)^T, \dots, \mathbf{Y}(x_N)^T)^T$. The conditional probability functions of $\mathbf{Y}(x_i) \mid (\mathbf{Y}(x_j), j \neq i)$, called *full-conditional* (that is used in the Gibbs

sampler) are written as $h(\mathbf{Y}(x_i) \mid (\mathbf{Y}(x_j), j \neq i))$, $i = 1, 2, \dots, N$ and are given by $\frac{h_N(\mathbf{Y})}{h_{N-1}(\mathbf{Y}_i)}$, where $\mathbf{Y}_i$ is $\mathbf{Y}$ without the vector $\mathbf{Y}(x_i)^T$, that is, without the i^{th} vector. When $\mathbf{Y}$ is multi-Gaussian, $\mathbf{Y}_i$ is multi-Gaussian, but the probability functions $h_N(\mathbf{Y})$, $h_{N-1}(\mathbf{Y}_i)$ and $\frac{h_N(\mathbf{Y})}{h_{N-1}(\mathbf{Y}_i)}$ are not multi-Gaussian.

The first step of plurigaussian simulation consists of generating observations from $h_N(\mathbf{Y})$. When this is done, the latent variable will be available at all $x_1, x_2, \dots, x_N$ respecting the constraints $\mathbf{Y}(x_l) \in A_i$ if $F(x_l) = i$, $l = 1, 2, \dots, N$. The second step consists of generating $F(x)$ at locations $s_1, s_2, \dots, s_M$. This reduces to one of generating the Gaussian vectors $\mathbf{Y}(s_1), \mathbf{Y}(s_2), \dots, \mathbf{Y}(s_M)$ conditional to the vectors $\mathbf{Y}(x_1), \mathbf{Y}(x_2), \dots, \mathbf{Y}(x_N)$ that are now available and obtaining the corresponding category using $F^{-1}(\mathbf{Y}(s))$. The method of performing the first step of the simulation is a modification of the Gibbs Sampler and was first suggested in the geostatistical context by Freulon and de Fouquet (1993). For the second step any of the methods that perform conditional simulation can be used, such as the sequential Gaussian simulation method, the L-U decomposition method, Accept-Reject method or the Gibbs Sampler. Since the Gibbs sampler is used in the first step, it is often convenient to use the Gibbs Sampler for the second step also. The details of the Gibbs Sampler are given in the section below.

The Gibbs Sampler

The *Gibbs Sampler* is a particular instance of the *Metropolis-Hastings Algorithm* (M-H algorithm) used to generate a Markov chain with a given stationary distribution f . If it is desired to generate an observation from $f_{\mathbf{Y}}$, then the M-H algorithm constructs an ergodic Markov chain whose state-space is the range $\mathbf{S}$ of $\mathbf{Y} = (\mathbf{Y}_1^T, \mathbf{Y}_2^T, \dots, \mathbf{Y}_N^T)^T$ and whose stationary distribution, called the *target distribution*, is $f_{\mathbf{Y}}$. Let $q(\mathbf{y} \mid \mathbf{v})$ be a conditional density called the *instrumental distribution*, where $\mathbf{y} = (\mathbf{y}_1^T, \mathbf{y}_2^T, \dots, \mathbf{y}_N^T)^T$ and $\mathbf{v} = (\mathbf{v}_1^T, \mathbf{v}_2^T, \dots, \mathbf{v}_N^T)^T \in \mathbf{S}$. Suppose the Markov chain $\mathbf{X}^{(t)}$ is in state $\mathbf{v}$ at time t , that is, $\mathbf{X}^{(t)} = \mathbf{v}$. An observation $\mathbf{y}$ is generated from $q(\cdot \mid \mathbf{v})$.

At time $t + 1$, $\mathbf{X}^{(t+1)} = \begin{cases} \mathbf{y} & \text{with probability } \rho(\mathbf{v}, \mathbf{y}) \\ \mathbf{v} & \text{with probability } 1 - \rho(\mathbf{v}, \mathbf{y}) \end{cases}$ with $\rho(\mathbf{v}, \mathbf{y}) = \min \left[\left(\frac{f_{\mathbf{Y}}(\mathbf{y}) \cdot q(\mathbf{v} \mid \mathbf{y})}{f_{\mathbf{Y}}(\mathbf{v}) \cdot q(\mathbf{y} \mid \mathbf{v})} \right), 1 \right]$. The kernel of the Markov chain $\mathbf{X}^{(t)}$ so generated verifies the detailed balance condition with $f_{\mathbf{Y}}$ as its stationary distribution (Robert and Casella 2004). Thus, for large t , $\mathbf{X}^{(t)}$ converges in distribution to $f_{\mathbf{Y}}$. The important feature of the M-H algorithm is that an observation from $f_{\mathbf{Y}}$ is generated by generating observations from $q(\cdot \mid \mathbf{v})$.

The Gibbs sampler is a special case of the M-H algorithm where the transition from $\mathbf{X}^{(t)} = \mathbf{v}$ to $\mathbf{X}^{(t+1)} = \mathbf{y}$ where

change from $\mathbf{v} = (\mathbf{v}_1^T, \mathbf{v}_2^T, \dots, \mathbf{v}_N^T)^T$ to $\mathbf{y} = (\mathbf{y}_1^T, \mathbf{y}_2^T, \dots, \mathbf{y}_N^T)^T$ occurs by change of a single vector $\mathbf{v}_i$, say, from $\mathbf{v}_i$ to $\mathbf{y}_i$, all other components remaining the same, with $\mathbf{y}_i$ drawn from $q_i(\mathbf{y}_i | (\mathbf{v}_1^T, \mathbf{v}_2^T, \dots, \mathbf{v}_{i-1}^T, \mathbf{v}_{i+1}^T, \dots, \mathbf{v}_N^T)^T)$ which is the conditional distribution of $\mathbf{Y}_i | \mathbf{Y}_j, j \neq i$, that is, the full-conditional. This gives $\rho(\mathbf{v}, \mathbf{y}) = 1$, implying a change from $\mathbf{X}^{(t)} = \mathbf{v}$ to $\mathbf{X}^{(t+1)} = \mathbf{y}$ within every step of the algorithm.

Thus, the algorithm is as follows: (i) Suppose $\mathbf{X}^{(t)} = \mathbf{v} = (\mathbf{v}_1^T, \mathbf{v}_2^T, \dots, \mathbf{v}_N^T)^T$ at time t . An integer is drawn at random between 1 and N . Let that integer be i . (ii) Generate $\mathbf{y}_i$ from the full-conditional $q_i(\cdot | (\mathbf{v}_1^T, \mathbf{v}_2^T, \dots, \mathbf{v}_{i-1}^T, \mathbf{v}_{i+1}^T, \dots, \mathbf{v}_N^T)^T)$. Then $\mathbf{X}^{(t+1)} = \mathbf{y} = (\mathbf{v}_1^T, \mathbf{v}_2^T, \dots, \mathbf{v}_{i-1}^T, \mathbf{y}_i^T, \mathbf{v}_{i+1}^T, \dots, \mathbf{v}_N^T)^T$. (iii) Assign $\mathbf{v} = (\mathbf{v}_1^T, \mathbf{v}_2^T, \dots, \mathbf{v}_{i-1}^T, \mathbf{y}_i^T, \mathbf{v}_{i+1}^T, \dots, \mathbf{v}_N^T)^T$ and loop to the first step of the algorithm with $t = t + 1$. The procedure may be started by setting $\mathbf{X}^{(0)} = \mathbf{0}^T$.

In Plurigaussian (or Truncated Gaussian) simulation, the full-conditionals $q_i(\mathbf{y}_i^T | (\mathbf{v}_1^T, \mathbf{v}_2^T, \dots, \mathbf{v}_{i-1}^T, \mathbf{v}_{i+1}^T, \dots, \mathbf{v}_N^T)^T)$ are $h(\mathbf{Y}(x_i) | (\mathbf{Y}(x_j), j \neq i)), i = 1, 2, \dots, N$. In practice, each component $Y_l(x_i), l = 1, 2, \dots, k$ of $\mathbf{Y}(x_i)$ is updated by drawing an observation from the conditional distribution $h_l(Y_l(x_i) | (\mathbf{Y}(x_j), j \neq i)), i = 1, 2, \dots, N, l = 1, 2, \dots, k$ which is a Gaussian $N(y_{l,SK}^*(x_i), \sigma_{l,SK}^2(x_i))$, where $y_{l,SK}^*(x_i)$ and $\sigma_{l,SK}^2(x_i)$ are the simple kriging estimate and variance of estimation, respectively, of the l^{th} component of $\mathbf{Y}(x_i)$ obtained using $Y_l(x_j), j \neq i$. The estimation of parameters of the Gaussian conditional distribution reduces to simple (auto-) kriging if the components of the process $\mathbf{Y}$ are independent and hence spatially orthogonal. At the first stage of Plurigaussian or Truncated Gaussian Simulation, the simulation proceeds along the lines indicated previously. At time t let the random vector $\mathbf{Y}^{(t)} = (\mathbf{y}^{(t)}(x_1)^T, \mathbf{y}^{(t)}(x_2)^T, \dots, \mathbf{y}^{(t)}(x_N)^T)^T$. A number between 1 and N is chosen randomly and i is set equal to this number. Values $y_l^{**}(x_i) = y_{l,SK}^*(x_i) + \sigma_{l,SK}(x_i) \cdot u, l = 1, 2, \dots, k$, where u is drawn independently from $N(0, 1)$ are generated. The random vector $\mathbf{Y}^{(t+1)}$ is generated by setting $\mathbf{y}^{(t+1)}(x_i) = \begin{cases} \mathbf{y}^{**}(x_i) & \text{if } \mathbf{y}^{**}(x_i) \in A_{F(x_i)} \\ \mathbf{y}^{(t)} & \text{otherwise} \end{cases}$, where $\mathbf{y}^{**}(x_i) = (y_1^{**}(x_i), y_2^{**}(x_i), \dots, y_k^{**}(x_i))^T$ and retaining $\mathbf{y}^{(t+1)}(x_j) = \mathbf{y}^{(t)}(x_j), j \neq i$ otherwise. The simulation is started by arbitrarily setting $\mathbf{Y}^{(0)} = (\mathbf{y}^{(0)}(x_1)^T, \mathbf{y}^{(0)}(x_2)^T, \dots, \mathbf{y}^{(0)}(x_N)^T)^T$ where $\mathbf{y}^{(0)}(x_i) \in A_{F(x_i)}, i = 1, 2, \dots, N$, that is, respecting the constraints at each x_i . When t is large, a kN -dimensional vector $\mathbf{Y} = (\mathbf{y}(x_1)^T, \mathbf{y}(x_2)^T, \dots, \mathbf{y}(x_N)^T)^T$ is generated that respects constraints $A_{F(x_i)}, i = 1, 2, \dots, N$ and has the desired covariance. This is the first stage of Plurigaussian Simulation. The generated vector can be then used to condition the simulation $\mathbf{Y} = (\mathbf{Y}(s_1)^T, \mathbf{Y}(s_2)^T, \dots, \mathbf{Y}(s_M)^T)^T$ in the second stage of the

simulation procedure. The final step consists of obtaining the simulated categories at $s_1, s_2, \dots, s_M$ as $F(s_i) = F^{-1}(\mathbf{Y}(s_i))$.

If the Gibbs sampler is used for both the first and second stages of the Plurigaussian (or Truncated Gaussian) procedure, both stages can be incorporated into a single operation by considering simulation of the $(k \times (M + N))$ random vector with the first corresponding to data locations and the acceptance criteria applies only to the first kM components. If M or N are very large windowing will help in reducing computation at the cost of accuracy of reproduction of the covariance function. This can be remedied by methods such as simulated annealing or placing restrictions on the transition probabilities (Emery et al. 2014).

In the Plurigaussian (or Truncated Gaussian) technique, at time $t + 1$, the Gibbs sampler generates $Y_l^{t+1}(x_i)$, the l^{th} component of the randomly chosen vector $\mathbf{Y}^{(t+1)}(x_i)$ independent of components $j \neq l$. If components of the plurigaussian process are chosen to be independent, calculation of the simple (auto) kriging estimate and variance of estimation of $Y_l(u), u \in (x_1, x_2, \dots, x_N, s_1, s_2, \dots, s_M), l = 1, 2, \dots, k$ requires a one-time computation of each $\mathbf{C}_{Y_l}^{-1}, l = 1, 2, \dots, k$. An alternate method after Galli and Gao (2001) consists of using the Gibbs sampler to generate an M and N -dimensional vector $X_l, l = 1, 2, \dots, k$ with respective covariance matrices $\mathbf{C}_{Y_l}^{-1}, l = 1, 2, \dots, k$ and then taking $Y_l = \mathbf{C}_{Y_l} X_l$. This gives $\text{Cov}[Y_l] = \mathbf{C}_{Y_l} \mathbf{C}_{Y_l}^{-1} \mathbf{C}_{Y_l} = \mathbf{C}_{Y_l}$, the desired covariance. Generating the random vector X_l using the Gibbs Sampler requires only $[\mathbf{C}_{Y_l}^{-1}]^{-1} = \mathbf{C}_{Y_l}$ and hence avoids the one-time inversion of matrices $\mathbf{C}_{Y_l}, l = 1, 2, \dots, k$. Examples of the use of the Gibbs sampler in Truncated Gaussian and Plurigaussian simulation can be seen in Emery (2007a, b) or Arroyo et al. (2012),

If a categorical random variable $F(x)$ is observed along with a numerical random vector $\mathbf{Z}(x)$ at locations $x_1, x_2, \dots, x_N$, then $F(x)$ and $\mathbf{Z}(x)$ can be co-simulated at locations $s_1, s_2, \dots, s_M$ (see for example, Maleki and Emery 2015). After generating $\mathbf{Y}(x)$ at $x_1, x_2, \dots, x_N$ obeying constraints with respect to $F(x)$, as discussed previously, the cross covariance matrix functions $\mathbf{C}_{YZ}(\vec{h})$ of $\mathbf{Y}$ and $\mathbf{Z}$ can be modeled using data at $x_1, x_2, \dots, x_N$ and used for simulation (by cokriging) at $s_1, s_2, \dots, s_M$.

Multi-Point Simulation (SNESIM)

As mentioned previously, two-point statistics such as covariance or variogram only ensures reproduction of features of the random function Z that can be described by its second-order moments. Complex features described by physical geometries of specific nature require inference of higher dimensional joint distributions which can only be done by using multi-point statistics, that is, statistics defined in terms of more than two random variables (Journel and Alabert 1989). The distributions of these statistics need to be modeled. Iterative

methods such as neural-networks have been used to infer parameters of the distributions of such mp-statistics. A non-parametric multipoint method for estimating the distribution of mp-statistics was given by Guardiano and Srivastava (1993). Although innovative, this had the limitation of requiring repeated scanning of training images for determining relevant conditional distributions at each location and hence was very CPU-intensive. Strebelle (2002) introduced *SNESIM* (Single Normal Equation Simulation) which greatly reduced the CPU requirement by using search-trees for determining conditional probability distribution functions of the required multipoint statistics. More recent advances, in the use of multipoint statistics for joint simulation can be seen in Mariethoz and Caers (2015), Minniakhmetov and Dimitrakopoulos (2017), and Yao et al. (2020).

Let Z be a one-dimensional random process where the random variable $Z(x)$ at location x is either a categorical variable taking K different values, or a numerical random variable discretized into K levels, labeled as $1, 2, \dots, K$. Let s be a location where Z has not been observed. It is desired to obtain an observation from $Z(s) \mid \left(Z(s + \vec{h}_\alpha) = k_\alpha, \alpha = 1, 2, \dots, N \right)$ where $k_\alpha \in [1, K]$. The set $\{ \vec{h}_\alpha, \alpha = 1, 2, \dots, N \} = \tau_N$, say, is called the data-geometry. The data-geometry around the location s is chosen so that it guides reproduction of the physical geometry around the location s . Let $T = \prod_{\alpha=1}^N$

$(I_{Z(s+\vec{h}_\alpha)=k_\alpha}(s+\vec{h}_\alpha))$ and let $W_k = I_{Z(s)=k}(s)$, where $I_{Z(s+\vec{h}_\alpha)=k_\alpha}(s+\vec{h}_\alpha) = \begin{cases} 1 & \text{if } Z(s+\vec{h}_\alpha) = k_\alpha \\ 0 & \text{otherwise} \end{cases}$ and $I_{Z(s)=k}(s) = \begin{cases} 1 & \text{if } Z(s) = k \\ 0 & \text{otherwise} \end{cases}$. The random variable T takes the

value 1 whenever the random variables $Z(s + \vec{h}_\alpha) = k_\alpha, \alpha = 1, 2, \dots, N$ in the data geometry τ_N around s . The problem now is to estimate $P[W_k | T]$. At the heart of the SNESIM algorithm is the *single normal equation* given by $P[W_k = 1 | T] = E[W_k] + \frac{\text{Cov}(W_k, T)}{\text{Var}(T)}(T - E[T])$. When $T = 1$, this becomes $P[W_k = 1 | T = 1] = E[W_k] + \frac{\text{Cov}(W_k, T)}{\text{Var}(T)}(1 - E[T]) = E[W_k] + \frac{\text{Cov}(W_k, T)}{E[T]}$. The mp-statistics used in the SNESIM algorithm are $\text{Cov}(W_k, T)$ and $E[T]$. Let $I(u, j) = \begin{cases} 1 & \text{if } (Z(u) = j, Z(u + \vec{h}_\alpha) = k_\alpha, \alpha = 1, 2, \dots, N) \\ 0 & \text{otherwise} \end{cases}$ and let $\hat{f}_k = \frac{\#\{u: I(u, k) = 1\}}{\sum_{j=1}^K \#\{u: I(u, j) = 1\}}$ where u are locations in the training

data sets. The $\hat{f}_k, k = 1, 2, \dots, K$ are estimates of the conditional probability $f_k = P[W_k = 1 | T = 1]$. The simulated value $z^s(s)$ is obtained by drawing from $\hat{f}_k$. The SNESIM algorithm proceeds by tracing a random path over the set $\{s\}$ of all locations at which observations on Z are required to be generated. When a value is generated at any location s , that location is treated as a data location in defining the data-

geometry for any subsequent location to be simulated. If in addition to Z , soft data are available, this information can also be included by taking it as part of the conditioning data. In practice a data-geometry $\tau_{n'}$ with $n' \leq n$ may be chosen so that $\hat{f}_k, k = 1, 2, \dots, K$ can be estimated in a stable manner.

Other Simulation Methods

Historically, in geosciences, simulations of random fields Z were first performed using the “Turning Bands” algorithm initially proposed by Matheron (1973). This method generates non-conditional simulations in 2D and 3D by first simulating values along a finite set of N lines spread uniformly in 2D or 3D with appropriate one-dimensional covariances $C^{(1)}\left(\left|\vec{s}\right|\right)$ corresponding to a desired isotropic covariance $C\left(\left|\vec{h}\right|\right)$ using standard techniques such as moving averaging, where $|\vec{s}|$ is the projection of $\vec{h}$ onto the lines. Let $z_i(x)$ be the value on the i^{th} line at a point closest to the point x at which simulation is desired, that is, at the foot of the perpendicular from the point x onto the line, or within a “band” around the foot. The simulated value $z(x)$ is obtained as $z(x) = \frac{1}{\sqrt{N}} \sum_{i=1}^N z_i(x)$ (Journal and

Huijbregts 1978). For simulation in 3D due to symmetry, simulation along 15 lines is sufficient to generate non-conditional simulations at desired locations (such as at grid-nodes or data-locations) with the desired covariances. Conditional simulations (i.e., conditioned to observations on Z) are obtained by using the non-conditional simulation to compute “isomorphous” errors at grid nodes and adding them to the kriged values at grid nodes as explained previously in section “Simulation of a Second-Order Stationary Process”.

Probability Field Simulation is another technique first proposed by Froidevaux (1993) that is used to generate multiple maps conditional to given observations on a random field Z . Here, it is assumed that at all points s where a simulated value is to be generated, an estimate $\hat{F}_{Z(s)}(z|N)$ of the conditional distribution of $Z(s) \mid N$, where N denotes data $z(x_1), z(x_2), \dots, z(x_N)$, is known. First, realizations $u(x)$ of random variables $U(x) \sim U(0, 1)$ are generated at each x with a given covariance function $C_U(\vec{h})$. The simulated value at s is then $z^s(s) = \hat{F}^{-1}(u(x))$. The covariance $C_U(\vec{h})$ is modeled using an appropriate transform $U = \phi(Z)$ such that $U(x) \sim U(0, 1)$ to generate $u(x_1), u(x_2), \dots, u(x_N)$ at data locations. A commonly used transform is $u(x) = \frac{r(x)}{N}$, where $r(x)$ is the rank of $z(x)$ among the N data. Although lacking in strict theoretical justification, Probability Field Simulation is used in applications for rapid generation of multiple realizations of maps.

Substitution Random Fields are also used to generate models for random functions by the composition of a directing function and a coding function. The random function Z is written as $Z(x) = Y(T(x))$, where $T(x)$ is the directing function and Y is the coding process. This enables the

generation of wide variety of stochastic processes. The structure can be used for non-conditional and conditional simulations. Details of algorithms are given in Lantuéjoul (2002).

Other stochastic approaches to simulation that emphasize the incorporation of geometric features that need to be reproduced include the Boolean models that consist of the union of independent geometric bodies (compact sets in $\mathfrak{R}^d$) of varied sizes, shapes, orientations, and internal geometry as per a specified probability distribution around random (Poisson) seeds generated in space. Boolean models form the basis for generation of random sets and are an integral part of several other models used in geostatistical simulations that take a direct approach to reproducing geometric features characteristic of a natural process. Object-based models are a kind of random genetic models wherein geological information is incorporated in simulations through generalizations of the Boolean model. The basic steps in object-based simulation consist of the following: (i) Generating points x , called germs, in the domain $D \subset \mathfrak{R}^d$. This may be done using a Poisson point-process X with a chosen intensity. (ii) An object $A(x)$ is associated with each germ x from a finite collection of such objects which are compact sets in $\mathfrak{R}^d$. To each $A(x)$ a value $V(x)$ drawn from a random process is assigned (iii) At a point $y \in \mathfrak{R}^d$, the random function Z to be simulated is assigned the value $Z(y)$ by considering all objects that contain y . Ways in which the value of $Z(y)$ can be assigned include addition or maximum of all values assigned to objects containing y or the value assigned to the last object that covers y when the Boolean objects are generated multiple times (Serra 1987, 1988; Lantuéjoul 2002). For example, in the

“dilution model,” $Z(y) = \sum_x V(x) I_{A(x)}(y)$ where $I_{A(x)}(y) = \begin{cases} 1 & \text{if } y \in A(x) \\ 0 & \text{otherwise} \end{cases}$. In the Boolean random function model $Z(y) = \max_x V(x) I_{A(x)}(y)$. In the “dead-leaves model” of Matheron (1968) and Jeulin (1980, 1989) Boolean random sets are generated multiple times and indexed by t and objects $A(x, t)$ refer to any object placed at x at time t with a corresponding value $V(x, t)$. The value $Z(y) = V(x) I_{A(x, T)}(y)$ where T corresponds to the most recent object (leaf) that contains y .

Object-based methods have been widely used for simulating fluvial reservoirs. These methods incorporate details of fluvial architecture and the hierarchy of geological processes in the simulation procedure in the form of geological rules translated into stochastic terms (Pycrz and Deutsch 2014). The steps mentioned above are modified accordingly (Haldorsen and Lake 1984; Haldorsen and Damsleth 1990; Georgsen and Omre 1993). Although object-based methods generate very realistic 2 and 3 dimensional images, they have the drawback that conditioning to data is difficult (Matheron et al. 1987). Object-based simulation techniques are also often

combined with other simulation techniques wherein internal features of objects are in-filled using pixel-based techniques.

Simulated annealing is an optimization technique where a chosen objective function is minimized proceeding through an accept-reject procedure at each step where the probabilities of acceptance of the simulation step is gradually reduced till convergence of the objective function to the accepted minimum occurs. The objective function is chosen to reflect the purpose of simulation (Pycrz and Deutsch 2014).

Concluding Remarks

The principal aim of this chapter has been to explain the theoretical foundations of the most important techniques and developments in, what may be called, “classical (two-point) geostatistics.” Methods based on higher order moments, that is, multipoint statistics are briefly reviewed. The importance of these developments lies in the fact that many of the advances in theory and aspects of their practical implementation were made in response to requirements that arose in industrial applications. The assumptions needed to make use of theoretical results may often seem restrictive when dealing with outcomes of some of the complex processes that require to be modeled. With advances in computing power, the need for making simplifying assumptions has diminished and more direct methods of learning from data have been gaining importance. Exploratory data analysis in higher dimensions and extracting information from “big-data,” the use of artificial intelligence (AI) and machine-learning (ML) are increasingly applied to capture the essence of patterns in complex processes for the purpose of modeling, prediction and decision-making in geoscience applications. Nevertheless, classical techniques provide elegant tools that can be applied very effectively whenever data-analysis shows that assumptions required for modeling are satisfied.

Cross-References

A Category

- [Big Data](#)
- [Computational Geoscience](#)
- [Geomathematics](#)
- [High-Order Spatial Stochastic Models](#)
- [Kriging](#)
- [Mathematical Morphology](#)
- [Multiple Point Statistics](#)
- [Spatial Data](#)
- [Spatial Statistics](#)
- [Statistical Computing](#)
- [Statistical Rock Physics](#)

B Category

- ▶ [Autocorrelation](#)
- ▶ [Plurigaussian Simulations](#)
- ▶ [Simulation](#)
- ▶ [Simulated Annealing](#)
- ▶ [Spatial Autocorrelation](#)
- ▶ [Spatial Statistics](#)
- ▶ [Variogram](#)

References

- Armstrong M, Matheron G (1986) Disjunctive kriging revisited: part I. *Math Geol* 18(8):711–728
- Armstrong M, Galli A, Beucher H, Le Loc'h G, Renard D, Doligez B, Eschard R, Geffroy F (2011) *Plurigaussian simulations in geosciences*. Springer, Berlin, 176p
- Arroyo D, Emery X, Peláez M (2012) An enhanced Gibbs sampler algorithm for non-conditional simulation of Gaussian random vectors. *Comput Geosci* 46:138–148
- Chiles JP, Delfiner P (2012) *Geostatistics: modelling spatial uncertainty*, 2nd edn. Wiley, New York, 699p
- Cressie N (1991) *Statistics for spatial data*. John Wiley & Sons, New York, 900p
- Delfiner P (1982) The intrinsic model of order k, Internal note no. C-97. Centre de Géostatistique, Fontainebleau, Paris, 159p
- Emery X (2007a) Simulation of geological domains using the plurigaussian model: new developments and computer programs. *Comput Geosci* 33:1189–1201
- Emery X (2007b) Using the Gibbs sampler for conditional simulation of Gaussian-based random fields. *Comput Geosci* 33:522–537
- Emery X, Arroyo D, Peláez M (2014) Simulating large Gaussian random vectors subject to inequality constraints by Gibbs sampling. *Math Geosci* 46:265–283
- Freulon X, de Fouquet C (1993) Conditioning a Gaussian model with inequalities. In: Soares A (ed) *Geostatistics Tróia'92*. Kluwer Academic, Dordrecht, pp 201–212
- Froidevaux R (1993) Probabilty field simulation. In: Soares A (ed) *Geostatistics Tróia, 92, 1*. Kluwer, Dordrecht, pp 73–84
- Galli A, Gao H (2001) Rate of convergence of the Gibbs sampler in the Gaussian case. *Math Geol* 33(6):653–678
- Galli A, Beucher H, Le Loc'h G, Doligez B (1994) The pros and cons of the truncated Gaussian method. In: Armstrong M, Dowd PA (eds) *Geostatistical simulations*. Kluwer, Dordrecht, pp 217–233
- Georgsen F, Omre H (1993) Combining fibre processes and Gaussian random functions for modelling fluvial reservoirs. In: Soares A (ed) *Geostatistics Tróia '92, 2*. Kluwer, Dordrecht, pp 425–439
- Goovaerts P (1997) *Geostatistics for Natural resources characterization*. Oxford University Press, New York, 483p
- Green AA, Berman M, Switzer P, Craig MD (1988) A transformation for ordering multispectral data in terms of image quality for noise removal. *IEEE Trans Geosci Remote Sensing* 26(1):65–74
- Guardiano F, Srivastava RM (1993) Multivariate geostatistics: beyond bivariate moments. In: Soares A (ed) *Geostatistics Tróia '92, 1*. Kluwer, Dordrecht, pp 133–144
- Guibal D, Zapua-Remacre A (1984) Local estimation of the recoverable reserves: comparing various methods with reality on a porphyry copper deposit. In: Verly G, David M, Journel AG, Marechal A (eds) *Geostatistics for natural resources characterization*. Riedel, Dordrecht, pp 435–448
- Haldorsen HH, Damsleth E (1990) Stochastic modelling. *J Pet Technol* 42:404–412
- Haldorsen HH, Lake LW (1984) A new approach to shale management in field-scale models. *SPE* 24:447–457
- Jeulin D (1980) Multi-component random models for the description of complex micro structures. *Mikroskopie* 37(s):130–137
- Jeulin D (1989) Sequential random function models. In: Armstrong M (ed) *Geostatistics*, vol 1. Kluwer Academic Publishers, Dordrecht, pp 189–200
- Journel AG (1983) Non-parametric estimation of spatial distributions. *Math Geol* 15(3):445–468
- Journel AG (1984) The place of non-parametric geostatistics. In: Verly G, David M, Journel AG, Marechal A (eds) *Geostatistics for natural resources characterisation part I*. Riedel, Dordrecht, pp 307–336
- Journel A (1999) Markov models for cross-covariances. *Math Geol* 31(8):955–964
- Journel AG, Alabert F (1988) Focussing on spatial connectivity of extreme-valued attributes: stochastic indicator-models of reservoir heterogeneities. *SPE Paper* 18324:621–632
- Journel AG, Alabert F (1989) Non-Gaussian data expansion in the earth sciences. *Terra Nova* 1:123–134
- Journel AG, Huijbregts CJ (1978) *Mining geostatistics*. Academic Press, London, 600
- Krige DG (1951) A statistical approach to some basic mine valuation problems in the Witwatersrand. *J Chem Metall Min Soc South Afr*, December:119–139
- Krige DG (1952) A statistical analysis of some borehole values in the Orange Free State goldfield. *J Chem Metall Min Soc South Afr*, September:47–64
- Lajaunie C, Lantuéjoul C (1989) Setting up the general methodology for discrete isofactorial models. In: Armstrong M (ed) *Geostatistics*, vol 1, pp 323–334
- Lantuéjoul C (2002) *Geostatistical simulation, models and algorithms*. Springer-Verlag, Berlin, 256p
- Maleki M, Emery X (2015) Joint simulation of grade and rock-type in a stratabound copper deposit. *Math Geosci* 47(4):471–496
- Marechal A (1984) Recovery estimation: a review of models and methods. In: Verly G, David M, Journel AG, Marechal A (eds) *Geostatistics for natural resources characterisation*. Riedel, Dordrecht, pp 385–420
- Mariethoz G, Caers J (2015) Multiple-point geostatistics: stochastic modeling with training images. John Wiley-Blackwell, Chichester, 376p
- Matheron G (1960) Krigeage d'un panneau rectangulaire par sa périphérie, Note géostatistique no. 28. Centre de Géostatistique, Fontainebleau, Paris
- Matheron G (1963) Principles of geostatistics. *Econ Geol* 58(8):1246–1266
- Matheron G (1968) Schema boolean sequential de partition aleatoire, Internal report N-83. Centre de Géostatistique et Morphologie Mathématique Fontainebleau, Paris
- Matheron G (1973) The intrinsic random functions and their applications. *Adv Appl Prob* 5:439–468
- Matheron G (1976a) A simple substitute for conditional expectation: disjunctive kriging. In: Gurascio M, David H, Huijbregts C (eds) *Advanced geostatistics in the mining industry*. Riedel, Dordrecht, pp 221–236
- Matheron G (1976b) Forecasting block grade distribution: the transfer functions. In: Gurascio M, David M, Huijbregts C (eds) *Advanced geostatistics in the mining industry*. Riedel, Dordrecht, pp 237–251
- Matheron G (1982) La destructurisation des hautes teneurs et le krigeage des indicatrices, Internal note N-761. Centre de Géostatistique, Fontainebleau, Paris, 33p

- Matheron G (1984a) Isofactorial models and change of support. In: Verly G, David M, Journel AG, Marechal A (eds) *Geostatistics for natural resources characterisation part-I*. Dodrecht, Riedel, pp 449–468
- Matheron G (1984b) The selectivity of the distribution and “the second principle of geostatistics”. In: Verly G, David M, Journel AG, Marechal A (eds) *Geostatistics for natural resources characterisation*. Riedel, Dodrecht, pp 421–433
- Matheron G (1984c) Une méthodologie générale pour les modèles isofactoriels discrets. *Science de la Terre, Series Inf* 21:2–64
- Matheron G (1984d) Changement de support en modèle mosaïque. In: Royer JJ (ed) *Science de la Terre, Ecole Supérieure de Géologie Appliquée et de la prospection Minière*, pp 435–454
- Matheron G (1989) Two classes of isofactorial models. In: Armstrong M (ed) *Geostatistics*, vol 1. Kluwer Academic Publishers, Dodrecht, pp 309–322
- Matheron G, Beucher H, Fouquet Ch de, Galli A, Guérillot D, Ravenne C (1987) Conditional simulation of the geometry of fluvio-deltaic reservoirs. *SPE Paper* 16753:591–599
- Minniakhmetov I, Dimitrakopoulos R (2017) Joint high-order simulation of spatially correlated variables using high-order spatial statistics. *Math Geosci* 49(1):39–66
- Myers DE (1989) To be or not to be . . . stationary ? That is the question. *Math Geol* 21(3):347–362
- Pycrz MJ, Deutsch CV (2014) *Geostatistical Reservoir Modelling*. Oxford University Press, New York, 496p
- Rivoirard J (1989) Models with orthogonal indicator residuals. In: Armstrong M (ed) *Geostatistics*, vol 1. Kluwer Academic Publishers, Dodrecht, pp 91–105
- Rivoirard J (1994) *Introduction to disjunctive kriging and non-linear geostatistics*. Clarendon Press, Oxford, 180p
- Rivoirard J, Freulon X, Demange C, Lécureuil A (2014) Kriging, indicators, and nonlinear geostatistics. *J South Afr Inst Min Metall* 114: 246–250
- Robert CP, Casella G (2004) *Monte Carlo statistical methods*, 2nd edn. Springer-Verlag, New York, 645p
- Serra J (1987) Boolean random functions. *Acta Stereologica* 6/III:325–330
- Serra J (1988) *Image analysis and mathematical morphology*, vol 2. Academic Press, London
- Strebelle S (2002) Conditional simulation of complex geological structures using multiple-point statistics. *Math Geol* 34(1):1–23
- Subramanyam A, Pandalai HS (2001) Characterization of symmetric isofactorial models. *Math Geol* 33(1):103–114
- Subramanyam A, Pandalai HS (2004) On the equivalence of the cokriging and kriging systems. *Math Geol* 36(4):507–523
- Subramanyam A, Pandalai HS (2008) Data configurations and the Cokriging system: simplification by screen effects. *Math Geosci* 40(3):425–443
- Switzer P, Green AA (1984) Min/Max autocorrelation factors for multivariate spatial imagery, Technical report no 6. Department of Statistics, Stanford University, Stanford, 14p
- Vargas-Guzmán JA, Dimitrakopoulos R (2002) Conditional simulation of random fields by successive residuals. *Math Geol* 34(5):597–611
- Vargas-Guzmán JA, Dimitrakopoulos R (2003) Computational properties of min/max autocorrelation factors. *Comput Geosci* 29:715–723
- Verly G (1983) The multigaussian approach and its applications to the estimation of local recoveries. *Math Geol* 15(2):263–290
- Wackernagel H (2003) *Multivariate geostatistics, an introduction with applications*, 3rd edn. Springer-Verlag, Berlin, 387p
- Yao L, Dimitrakopoulos R, Gamache M (2020) High-order sequential simulation via statistical learning in reproducing kernel Hilbert-space. *Math Geosci* 52(5):693–724

GeoSyntax

Cedric M. Griffiths

StrataMod Pty. Ltd., and Curtin University, Perth, WA, Australia

Definition Consideration of Gressly’s original facies concept, and the formal application of “Walthers Law” suggested the basis of a succession analysis tool that could be used to: (a) match or correlate adjacent sections or wells via “Syntactic Pattern Recognition,” (b) act as the basis of quantitative hypothesis testing in stratigraphic evolution, (c) enable forward simulation of stratigraphic successions that test current understanding of facies association.

“GeoSyntax” is based on the development of a grammar of unit associations, and the parsing of lithological units defined either by hand specimen description, wire-line log characteristics, or theoretical facies associations.

Background

The philosophical basis for the geoSyntax approach to machine-assisted interpretation was first published by Griffiths (1989, 1990). Further work by Duan (1996 to 2001) expanded this to two-dimensional sections, and Hill (2007 to 2009) expanded the concept further to conditioned two-dimensional channel map forms and three-dimensional volumes using 3D-plex grammars.

Because of the human psychological necessity to reduce information to “chunks” for efficient processing (Griffiths 1989, 1990) as described by Miller (1956), classification forms an essential role in the observational sciences, and yet the theoretical process of classification itself has not been subjected to much debate in the geological literature, despite playing a central role in geological interpretation. The language of geological rules, the lingua franca of geological discourse, is composed largely of classes derived from subdivisions of observations. As an example, take this description of a canyon-fill deposit: “localised sand-prone intervals, possibly sand filled leveed feeder channels . . . local areas of high amplitude, internally mounded, discontinuous reflections.”

This description contains at least eight classifications (“local,” “sand,” “possible,” “levee,” “feeder,” “channel,” “high,” “mounded,” “discontinuous”), and a mixture of genetic, descriptive, fuzzy, and “hard” classes. A machine-based interpreter would have severe problems with many of these classifications, as would a geologist attempting to put numbers on these characteristics for quantitative testing of hypotheses.

Essential Concepts of the GeoSyntax Approach

A formal syntactic system has been developed for combining such qualitative/semiquantitative classes without the need for rigorous numerical treatment. The need to provide fast and efficient compilers for computer languages led to various ways of representing this structure, or “syntax.” One widely used syntactic system uses the “Backus- Naur-Form” (BNF) notation which was first developed to define the syntax of Algol 60 (Chomsky 1957). In the fields of stratigraphy and basin analysis there are a few stated principles such as:

- (i) Walther's Law of Succession of Facies (Walther 1893): which states that in a succession with no breaks (i.e., faults or erosional unconformities) only those facies that are to be found beside each other at the present time can be superimposed in vertical section.
- (ii) Law of Superposition of Strata: which states that in any succession of strata the beds were deposited in a simple vertical order from base to top.
- (iii) Gressly's Facies concept (Gressly 1838): which states that lateral changes in physical characteristics of rocks at a mappable scale are identified as discrete units.

The facies concept has been subjected to various interpretations, but Gressley's original sense, of being a collection of physical properties that are distinguishable from those of neighboring rocks at a chosen scale, is suitable for GeoSyntax purposes (Moore 1949). Walther's Law has potential as a predictive tool if used carefully, and may be used to form the basis of a syntactic rule set that translates vertical well or outcrop observations to three-dimensional facies volumes. For any given depositional system at a given scale, there will be a set of facies relationships such that a traverse in a given direction in relation to the direction of flow of the transport medium would be expected to encounter a succession of descriptive facies units with an estimated probability. In stratigraphy we thus have a set of terminal symbols (facies) that, if they are capable of being unambiguously mapped within a present-day depositional system, should also be capable of being identified, using the same syntax, from borehole successions if there were no intervening periods of erosion.

```

<Upper.channel> ::= <Bar.top> | <VA>
<VA> ::= f | <VA> g
<Bar.top> ::= f | e | d
<Lower.channel> ::= <Channel.floor> |
<Lower.channel> c | <Lower.channel> b
<Channel.floor> ::= ss | a |
<Channel.floor> a
<Channel.fill> ::= <Lower.channel>
<Upper.channel> | <Channel.fill> <Upper.
channel>

```

```

<Channel.succession> ::= <Channel.fill>
<Channel.fill> | <Channel.succession>
<Channel.fill>
<Channel.succession> ::= <Channel.
succession> <Upper.channel>

```

where: ss = scoured surface, a = poorly defined trough cross-bedding, b = well-defined trough cross-bedding, c = large planar tabular cross beds, d = small planar tabular cross beds, e = isolated scour fills, f = ripple cross laminated silts and muds, g = low angle inclined stratification, VA = vertical accretion unit.

The simple one-dimensional Battery Point channel fill model (Fig. 1), which is itself derived from a Markov analysis of several field sections, can provide an example of a simple syntax derived from theoretical considerations. [GeoSyntax, Fig. 1.tif].

Examination of the “summary” succession might lead to the design of the following context-free BNF syntax (from the base upwards).

Each BNF production rule contains:

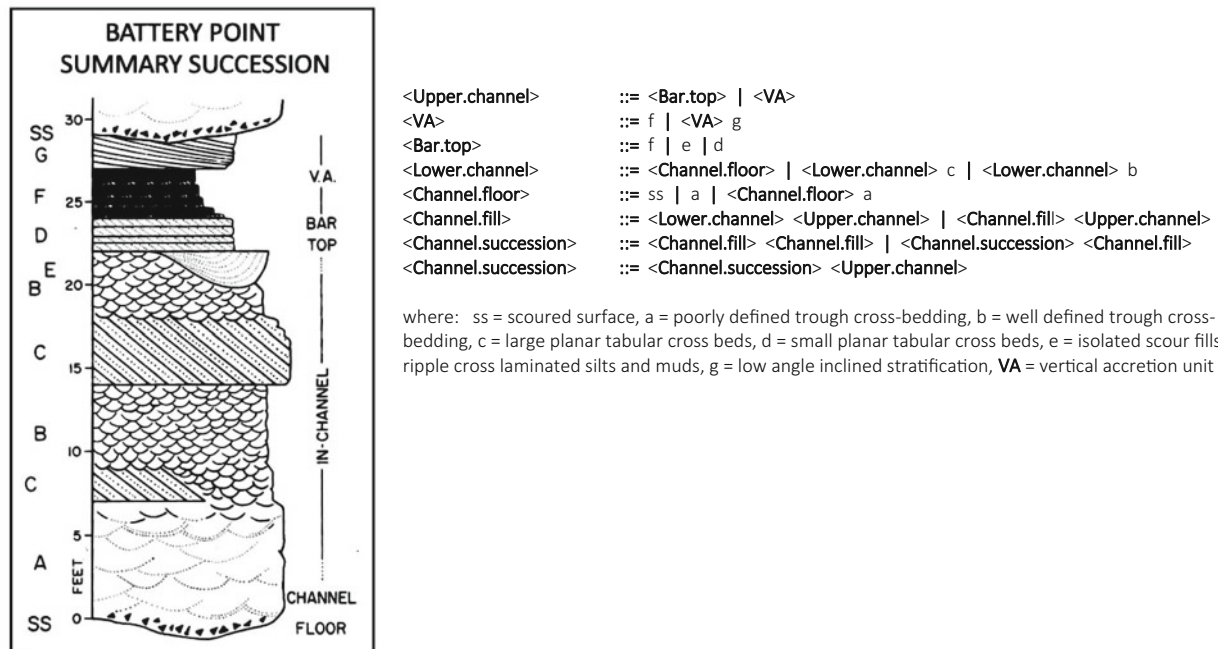
a symbol “::=”, which means “is comprised of”.

a nonterminal symbol (symbol to the left of::=).

one or more terminal symbols (symbols to the right of::=).

Terminal symbols are the elementary symbols of the language defined by a formal grammar. Nonterminal symbols are replaced by groups of terminal symbols according to the production rules. The terminals and nonterminals of a particular grammar are two disjoint sets. Where there is more than one terminal symbol these are separated by the “or” symbol (|), or a space which is interpreted as “followed by.” This implies that the succession is directional and read from base to top in an outcrop or borehole, or (after Walther's Law) from proximal to distal.

The simple grammar and syntax described above has captured the essence of a “Channel Fill” succession as understood by Cant and Walker. This understanding can be used in several ways: (a) to simply and formally define the essence of “Channel Fill” and its possible variability in a way that can be understood by both human and machine interpreters; (b) to enable multiple synthetic “Channel Fill” units to be used to forward model channel successions; (c) to enable an unknown section to be tested automatically for degree of “Channel Fill”-ness. Ranges of physical properties and probabilities (such as thickness, lateral extent, porosity, and permeability) may be attributed to each of the terminal symbols as described in Duan et al. (1999) for generating probabilistic 2D stratigraphic sections. To produce a synthetic succession the highest level nonterminal symbol, in this case “Channel.Succession,” must be completely replaced by terminal symbols using the grammatical rules. Terminal symbols represent basic geological observations: scoured surface, poorly defined trough cross-bedding, well-defined trough cross-



GeoSyntax, Fig. 1 Battery point summary channel fill succession. (Derived from Walker and Cant 1984, Fig. 20, p. 83)

bedding, large planar tabular cross beds, small planar tabular cross beds, isolated scour fills, ripple cross-laminated silts and muds, and low angle inclined stratification. By convention, names of terminal and nonterminal symbols start with lower-case and upper-case letters, respectively.

Formally speaking, this is an inverted BNF grammar. It is presented this way for two reasons. The first is that computationally it is more efficient, and secondly it is easier to follow the progressive interpretation of nonterminal to terminal symbols while reading. It is worth noting that grain size is included explicitly in only one terminal symbol (f). The remainder are a mixture of genetic and descriptive structural/textural facies. When reconstructing a sedimentary body from limited observational data, geologists will typically produce the simplest model that honors the available data.

The geosyntactic method is not only object-based but it is also object-oriented. Object-oriented means that a geological object can be represented by a single symbol of variable size rather than a collection of rectangular cells, and hence the simulation does not have a fixed resolution dependent on cell size. The syntactic method provides a holistic framework to describe a complete depositional environment in a single grammar and has the potential to describe sophisticated patterns in a simple format. Once the user has selected or designed a suitable grammar the generation of geological models is performed automatically by a computer program

called the parser. Parsing is the process of applying the legal components of a grammar according to the rules described in the syntax to either interpret or generate “legal” successions of terminal symbols. In a stratigraphic context the grammar is the set of rules specifying the sedimentary units that are associated with a specific environment and their relationships as defined by the syntax. The basic principles of geosyntax as outlined above may be expanded in application and complexity to address the simulation of two- and three-dimensional sediment bodies.

Current Status

GeoSyntax allows for the concise definition of complex facies relationships, and enables their quantitative regional comparison by combining an attributed grammar with a “Generalised Levenshtein Distance.” In this case, each of the facies symbols contains one or more attributes, (a “string with attributes” in symbolic computation language). For instance, the bottom to top facies type of a typical channel sedimentary facies succession can be coded by symbols as shown in the Channel Fill case above, and each facies (terminal symbol) can be associated with a number or an attribute representing the thickness of each facies as described in detail by Duan et al. (2001a, b; Duan 2017) combining an attributed grammar, with a “Generalised Levenshtein Distance.”

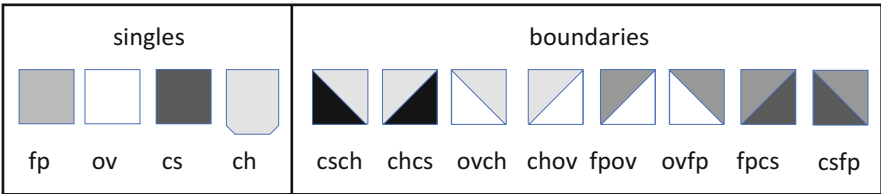
Extension to Two- and Three-Dimensional Sedimentary Bodies

The above examples used “facies” as terminal symbols (e.g., “small planar tabular cross beds”), and environments (e.g., “Upper.channel”) as nonterminal symbols. However, it proved also possible to use geometric shapes as symbols in a

grammar describing channel behavior in two-dimensional vertical section as shown in Figs. 2 and 3, and two-dimensional plan view as shown in Fig. 4.

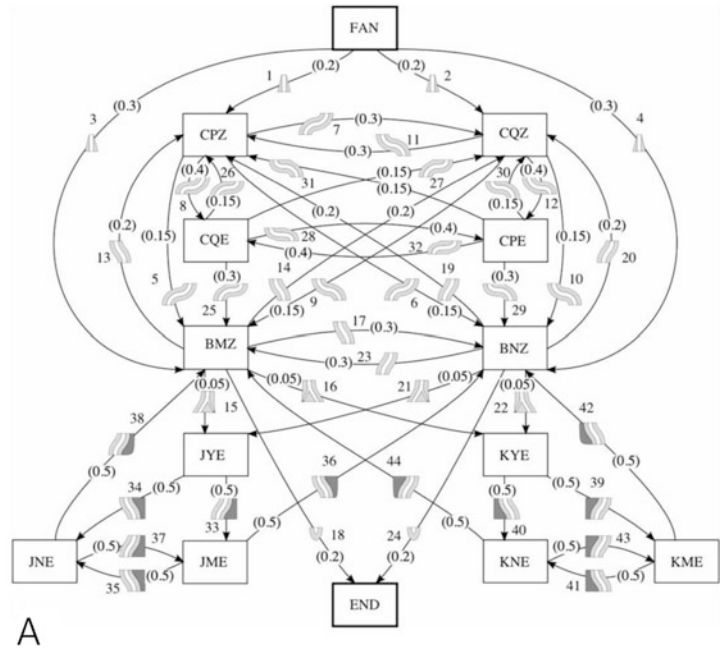
A set of probabilistic rules for a meandering channel system. Nonterminal symbols are written in capitals, MDCH is the start symbol (Table 1).

GeoSyntax, Fig. 2 Basic shape of terminal symbols: *fp* floodplain, *ov* overbank, *cs* crevasse splay and *ch* channel fill deposits. (Derived from Hill and Griffiths 2006))



GeoSyntax, Fig. 3 Two-dimensional simulated vertical section based on the syntax shown in Fig. 2, conditioned on 2 hypothetical wells (black lines): 10 time slices using a grammar which generates nested channels.

Dark gray = floodplain muds, light gray = channel fill, black = crevasse splay deposit, white = overbank deposits. (Derived from Hill and Griffiths 2006))



GeoSyntax, Fig. 4 Two-dimensional plan-view conditioned geosyntax simulation. (Derived from Hill and Griffiths 2009)
A – A finite-state diagram illustrating the rules which generate parent symbols. The states are outlined in boxes and have 3 letter labels, the start state for the grammar is FAN. The rules are represented by arrows which go from initial state to final state. Each rule has (a) a number, (b) the probability of the rule in brackets, and (c) a simple representation

of the parent symbol generated by that rule.
B – Examples of maps of channel-related deposits generated by the parser: (a) an unconditioned map; (b) and (c) maps generated using the same two conditioning symbols. Light gray 1/4 overbank deposits; dark gray 1/4 splay deposits; white 1/4 channel fill; black 1/4 channel conditioning symbol.

GeoSyntax, Table 1 A set of rules for a meandering channel system (MDCH). Terminal symbols (facies) start with small letters and non-terminal symbols (interpretation) are in capital letters. MDCH is the start symbol for a forward simulation, or the final interpretation for an inverse model

No.	Head		Tail(s)	Probability of the first tail
1	MDCH	→	fp fp + MD	1.0
2	MD	→	MDL MDR	0.5
3	MDR	→	fp + fpOv + ov + ovCH + chR + CHFP	1.0
4	MDL	→	FPCH + chL + chOv + ovR + fp + MDCH	1.0
5	CHFP	→	CSFP OVFP	0.7
6	FPCH	→	FPCS FPOV	0.7
7	CSFP	→	chCs + cs + CSFP + fp + MDCH	1.0
8	OVFP	→	chOv + ov + OVFP + fp + MDCH	1.0
9	FPCS	→	fp + FPCS + cs + csCh	1.0
10	FPOV	→	fp + FPOV + ov + ovCh	1.0

Summary

GeoSyntax as a novel geological concept and technology evolved from the work of Griffiths (1989, 1990); Griffiths et al. 1992) and Duan et al. (1996, 1999, 2001a, b) who successfully demonstrated 1D syntactic interpretations of logs and generation of conditional simulations of simple 2D parasequences using syntactic techniques. This initial work further evolved at CSIRO (Hill and Griffiths 2006, 2007, 2008, 2009) and Sinopec (Duan 2017) to the point where it can be used for 3D applications in both forward and inverse modelling.

The computer programs and reports resulting from the CSIRO research can be accessed from the following websites.

Hill, June (2014): GeoSyntax Project and Software. v1.

CSIRO: <https://doi.org/10.4225/08/543C9519616FB>

Hill, June (2014): GeoSyntax Java Code. v1. CSIRO: <https://doi.org/10.4225/08/543CA14A6A2C7>

Bibliography

- Chomsky N (1957) Syntactical structures. Mouton, The Hague
- Duan T (2017) Similarity measure of sedimentary successions and its application in inverse stratigraphic modeling. *Pet Sci* 14:484–492
- Duan T, Griffiths CM, Johnsen SO (1996) Syntactic simulation of 1-D sedimentary sequences in a coal-bearing succession using stochastic context-free grammars. *J Sed Research* 66(6):1091–1101
- Duan T, Griffiths CM, Johnsen SO (1999) A new approach to reservoir heterogeneity modelling: conditional simulation of 2-D parasequences in shallow marine depositional systems using an attributed controlled grammar. *Comput Geosci* 25(6):667–681. [https://doi.org/10.1016/S0098-3004\(98\)00162-9](https://doi.org/10.1016/S0098-3004(98)00162-9)
- Duan T, Griffiths CM, Johnsen SO (2001a) High frequency sequence stratigraphy using syntactic methods and clustering applied to the upper limestone coal group (Pendleian, E1) of the Kincardine Basin, United Kingdom. *Math Geol* 33(7):825–884
- Duan T, Cross TA, Lessenger MA (2001b) Reservoir- and exploration-scale stratigraphic prediction using a 3-D inverse carbonate model. In: AAPG annual convention, Denver, Colorado
- Gressly A (1838) Observations géologiques sur le Jura Soleurois. *Neue Denkschr Allg Schweiz Ges Gesamten Naturwiss* 2:1–112
- Griffiths CM (1989) The nature of the geological representation language and consequent constraints on machine interpretation - preliminary discussion and suggestions. In: Simaan M, Aminzadeh F (eds) *Advances in geophysical data processing Vol. 3, artificial intelligence and expert Systems in Petroleum Exploration*. JAI Press Inc., Connecticut
- Griffiths CM (1990) The language of rocks - an example of the use of syntactic analysis in the interpretation of sedimentary environments from wireline logs. In: Hurst A, Lovell MA, Morton AC (eds) *Geological applications of wireline logs*. Geological Society of London
- Griffiths CM, Kopaska-Merkel, DC, Schott M (1992) Sedimentary sequences influenced by submarine fan deposition: Argo Abyssal Plain, northwestern Australia. *Proceedings of the Ocean Drilling Program, Scientific Results* 123:601–623
- Hill EJ, Griffiths CM (2006) A stochastic grammar and a predictive parser for the description and prediction of sedimentary successions. *Society for Mathematical Geology XIth international congress*. Université de Liège, Belgium
- Hill EJ, Griffiths CM (2007) Simulating sedimentary successions using syntactic pattern recognition techniques. *Math Geol* 39(2):141–157. <https://doi.org/10.1007/s11004-006-9074-4>
- Hill EJ, Griffiths CM (2008) Formal description of sedimentary architecture of analog models for use in 2D reservoir simulation. *J Mar Petr Geol* 25:131–141. <https://doi.org/10.1016/j.marpetgeo.2007.05.001>
- Hill EJ, Griffiths CM (2009) Describing and generating facies models for reservoir characterisation: 2D map view. *J Marine Petr Geol* 26(8): 1554–1563
- Miller GA (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol Rev* 63(2): 81–97
- Moore RC (1949) Meaning of facies. *Mem Geol Soc Am* 39:1–34
- Walker RG, Cant DJ (1984) *Facies Models*, ed. RG Walker, Repr. Ser. 1.: Toronto: Geosci. Can. 2nd ed.
- Walther J (1893) “Einleitung in die Geologie als Historische Wissenschaft,” vol I. Fischer, Jena, 196pp

Geotechnics

Pejman Tahmasebi

College of Engineering and Applied Science, University of Wyoming, Laramie, WY, USA

Definition

Geotechnics is the science of application of applied geology and mechanics for the problems related to the earth's surface or near-surface. As such, their target materials are categorized into two groups: soil and rock. Soil is considered as loose materials and sediments. Also, soil can be broken into pieces easily. Unlike soil, rock materials are very rigid materials, and the strong bond between grains can be found in the cohesion and molecular forces. The science generated in geotechnics can be useful in many natural phenomena and construction of facilities on the earth and beneath it.

Introduction

Geotechnics deals with geomaterials located on/near the surface of the earth. Geotechnics has influenced on several areas, such as rock mechanics which describes the properties and mechanics of rocks, soil mechanics that is concerned with soil's dynamic behavior, mechanics of materials, and kinematics of the soil. One often studies the soils and rocks since their behaviors are very different. Furthermore, all the materials that exist on the surface are not entirely soft or hard. In fact, most of the geomaterials can be categorized somewhere between soft and hard materials. Foundation science and engineering is another branch of geotechnics that deploys geology, rock mechanics, soil mechanics, and structural engineering to secure a structure from failure. The knowledge in geotechnics has also been used extensively to model large foundations to avoid the collapse of structures. In recent years, researchers have also considered environmental problems as well. This special attention has made new fields called geo-environmental engineering where the issues related to geotechnical engineering that affect the environments, such as groundwater quality, waste of industries disposal, nuclear waste disposal, etc. are studied. In summary, geotechnics is a combination of physics, mathematics, geology, environmental sciences, hydraulic, and mining engineering, to predict the behavior of soils or rocks.

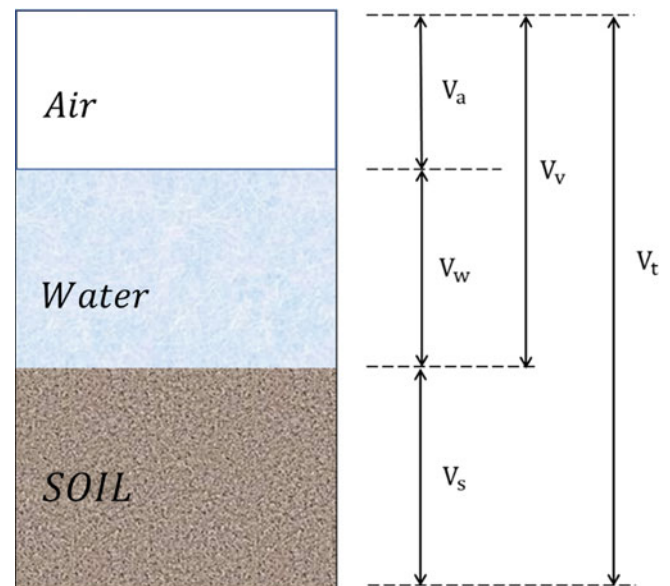
All the ancient civilizations built many facilities on the surface or even underground. Coulomb, which is named as the leader of soil mechanics, was the first famous engineer that studied the stability of walls when they are under pressure. Coulomb's essay is the earliest soil mechanics study

published in 1776 by the Academy of Science. Collin was the first investigator that studies the stability of clay slope and the shear strength of clays, in 1846. Many researchers such as Darcy (1856), Rankine (1857), and Gregory worked in this field in the late eighteenth century. In the nineteenth century, the Swedish researcher, Atterberg, introduced the limits for clay soil, which is currently used by many engineers. In this period of time, the Swedish State Railways developed methods that can be used to calculate the stability of slopes. Moreover, they studied the subsurface techniques for soil sampling. In 1925, the German Professor, Karl Terzaghi who is named as the pioneer of soil mechanics, published his and the world's first textbook on soil mechanics (Terzaghi 1925). He wrote almost 250 technical reports and papers on soil mechanics. Professor Arthur Casagrande was a Harvard University professor between 1932 and 1969 who had an essential role in expanding and developing the science of geotechnics (Casagrande 1936a, b, 1948, 1965).

Geomaterials mostly contain pore spaces. As such, the total volume of a soil mass (V_t) consist of solid (V_s) and void (V_v). The void space can be filled by water (V_w) and/or air (V_a). This concept is shown in Fig. 1. Given the above parameters, the void ratio (e) is expressed as:

$$e = \frac{V_v}{V_s}. \quad (1)$$

The void ratio can vary from 0 to ∞ . For instance, the typical value for void ratio in sand samples is between 0.4 and 1.0. Similarly, the porosity (n) can be defined as:



Geotechnics, Fig. 1 Volumetric relationship for a soil using a phase diagram

$$n = \frac{V_v}{V_t} \times 100. \quad (2)$$

From Eqs. (1) and (2), one can write:

$$n = \frac{e}{1 + e} \quad (3)$$

$$e = \frac{n}{1 - n} \quad (4)$$

Also, saturation (S) can also be calculated using:

$$S = \frac{V_w}{V_v} \times 100. \quad (5)$$

In geotechnics, it is very common to use the unit weight (γ) instead of density g :

$$\gamma = \frac{W}{V} = \frac{\rho V g}{V} = \rho g \text{ [N/m}^3\text{]} \quad (6)$$

In small-scale problems where the soil particles are studied, the size of such instances is very important and can be used in computational modeling as input. Some of the available classifications are shown in Table 1. The behavior of the clay soil, for example, can be different based on the amount of water in the soil. When the amount of water increases, clay experiences four different stages, solid, semi-solid, plastic, and liquid. For each of the abovementioned stage, the behavior and general properties of the soil can be changed. Therefore, it is necessary to identify in which stage the clay and silt soil are found. This allows having an accurate estimation of its strength, other mechanical properties. To find the state of the soil, the Atterberg limit test can accurately characterize the stage that the soil is in White (1949). This concept is shown in Fig. 2.

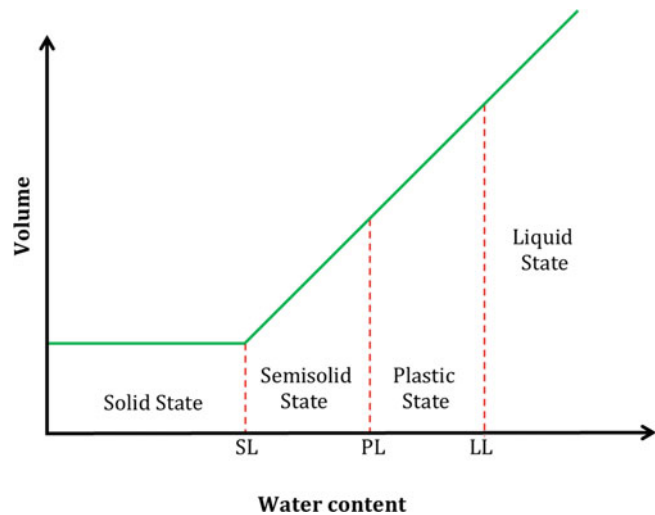
As demonstrated in Fig. 2, the loss of water content will not cause a volume change before the SL level. By increasing the amount of water, however, the soil shifts from solid state to plastic level, which is called semisolid state. When the water content reaches the liquid limit, the soil changes the state from plastic to a liquid state. Using this information, one can calculate the Atterberg soil index, or the *plasticity index* (PI), by:

$$PI = LL - PL \quad (7)$$

The PI is the range between the plastic limit and liquid limit. If the PI is high, the soil has a high content of clay. The liquidity index (LI) is another index which can characterize the soil's consistency and is defined by:

Geotechnics, Table 1 Well-known particle size classifications

Organization	Particle Size (mm)			
	Clay	Silt	Sand	Gravel
AASHTO	<0.002	0.002–0.075	0.075–2	2–76.2
USDA (Minasny and McBratney 2001)	<0.002	0.002–0.05	0.05–2	>2
MIT	<0.002	0.002–0.06	0.06–2	>2



Geotechnics, Fig. 2 Demonstration the various states of soil based on the Atterberg limit. Here, the SL, PL, and LL represent shrinkage limit, plastic Limit, and liquid limit, respectively

$$PI = \frac{PL - W}{PI}. \quad (8)$$

In Eq. (8), W is the water content in the soil sample. The LI range is shown in Fig. 3. As can be seen, the state of the solid can be characterized by the value of the LL.

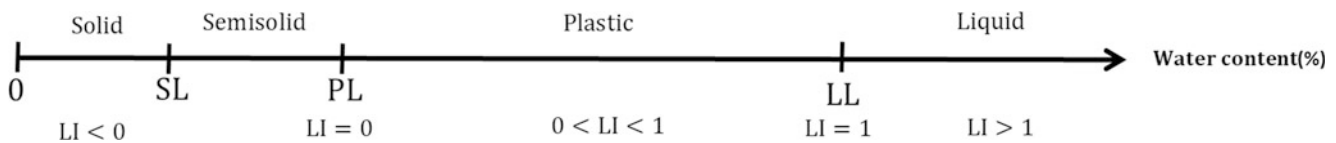
Another fundamental property of soil is their shear strength (S), which is expressed as:

$$S = \sigma \tan \theta + C, \quad (9)$$

where σ is the normal stress, θ is the friction angle of the soil, and C is the cohesion.

Research Path of Geotechnics

In recent years, environmental aspects are added to geotechnical application, which are most around water-soil interactions. However, there is still a gap in the geo-environmental numerical simulations. The computational power has increased in recent years, and it resulted in considering the effect of fluid to such modeling. The computational fluid



Geotechnics, Fig. 3 Various states of soil based on water content and LI

dynamic (CFD) is one of such methods that can help to simulate the fluid phase in geotechnical systems. For instance, groundwater pollution, which is one of the noticeable issues in geo-environmental science (Lopes and Quinta-Ferreira 2010; Wu et al. 2020), can be studied in this framework. Furthermore, to have more accurate simulations at the grain-scale, many investigators have applied the Lagrangian methods such as discrete element method (DEM) in geotechnical systems (Jiang et al. 2003; Ngo et al. 2014) wherein the particles are presented using spherical objects. Besides, these methods are coupled and a set of new CFD-DEM coupling can be used to represent both the solid and fluid phases, which can result in a more representative simulation (Guo and Yu 2017). This field of research still needs more investigation to better represent the morphology of particles and also provide a more feasible simulation when all such complexities are considered. More complex multiphysics simulation, such as those known as THMC (thermo-hydro-mechanical-chemical), can be developed for both small- and large-scale systems (Liu et al. 2014).

Conclusion

Geotechnics focuses on studying the mechanics of soils and rocks. The fundamental knowledge of this area was discussed. Due to the length limitation for this entry, all the necessary materials were not covered, but we restricted the topics to those through which one can study the behavior of soils. More interdisciplinary studies have been and will be conducted, and this topic can benefit various other fields, such as structural engineering, water resources engineering, hydraulic structures engineering, environmental engineering, mining engineering, etc.

Bibliography

- Casagrande A (1936a) Characteristics of cohesionless soils affecting the stability of slopes and earth fills. *J Boston Soc Civil Eng* 23(1):13–32
- Casagrande A (1936b) The determination of pre-consolidation load and its practical significance. *Proc Int Conf Soil Mech Found Eng Cambridge, Mass* 3:60
- Casagrande A (1948) Classification and identification of soils. *Trans Am Soc Civ Eng* 113(1):901–930
- Casagrande A (1965) Role of the calculated risk in earthwork and foundation engineering. *J Soil Mech Found Div* 91(4):1–40

- Guo Y, Yu XB (2017) Comparison of the implementation of three common types of coupled CFD-DEM model for simulating soil surface erosion. *Int J Multiphase Flow* 91:89–100
- Jiang MJ, Konrad JM, Leroueil S (2003) An efficient technique for generating homogeneous specimens for DEM studies. *Comput Geotech* 30(7):579–597
- Liu Y, Ma L, Ke D, Cao S, Xie J, Zhao X, Chen L, Zhang P (2014) Design and validation of the THMC China-Mock-Up test on buffer material for HLW disposal. *J Rock Mech Geotech Eng* 6(2):119–125
- Lopes RJG, Quinta-Ferreira RM (2010) Evaluation of multiphase CFD models in gas–liquid packed-bed reactors for water pollution abatement. *Chem Eng Sci* 65(1):291–297
- Minasny B, McBratney AB (2001) The Australian soil texture boomerang: a comparison of the Australian and USDA/FAO soil particle-size classification systems. *Soil Res* 39(6):1443–1451
- Ngo NT, Indraratna B, Rujikiatkamjorn C (2014) DEM simulation of the behaviour of geogrid stabilised ballast fouled with coal. *Comput Geotech* 55:224–231
- Terzaghi K (1925) *Erdbaumechnik auf bodenphysikalischer Grundlage*. F. Deuticke
- White WA (1949) Atterberg plastic limits of clay minerals. *Am Mineral: J Earth Planet Mater* 34(7–8):508–512
- Wu J, Liu Z, Yuan S, Cai J, Hu X (2020) Source term estimation of natural gas leakage in utility tunnel by combining CFD and Bayesian inference method. *J Loss Prev Process Ind* 68:104328

Geothermal Energy

Katsuaki Koike and Shohei Albert Tomita
Department of Urban Management, Graduate School of Engineering, Kyoto University, Kyoto, Japan

Definition

The Earth contains a large amount of heat energy. The planet's interior temperature has been shown to strongly and linearly increase with depth at shallow depths, exceeding 1000 °C at the lithosphere-asthenosphere boundary (~100-km depth) following a general temperature gradient of 20–30 °C/km, and gradually increase in the asthenosphere, mesosphere, and core. The Earth's interior heat is transferred to the surface with a mean surface heat flux of 87 mW/m², which mostly (~80%) originates from the decay of radioactive isotopes (mostly ²³⁵U, ²³⁸U, ²³²Th, and ⁴⁰K), and the remainder is mainly from primordial cooling (Turcotte and Schubert 2014). This heat production drives mantle convection and resultantly plate motion.

The heat in the shallow crust within several kilometers depth can be used as geothermal energy for electric power generation, agriculture, aquiculture, and recreation. The importance of geothermal energy has recently increased in terms of renewable energy-based power generation owing to its high power output, high operation rate, and low CO₂ emissions. A power-generating turbine is rotated by high-pressure steam that originates from a hydrothermal fluid, which is a mixture of water and gas originated generally from meteoric water heated by an interior heat source (e.g., deep magma chamber) or injection water through hot rocks. These fluids are stored in reservoirs that are generally formed by permeable and continuous fractures under a caprock layer. Heat sources, fluids, and fractures are therefore fundamental elements within high-potential geothermal energy zones.

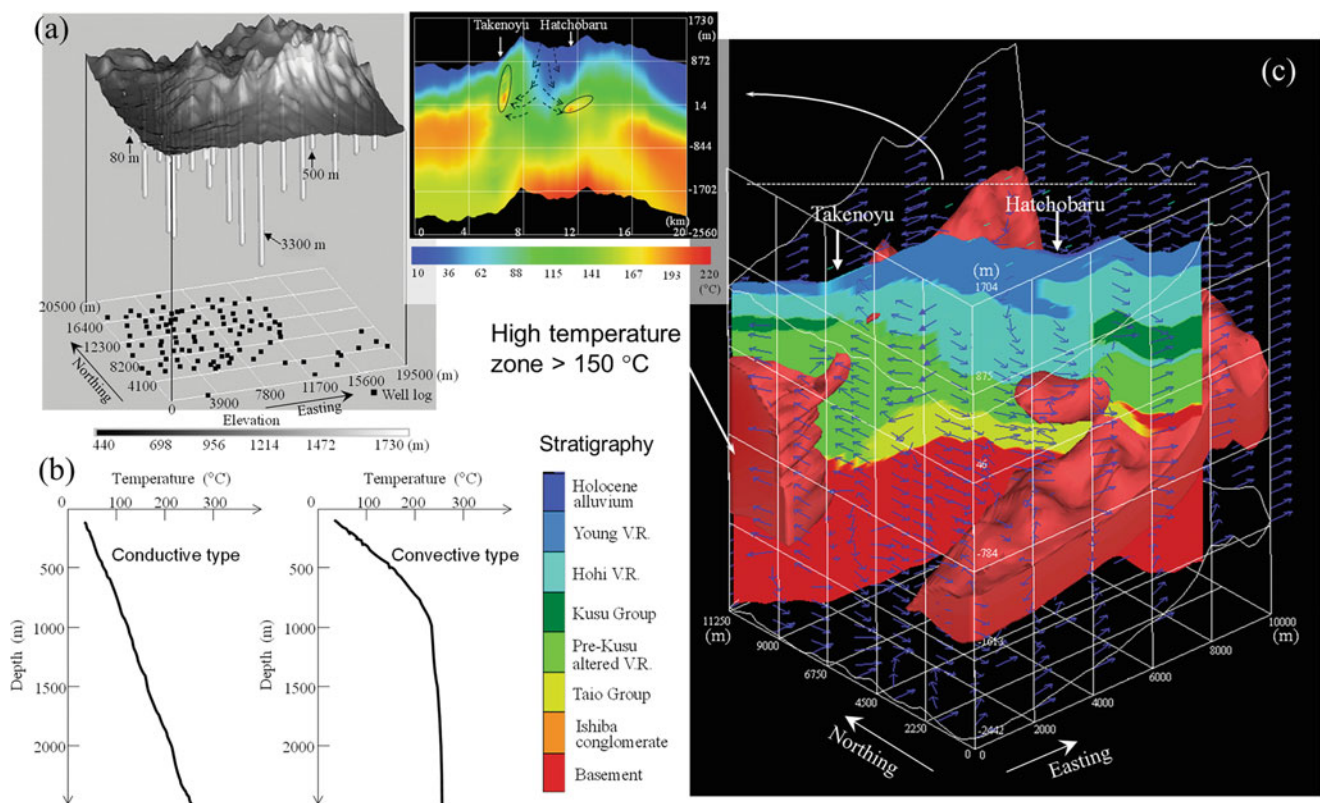
The main properties of a fluid include temperature, pressure, and flow pattern, which can be estimated using numerical simulations by adopting an optimal hydrogeologic model. The subsurface and surface temperature distribution is an essential property because it provides basic information regarding these three elements. Fluid path fractures and heat sources can be detected using remote sensing technology to

measure geologic structural and mineralogical features and surface displacement. Mathematical geologic methods have been applied to improve the modeling accuracy of fluid flow, temperature distribution, fracture systems, and regional geothermal systems caused by a heat source. An example of subsurface temperature modeling is outlined in Fig. 1.

Subsurface Temperature Modeling

Application of Geostatistics

The three-dimensional temperature distribution is the most fundamental piece of information for specifying large geothermal energy zones. The temperature distribution can be estimated or simulated by applying geostatistics using well-logging or bottom-hole temperature data, which are obtained in geothermal or petroleum/gas resources exploration operations and recognized as a regionalized variable. One precaution that should be considered when using this spatial modeling approach is the physics that controls the temperature-change pattern with depth. The patterns are generally classified into two types depending on whether the heat



Geothermal Energy, Fig. 1 Example of temperature modeling for the Hoho geothermal field in southwestern Japan using well logging data from different bottom depths. (Modified from Teng and Koike (2007)). (a) Topography and well location with lengths of 80–3300 m. (b) Schematic temperature profiles of conductive and convective thermal types. (c) Temperature model integration showing high-temperature

zones (>150 °C), a geologic model composed of eight formations, and large-magnitude fluid flow vectors for imaging the geothermal system. A NW-SE vertical cross-section passing through two hot springs along the dashed line in (c) schematically shows the downflows of meteoric waters and shallow oval-shaped reservoirs

is transmitted via conduction or transferred via fluid convection. A conductive pattern involves a linear increase of temperature with depth, whereas a convective pattern bears an inflection point that separates the linear increase with depth at shallow depths and the nearly constant temperature at greater depths. These temperature-depth trends result in temperature datasets to be non-stationary variables. Another precaution to consider is that the temperature data are quite sparse at great depth, and suitable measures are therefore required to extrapolate shallow data for temperature modeling at reservoir depths or deeper.

One effective measure is to approximate the temperature-variation pattern using a polynomial and kriging with external drift is then applied by regarding the polynomial as a trend. A global linear trend (i.e., constant geothermal gradient) is applicable in cases when most of the data follow a conductive pattern type over a sufficiently wide area (e.g., hundreds of thousands of km²) (Agemar et al. 2012; Tian et al. 2015). Convective patterns dominate the well-logging data of geothermal fields with high-output power stations. One example is the world-famous, Wairakei geothermal field in New Zealand in which the trend was approximated well by a third-order polynomial (Sepúlveda et al. 2012). Previous studies have demonstrated the effectiveness of incorporating this trend because the increased temperatures and hot zones are estimated in deeper zones beyond the well bottoms.

In addition to temperature data, geologic column data are also useful for constructing geologic models, typically by transforming column data into binary data (i.e., 1 for a selected rock type; 0 for other types). The spatial modeling of these binary datasets provides information regarding the rock type that is most probable to occur at a given location, and can thus be applied to construct geologic models. The superimposition of the estimated temperature distribution on a constructed geologic model enables the interpretation of fault structures, fault permeability, main fluid flow paths, flow patterns (up, down, or lateral flow), and the location, configuration, and rock type of a reservoir with high temperatures and dominant upflows, as demonstrated for the Wairakei field (Sepulveda et al. 2012) and Hoho field in Japan (Teng and Koike 2007).

Application of Machine Learning

Machine learning techniques are also applicable to logging data to obtain the temperature distribution. A deep neural network (DNN) is a typical method that uses the relationship between the data location and measured temperature. One measure to evaluate the DNN capability is the extrapolation accuracy of logging data because deep logging data are generally limited. Network learning is repeated until the total difference between the network outputs and data values are acceptably small, which is the chief factor of the learning

criterion. Multiple factors can also be incorporated into the learning process as constraints. A thermal-type semi-variogram is one example that assigns 0 or 1 to a drilling site of either conductive- or convective-type patterns and produces a binary dataset. This constraint can be used to locate high-temperature zones and derive a convection fluid pattern by honoring the spatial correlation of the thermal type (Koike et al. 2001). Other methods, such as XGBoost and Random Forest, are also powerful and accurate tools for temperature modeling. These methods can predict temperatures at great depths by extrapolation and generate geothermal gradient maps to specify high potential resource zones with a large gradient using a dataset of bottom-hole temperatures (Shahdi et al. 2021).

Remote Sensing Applications

Outline of Remote Sensing

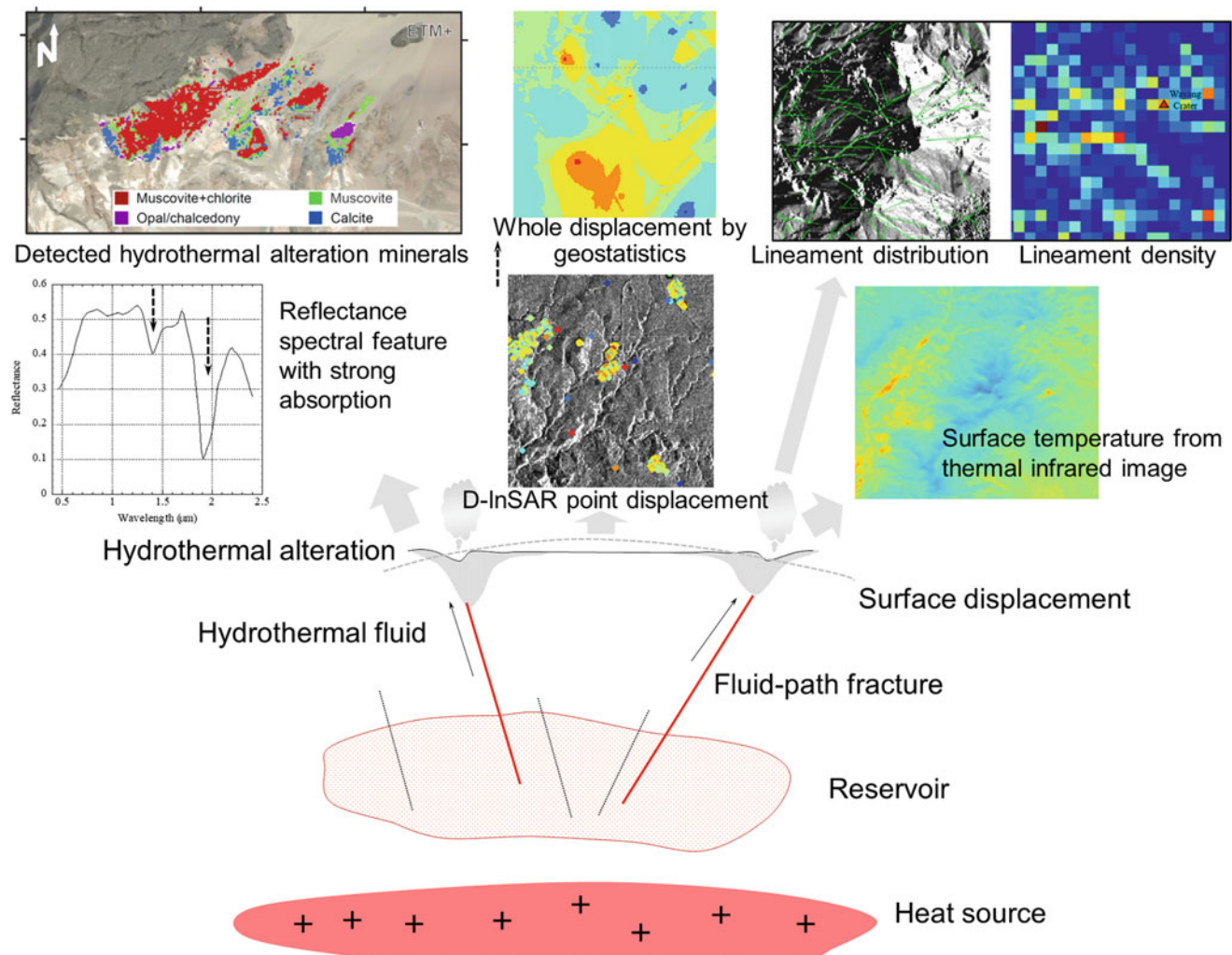
Remote sensing (RS) technology can provide extensive and repeated Earth surface observations in a short time period using electromagnetic waves sensed through optical and microwave sensors onboard satellites, airplanes, and unmanned aerial vehicles. The reflectance of surface materials in visible, near-infrared, and short-wave infrared regions and their radiance in infrared thermal region are the targets of optical sensors, whereas scattering is the main phenomenon for the longer wavelengths of microwave sensors. RS can effectively identify geothermal manifestations (e.g., hot springs, fumaroles, and hydrothermal alteration) over a wide area, which can be used to identify the presence of an underlying latent geothermal resource. RS applications to geothermal resource exploration are typically classified into four categories, as outlined in Fig. 2.

Surface Temperature Estimation

The most traditional application is the measurement of surface temperature based on Planck's law to detect high surface temperature zones. This law describes the spectral radiance B (Wm⁻²μm⁻¹sr⁻¹) of a black body of wavelength λ (m) of an electromagnetic wave at absolute temperature T (K) according to:

$$B = \frac{c_1 \lambda^{-5}}{\pi [\exp(c_2 / \lambda T) - 1]}$$

where c_1 and c_2 are constants derived from the Planck and Boltzmann constants, respectively ($c_1 = 3.7418 \times 10^{-16}$ Wm⁻², $c_2 = 1.4393 \times 10^{-2}$ mK). The surface temperature T_s can be calculated from B in the thermal infrared region and the surface material emissivity. T_s is used to calculate the radiant heat flux Q_r (Wm⁻²), which is proportional



Geothermal Energy, Fig. 2 Schematic diagram of RS applications to detect geothermal energy-related surface features. These include thermal infrared image analysis of surface temperatures and the extraction and mapping of hydrothermal alteration minerals using reflectance

absorption, lineament-based fracture characterization, and topographic change mapping by D-InSAR. These application targets are detected by the development of permeable fractures acting as fluid paths, the upflow of hot fluids, and the existence of deeply seated heat sources

to T_s^4 following the Stefan–Boltzmann law (Vaughan et al. 2012), and to assess geothermal energy.

Due to several factors such as atmospheric and topographic effects, heterogeneity of emissivity may vary over a short distance, and different observation dates, RS-based T_s values typically do not match with in situ measurements. By combining in situ measurements, T_m values, as hard and conditioning data, and RS-based T_s values as soft data, T_s can be corrected to match T_m using geostatistical techniques. The corrected T_s (T_s^*) is given as:

$$T_s^*(\mathbf{u}) = T_m^k(\mathbf{u}) + [T_s(\mathbf{u}) - T_s^k(\mathbf{u})]$$

where $\mathbf{u}$ is the location, $T_m^k(\mathbf{u})$ is the kriging estimate at $\mathbf{u}$ using the in situ measurement data, and $T_s^k(\mathbf{u})$ is another kriging estimate using T_s at the data location.

Another example is to downscale the T_s distribution from coarse- to fine-resolution images using geostatistics (Poggio and Gimona 2013). This approach can be used to effectively reveal the T_s distribution at a finer scale and shorter time interval than the actual data because the observation time intervals of fine images are substantially longer than those of coarse images.

Detection and Mapping of Hydrothermally Altered Minerals

A second traditional application is the detection and mapping of hydrothermal alteration minerals using reflectance band data (i.e., of a selected wavelength range). The analysis of alteration minerals can be used to indicate the temperature and chemical properties, typically acidity of the hydrothermal fluids that generate certain mineral types. For example, water-rock interaction leads to the formation of illite at

approximately 200 °C and Wairakite at approximately 300 °C. The abundance of high-temperature alteration minerals on the Earth's surface suggests the high potential of geothermal energy.

An effective indicator for mineral detection is the wavelength and depth of the reflectance absorption in the visible to short-wave infrared region because this feature is caused by specific ions (Fe^{2+} , Fe^{3+} , OH^- , CO_3^{2-}) and water molecules (H_2O), which are generally contained in hydrothermal alteration minerals. This absorption is enhanced by the band ratio. The simplest measure is described by A/B or $(A/B) \times (C/D)$, where B and D are the low reflectance values of the absorption bands and A and C are the high reflectance values of the reference (i.e., non-absorption) bands. Other band-ratio equations are possible as long as the equation value increases with increasing reflectance absorption at one, two, or more bands. The band-ratio equation for a mineral is termed its mineral index (e.g., alunite index). The band ratio only shows the possibility of a target mineral and has no definite criterion of its value for a mineral's presence or absence. Spectral matching with reference data of standard minerals is another standard method to discriminate minerals, for which machine learning and LASSO logistic regression analysis have been recently adopted to increase the accuracy (Shim et al. 2021).

A critical issue that reduces a mineral's detection accuracy is a limited number of bands in a multispectral sensor image. For example, the number of Terra ASTER sensors for the most widely used imagery in geological studies is nine. In contrast, hyperspectral sensors have hundreds of bands with high spectral resolution, which can be used to unravel the absorption features of a target mineral. However, the availability of hyperspectral images is quite limited in certain areas. Because multispectral images cover almost the entire Earth's surface, they have the possibility to simulate hyperspectral images. The pseudo-hyperspectral image transformation algorithm (PHITA: Hoang and Koike 2017) is a method that can meet this demand. PHITA combines band reflectance data between multispectral and hyperspectral images using a multiple linear regression model as:

$$\rho_{ij}^H(\lambda) = \beta_{0i} + \sum_{k=1}^m \beta_{ki} \cdot \rho_{kj}^M(\lambda) + \varepsilon_{ij}$$

where $\rho_{ij}^H(\lambda)$ represents the reflectance of the hyperspectral band i and pixel j , β_{0i} is the intercept of i , $\rho_{kj}^M(\lambda)$ is the reflectance of multispectral band k at j , β_{ki} is a regression coefficient between k and i , ε_{ij} is a residual, and m is the number of given multispectral bands. Geologic information, such as the mineral index or lithotype, can be incorporated into this equation as an auxiliary parameter to increase the transformation accuracy because reflectance spectra differ for different minerals and rocks. This feature can thus serve as

conditioning for the transformation. More advanced methods to correlate band reflectance data between multi- and hyperspectral images (e.g., machine learning) are applicable instead of linear regression. PHITA has demonstrated an increased detection accuracy of hydrothermal alteration minerals, such as muscovite and calcite, from multispectral images with the same degree of hyperspectral imaging (Hoang and Koike 2018).

Lineament-Based Fracture Characterization

Lineament extraction from optical and microwave sensor images and digital elevation models (DEMs) is another application from the aspect of fluid-path fractures. Lineament is a linear or slightly curved topographic feature that appears in images and shaded DEM images. Its geologic importance arises from the fact that lineaments are generally caused by continuous rock fractures, including faults. Lineaments have therefore been used to characterize fracture distributions in terms of direction, length, and density. In addition to these planar distribution features, the orientation (strike and dip) of dominant fractures can also be estimated by concatenating several lineaments with a similar direction that are located on almost the same line or gentle curve using vector analysis to determine the relationship between lineament direction and normal directions of the slopes in the lineament path (Koike et al. 1998).

A primary purpose of lineament extraction for geothermal studies is to specify permeable fractures that act as pathways of ascending fluid flow from a reservoir. One example is the correspondence of high-density zones of lineaments extracted from synthetic aperture radar (SAR) images with surface features of alteration, mud pools, and hot springs (Saepuloh et al. 2018). For comparison, a fine-scale density map is produced by kriging of the density data defined in a large (1×1 km) cell and assigned to the cell center.

Gas geochemistry using soil gas samples is a particularly effective tool to locate upflow zones. One essential component is radon-222 (Rn), which migrates through permeable fractures. However, the number of sampling sites is usually limited in a given study area because of topographical, temporal, and cost constraints. Detailed Rn concentration features can therefore not be clarified over most study areas. One solution is the application of cokriging using Rn concentration data and lineament density data as the primary and secondary variables, respectively. The effectiveness of this method has been confirmed by the good agreement of high Rn concentration zones with geothermal features (Heriawan et al. 2021).

Continuous Mapping of Topographic Change by D-InSAR

Differential interferometric SAR (D-InSAR) is a technique that measures the topographic change that accompanies crustal deformation, which occurs between image acquisition

dates and is typically caused by earthquakes, volcanic eruptions, landslides, and subsidence. This change is expressed as the displacement velocity (mm/year) from the D-InSAR processing of several image pairs. The applications of D-InSAR to geothermal studies can be categorized into two themes. The first is for resource management. Reservoir pressures are reduced by the overuse of geothermal fluids and insufficient return of fluids to the reservoir, which resultantly leads to surface subsidence. Topographic change monitoring, therefore, contributes to sustainable resource development by preventing the overuse of fluids. The second category is the detection of high potential zones of geothermal energy. Regional topographic change can also be induced by magmatic activity following an increase or decrease of pressure in a deep magma chamber. Geothermal reservoirs form above magma chambers. Areas that intermittently show large topographic changes that are unrelated to landslides may therefore reflect the presence of an active magma chamber that yields large geothermal energy.

One critical issue of D-InSAR is that topographic changes are only detected in pixels with high coherence between an image pair. High coherence cannot be obtained in vegetated areas because the scattering properties of plant surfaces are continuously changeable by growth and meteorological factors. Accordingly, pixels that show a displacement velocity are sparse over a SAR scene. Kriging methods can compensate for undetected displacements in the low-coherence pixels and clarify the overall displacement velocity (Yaseen et al. 2013). A displacement velocity map in a geothermal field produced by kriging contributes to identify the geologic structural controls on topographic change, locate advanced hydrothermal alteration zones, and estimate the sensitivity of a reservoir to production and fluid reinjection (Sabrian et al. 2021).

Fluid Flow Simulation

After constructing a conceptual model of a geologic structure, numerical simulations are implemented to unravel the fluid flow pattern and estimate the power generation by solving the balance equations of mass, momentum, and energy. Because flow is generally laminar and slow, it tends to follow Darcy's law. This simulation consists of three steps: data interpretation; model calibration; and forecasting (Franco and Vaccaro 2014). The first step is to build a reliable numerical model using geological, geophysical, and geochemical data analyses. The numerical model is then refined and its accuracy iteratively improved in the second step by two matchings of the natural state and production history. In the natural state matching, the thermophysical parameters (e.g., permeability, thermal conductivity, porosity, and specific heat capacity) and boundary conditions are determined by matching with the

well-logging temperature and pressure data over long simulation times (typically 10^5 – 10^6 years) for model convergence. Figure 3 depicts an example of the first and second steps. For geothermal power generation areas, the natural state model is iteratively improved by mainly matching with the history data of temperature and mass flow rate with production and reinjection. The third step is simulation for future geothermal fluid development scenarios.

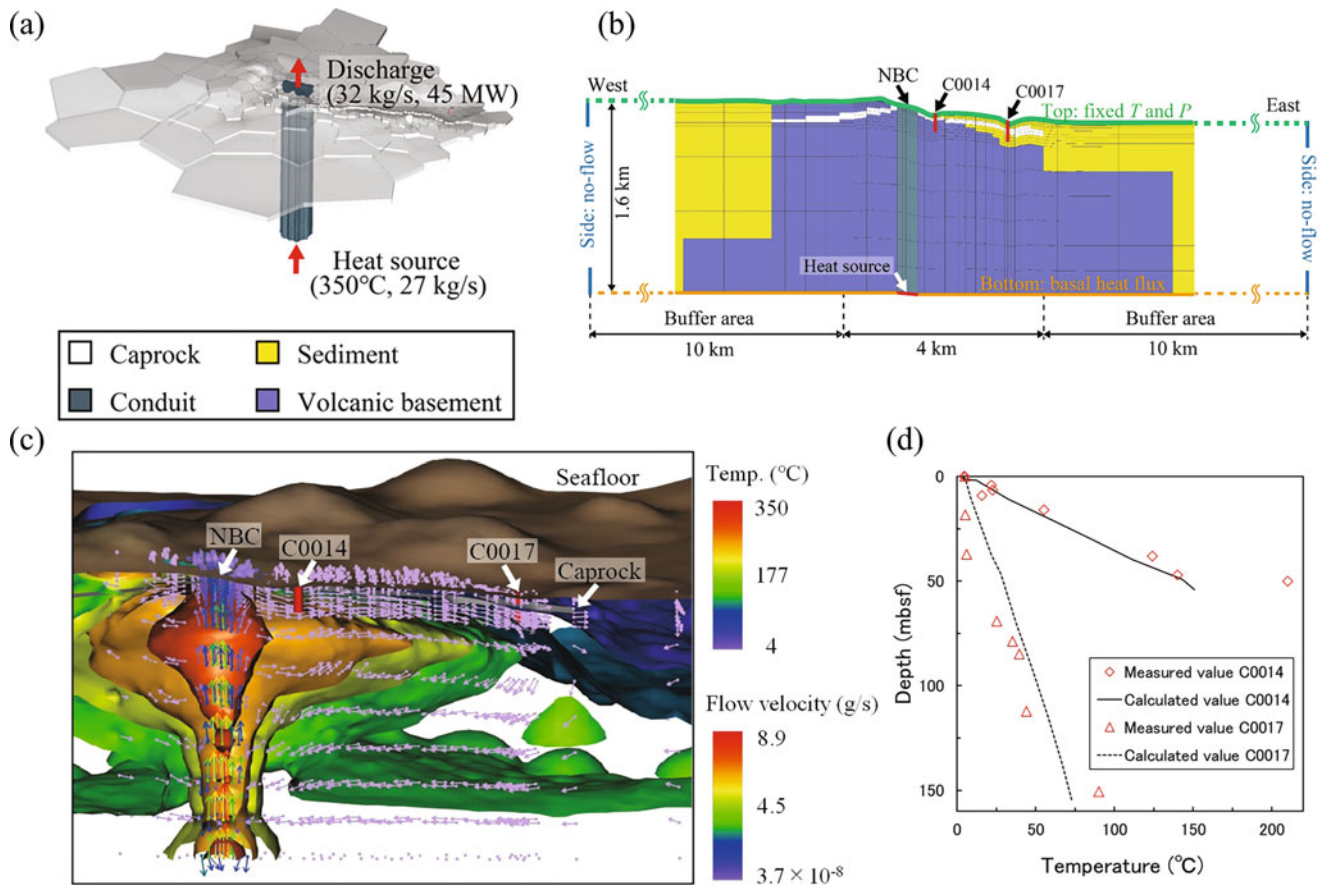
The correct incorporation of a fracture system is essential in the numerical models owing to its role as a path or barrier of fluid flow. Fracture systems in geothermal reservoirs are generally simplified and representatively expressed using a dual-continuum model (DCM) to improve the computational efficiency, and subdivided into dual or multi-porosity models, a dual-permeability model, and a more general multiple interaction continua approach (MINC) (Wu 2016). The DCM treats the matrix and fracture system as two separate continua over the model domain. The MINC can more rigorously handle transient flow between fractures and matrix by subdividing the matrix system in each grid block into a sequence of nested subblocks.

STAR, TETRAD, and TOUGH2 are geothermal reservoir simulators that can handle multiphase and multicomponent fluid flows (O'Sullivan et al. 2001). These simulators were developed to incorporate chemically reactive flow, inverse modeling of thermophysical parameters for model calibration optimization, and the coupling of thermal, hydraulic, mechanical, and chemical properties.

Summary

The potential of geothermal energy for power generation can be evaluated using subsurface and surface properties. The temperature distribution with depth is the most representative subsurface feature, which can be modeled by applying geostatistical and machine learning methods to well-log and bottom-hole temperature datasets. The model should consider the physical laws of heat transfer (conductive- versus convective-type) which govern the temperature-change pattern with depth. This behavior is indispensable because it forms the basis of the temperature trend and allows the temperature data to be used as a non-stationary variable. The incorporation of the thermal type enables the temperature to be estimated in deep regions below the well bottoms. The integration of temperature data and geologic models is effective for interpreting fault systems, fluid flow paths and patterns, and the location and configuration of a reservoir.

The principal surface properties include temperature, hydrothermal alteration minerals, fracture systems, and topographic change, which are detected based on the RS analysis of thermal-infrared images, reflectance-absorption based spectral features, lineaments, and D-InSAR, respectively.



Geothermal Energy, Fig. 3 Example of a calculation model, simulation result, and verification for a seafloor hydrothermal vent area. (Modified from Tomita et al. (2020)). (a) Division of area by Voronoi cells and settings of boundary conditions on the heat source, discharge, and a conduit zone. Only a caprock layer and vertical conduit zone are

shown. (b) Vertical cross-section showing the model size, buffer area, cell configuration, geologic structure composed of four elements, and boundary conditions. (c) 3D view of iso-temperature surfaces and fluid flows (arrows) with seafloor drillings (red lines). (d) Comparison between calculated and measured temperatures at two sites

Although these properties are indirect indicators of geothermal energy, they can provide valuable information regarding upflow zones at high temperature, the development of permeable fractures, and the location and extent of heat sources. Geostatistical methods can be applicable for the conditioning of in situ data for image analysis and the detection of surface displacement over an image scene. The detection accuracy of alteration minerals and permeable fractures can be increased by downscaling the reflectance spectral data using auxiliary geologic parameters and the co-kriging method to include a combination of lineament and geochemical data, respectively.

In addition to these mathematical geologic applications, subsurface physical properties modeled using geophysical methods are essential for imaging geothermal structures. Resistivity is the most common parameter measured by magnetotelluric (MT) and audio-frequency MT surveys. The combination of physical properties applied using the above methods increases the assessment accuracy of geothermal energy potential and provides a deeper understanding of geothermal systems.

Cross-References

- Deep Learning in Geoscience
- Flow in Porous Media
- Geostatistics
- Hyperspectral Remote Sensing
- Kriging
- Machine Learning
- Mineral Prospectivity Analysis
- Object-Based Image Analysis
- Remote Sensing

Bibliography

- Agemar T, Schellschmidt R, Schulz R (2012) Subsurface temperature distribution in Germany. *Geothermics* 44:65–77. <https://doi.org/10.1016/j.geothermics.2012.07.002>
- Franco A, Vaccaro M (2014) Numerical simulation of geothermal reservoirs for the sustainable design of energy plants: a review. *Renew*

- Sust Energ Rev 30:987–1002. <https://doi.org/10.1016/j.rser.2013.11.041>
- Heriawan MN, Syafi'AA, Saepuloh A, Kubo T, Koike K (2021) Detection of near-surface permeable zones based on spatial correlation between radon gas concentration and DTM-derived lineament density. *Nat Resour Res* 30:2989–3015. <https://doi.org/10.1007/s11053-020-09718-z>
- Hoang NT, Koike K (2017) Transformation of Landsat imagery into pseudo-hyperspectral imagery by a multiple regression-based model with application to metal deposit-related minerals mapping. *ISPRS J Photogramm Remote Sens* 133:157–173. <https://doi.org/10.1016/j.isprsjprs.2017.09.016>
- Hoang NT, Koike K (2018) Comparison of hyperspectral transformation accuracies of multispectral Landsat TM, ETM+, OLI and EO-1 ALI images for detecting minerals in a geothermal prospect area. *ISPRS J Photogramm Remote Sens* 137:15–28. <https://doi.org/10.1016/j.isprsjprs.2018.01.007>
- Koike K, Nagano S, Kawaba K (1998) Construction and analysis of interpreted fracture planes through combination of satellite-image derived lineaments and digital elevation model data. *Comput Geosci* 24:573–583. [https://doi.org/10.1016/S0098-3004\(98\)00021-1](https://doi.org/10.1016/S0098-3004(98)00021-1)
- Koike K, Matsuda S, Gu B (2001) Evaluation of interpolation accuracy of neural kriging with application to temperature-distribution analysis. *Math Geol* 33:421–448. <https://doi.org/10.1023/A:1011084812324>
- O'Sullivan MJ, Pruess K, Lippmann MJ (2001) State of the art of geothermal reservoir simulation. *Geothermics* 30:395–429. [https://doi.org/10.1016/S0375-6505\(01\)00005-0](https://doi.org/10.1016/S0375-6505(01)00005-0)
- Poggio L, Gimona A (2013) Modelling high resolution RS data with the aid of coarse resolution data and ancillary data. *Int J Appl Earth Obs Geoinf* 23:360–371. <https://doi.org/10.1016/j.jag.2012.10.010>
- Sabrian PG, Saepuloh A, Kashiwaya K, Koike K (2021) Combined SBAS-InSAR and geostatistics to detect topographic change and fluid paths in geothermal areas. *J Volcanol Geotherm Res* 416:107272. <https://doi.org/10.1016/j.jvolgeores.2021.107272>
- Saepuloh A, Haeruddin H, Heriawan MN, Kubo T, Koike K, Malike D (2018) Application of lineament density extracted from dual orbit of synthetic aperture radar (SAR) images to detecting fluids paths in the Wayang Windu geothermal field (West Java, Indonesia). *Geothermics* 72:145–155. <https://doi.org/10.1016/j.geothermics.2017.11.010>
- Sepúlveda F, Rosenberg MD, Rowland JV, Simmons SF (2012) Kriging predictions of drill-hole stratigraphy and temperature data from the Wairakei geothermal field, New Zealand: implications for conceptual modeling. *Geothermics* 42:13–31. <https://doi.org/10.1016/j.geothermics.2012.01.002>
- Shahdi A, Lee S, Karpatne A, Nojabaei B (2021) Exploratory analysis of machine learning methods in predicting subsurface temperature and geothermal gradient of Northeastern United States. *Geothermal Energy* 9:18. <https://doi.org/10.1186/s40517-021-00200-4>
- Shim K, Yu J, Wang L, Lee S, Koh S-M, Lee BH (2021) Content controlled spectral indices for detection of hydrothermal alteration minerals based on machine learning and lasso-logistic regression analysis. *IEEE J Selected Topics Appl Earth Observ Remote Sens* 14:7435–7447. <https://doi.org/10.1109/JSTARS.2021.3095926>
- Teng Y, Koike K (2007) Three-dimensional imaging of a geothermal system using temperature and geological models derived from a well-log dataset. *Geothermics* 36:518–538. <https://doi.org/10.1016/j.geothermics.2007.07.006>
- Tian B, Wang L, Kashiwaya K, Koike K (2015) Combination of well-logging temperature and thermal remote sensing for characterization of geothermal resources in Hokkaido, northern Japan. *Remote Sens* 7:2647–2667. <https://doi.org/10.3390/rs70302647>
- Tomita SA, Koike K, Goto T-N, Suzuki K (2020) Numerical simulation-based clarification of a fluid-flow system in a seafloor hydrothermal vent area in the middle Okinawa trough. *Geophys Res Lett* 47:e2020GL088681. <https://doi.org/10.1029/2020GL088681>
- Turcotte D, Schubert G (2014) *Geodynamics*, 3rd edn. Cambridge University Press, Cambridge, pp 164–167, 219–223, 623
- Vaughan RG, Keszthelyi LP, Lowenstern JB, Jaworowski C, Heasler H (2012) Use of ASTER and MODIS thermal infrared data to quantify heat flow and hydrothermal change at Yellowstone National Park. *J Volcanol Geotherm Res* 233–234:72–89. <https://doi.org/10.1016/j.jvolgeores.2012.04.022>
- Wu Y-S (2016) *Multiphase fluid flow in porous and fractured reservoirs*. Gulf Professional Publishing. <https://doi.org/10.1016/C2015-0-00766-3>
- Yaseen M, Hamm NAS, Woldai T, Tolpekin VA, Stein A (2013) Local interpolation of coseismic displacements measured by InSAR. *Int J Appl Earth Obs Geoinf* 23:1–17. <https://doi.org/10.1016/j.jag.2012.12.002>

Global and Regional Climatic Modeling

Yuriy Kostyuchenko^{1,2}, Igor Artemenko¹, Mohamed Abioui³ and Mohammed Benssaou³

¹Heat and Mass Exchange in Geo-systems Department, Scientific Centre for Aerospace Research of the Earth, National Academy of Sciences of Ukraine, Kiev, Ukraine

²International Institute for Applied Systems Analysis (IIASA), Laxenbourg, Austria

³Department of Earth Sciences, Faculty of Sciences, Ibn Zohr University, Agadir, Morocco

Definition

Global and regional climate models are formalized descriptions of the climate system, which differ by the resolution (generalization) or modeling step. The result of the simulation is global, regional (or local) distributions of climate data, with a spatial resolution of 500–100 or 10–50 km, respectively. Accordingly, these models refer to the impact of the processes of different scales and origins in the atmosphere, ocean, and land surface.

Introduction

The rapid development of computational methods in recent decades has led to the transformation of applied mathematics in geosciences into a successful tool for solving many important tasks (Sagar et al. 2018): from natural resources exploration to meteorology.

Among other problems connected with human well-being and security, there are several complex tasks, where the application of mathematical geosciences is required, such as support of sustainable development, climate change adaptation, disaster risk reduction or reducing vulnerability, and enhancing the resilience of ecosystems and communities.

Solving these tasks should be based on the multiscale climate predictions, which requires the development of climate models – a direct task of mathematical geoscience.

Climate: Basic Approaches to Analysis

When we speak about climate, usually we focus on the slowly varying aspects of the atmosphere-hydrosphere-land surface system (Miller 2019). Climate may be usually described by average parameters and variables of the climate system during the periods of a month or more, considering the time variability of parameters and variables. Classification of climatic types usually include the spatial variation of used averaged variables. The climate description concept has substantially evolved with the broadening of the understanding of the key climate-determining processes and its variations: covering more than annual local variations of averages of surface temperature and precipitation.

From the viewpoint of sustainability, usually, we are interested to study climate change and climate anomalies.

Climate anomaly may be described as the quantitative difference between the average decadal climatic parameters, and the climatic parameters during separate month/season.

At the same time, the long-term systematic statistics trends of the climatic parameters (such as temperature, pressure, or winds) should be enough sustained over a period of several decades. Climate change may be caused by the natural external forcing (for example, the changes of solar emission, or changes in the Earth's orbital elements), by natural internal processes in the climate system, or anthropogenic forcing.

To understand the behavior and variations of the climate system the climate predictions are usually applied.

Climate prediction is the calculation-based prediction of different parameters and aspects of the climate system during determined future time period. Climate predictions are usually presenting as the set of interconnected probabilities of anomalies of climate variables (temperature, precipitation, etc.), for the period of several seasons. Also, the term “climate projection” is now commonly used for long-term predictions rather than “climate prediction,” because described scenarios have a higher uncertainty and less specificity (for example, “projection” is preferentially used in analysis of climate reactions to variations of land use).

A climate system – the system, comprising the atmosphere, hydrosphere, lithosphere, and biosphere, interrelation and external forcing reaction which determine the planetary climate, and physical, chemical, and biological processes in these components define the interactions among the components of the climate system (Trenberth 1992).

Modeling is the usual method for analysis of climate system behavior (Howarth 2017). A model is a tool for simulation and simulation-based prediction of the behavior

of the dynamic systems, for example, the Earth's atmosphere. Varied models can be based on varied approaches: subjective heuristic methods, statistics, simplified physical systems, analogy, but now this term is most commonly applied to numerical models, wherein different models require varied methods of model calibration (Howarth 2017), such as different approaches to adjusting of the numerical model parameters to optimize the agreement between observed and simulated data (model's predictions).

Modeling: Tools and Approaches

Modern approaches to the analysis of climate system behavior are mainly based on the varied coupled ocean-atmosphere circulation models (Jursa 1985).

Coupled ocean-atmosphere general circulation model is a climate system model that includes coupled interacting atmospheric and oceanic components. The atmospheric and oceanic components in many cases presented as the general circulation models in global or regional scale. These types of models are used for study of the climate system response to time variations of external and internal forcing parameters, such as the atmospheric greenhouse gases concentration. Key models' control parameters include surface pressure, horizontal components of velocity, temperature and water vapor, solar and terrestrial short, long wave, infrared radiation, convection, Albedo, hydrology, and cloud cover.

Usually, such types of models include also geosphere-biosphere components, describing water and carbon balance inland and ocean's ecosystems.

A key component of climate models is the atmosphere models (Jursa 1985). The atmosphere model is a formal theoretical representation of the atmosphere, first of all, the vertical temperature distribution.

Now there are numerous atmosphere models used for analysis of climate system and predictions of its behavior: adiabatic atmosphere, homogeneous atmosphere, isothermal atmosphere, thermotropic model, equivalent barotropic model, barotropic model, and standard atmosphere. These models are traditionally used in climate system analysis.

Adiabatic atmosphere (or “dry-adiabatic atmosphere,” “convective atmosphere”) is a traditional widely used atmosphere model with a dry-adiabatic lapse rate along the vertical axis, and the pressure in the adiabatic atmosphere decreases with height. Practically, such a condition is never truly observed, and also this description is not correct enough, because “adiabatic” represents a process, not a condition.

A homogeneous atmosphere is a hypothetical atmosphere representation, in which the density is constant with height. In this atmosphere model the lapse rate of temperature is presented as the autoconvective lapse rate and is equal to g/R (approximately $3.4^{\circ}\text{C}/100\text{ m}$), where g is the acceleration of

gravity, and R is the gas constant. A homogeneous atmosphere has a finite total thickness, determined as $R_d T_v / g$ (where R_d is the gas constant for dry air, and T_v is the virtual surface temperature). For a surface temperature of 273 K, the vertical extent of the homogeneous atmosphere is approximately 8000 m. Formally, a homogeneous atmosphere model is similar to the adiabatic atmosphere.

Isothermal atmosphere (also called “exponential atmosphere”) is an idealized atmosphere representation in hydrostatic equilibrium. In this model the temperature is constant with a height, and the pressure decreases exponentially.

The thermotropic atmosphere model is an atmosphere model for numerical forecasting, in which the simulated parameters are the height of the constant-pressure surface (usually 500 mb) and the temperature (usually the mean temperature). In this model the quasi-geostrophic approximation is applied, and the thermal wind is assumed height-constant. Using this model, surface prognostic maps can be constructed.

The equivalent barotropic atmosphere model is an advanced representation of the barotropic model. In this model the wind height variation is averaged, based on the assumption that the thermal and geostrophic wind has the same directions at all heights. The resulting model is simulating the geostrophic stream function as the single-parameter representation of the barotropic vorticity equation. Due to its simplicity, the equivalent barotropic model is the very popular barotropic model for numerical forecasts using real observation data.

Barotropic atmosphere model is a designation of the cluster of atmosphere models, which is characterized by the following conditions: when pressure and temperature surfaces coincide; the vertical wind shear is absent; the vertical motions are absent; the horizontal velocity divergence is absent; and the vertical component of absolute vorticity is conserved. There are two usual classes of the barotropic models: the nondivergent and the divergent barotropic model (which is also called the “shallow-water equations models”); the simplest form of the barotropic model is based on the barotropic vorticity advection equation. Barotropic atmosphere models cannot predict the behavior of weather systems, because there are no vertical variations or thermal advection processes in this type of model.

The standard atmosphere is a formal description of hypothetical vertical distribution of the determined set of parameters: atmospheric temperature, pressure, and density. These parameters, by common agreement, are assumed as the representative for instrumental purposes: altimeter calibrations, aircraft performance calculations, aircraft and missile design, ballistic tables, etc. (Jursa 1985).

In this model, the air is assumed to obey the perfect gas law and the hydrostatic equation. These relations define together

interconnected temperature, pressure, and density vertical variations. It is also assumed that the air contains no water vapor and that the acceleration of gravity is constant with height (so, for representing the measure of vertical displacement a particular unit of geopotential height may be adopted in place of a unit of geometric height, because these two units are numerically equivalent). The current standard atmosphere model was adopted in 1976 and is a slight modification of 1952 model adopted by the International Civil Aeronautical Organization (ICAO), which, replaced the NACA Standard Atmosphere (or US Standard Atmosphere) of 1925. This model assumes the following parameters for sea level: temperature is 288.15 K (15°C), the pressure is 101 325 Pa (1013.25 mb, 760 mm of Hg, or 29.92 of Hg), density is 1225 g m⁻² (1.225 g L⁻¹), and mean molar mass is 28.964 g mole⁻¹.

The assumptions about parameters and constants in the standard atmosphere model are the following: zero pressure altitude is 760 mm high of mercury column (this pressure is equal to 1.013250×10^6 dynes cm⁻² or 1013.250 mb, and is known as “one standard atmosphere” or “one atmosphere”), the temperature at zero pressure altitude is 15°C or 288.15 K, the gas constant for dry air is 2.8704×10^6 ergs gm⁻¹K⁻¹, the ice point at the standard atmosphere pressure is 273.16 K, the density at zero pressure altitude is 0.0012250 gm cm⁻³, the lapse rate of temperature in the tropopause is 6.5°C km⁻¹, the pressure altitude of the tropopause is 11 km, the temperature at the tropopause is -56.5°C, and the acceleration of gravity is 980.665 cm s⁻².

The ARDC (Air Research and Development Command) Model Atmosphere, 1959, extended the above standard as follows: the lapse rate from 11 km to 25 km is 0°C km⁻¹, the lapse rate from 25 km to 47 km is +3.0°C km⁻¹, and temperature at 47 km is +9.5°C, the lapse rate from 47 km to 53 km is 0°C km⁻¹, the lapse rate from 53 km to 75 km is -3.9°C km⁻¹, and temperature at 75 km is -76.3°C, the lapse rate from 75 km to 90 km is 0°C km⁻¹, the lapse rate from 90 km to 126 km is +3.5°C km⁻¹; the temperature at 126 km is +49.7°C (molecular-scale temperatures), the lapse rate from 126 km to 175 km is +10.0°C km⁻¹; the temperature at 175 km is 539.7°C (molecular-scale temperatures), the lapse rate from 175 km to 500 km is +5.8°C km⁻¹; the temperature at 500 km is 2424.7°C (molecular-scale temperatures).

In the contemporary version of this model, a standard unit of atmospheric pressure, the 45° atmosphere, is defined as that pressure of a 760-mm mercury column at 45° latitude at sea level at temperature 0°C. This is a unit recommended for meteorological use.

So, as we can see, the existing complex multicomponent models mainly are oriented to the analysis of the climatic system as a whole, on the global scale, and need to be reduced to subregional or especially local scales in varied ways.

Global Versus Local

Classic climatology traditionally operates on macro-, meso-, and microclimate terms to analyze varied scales specific of the climatic parameters (Miller 2019; Arias 1942).

Macroclimate is the general large-scale climate of a large area (for example, country), as distinguished from the mesoclimate and microclimate.

Mesoclimate is the climate of a natural region of a relatively small extent (for example, valley, forest, plantation, and park). Due to insignificant elevation and exposure variations, the climate may not be representative of the general regional climate.

A microclimate is the fine climatic structure of the near-surface air space that extends from the Earth surface to a height where the effects of the underlying surface no longer can be distinguished from the general local climate (mesoclimate or macroclimate). The microclimate varies with and is superimposed upon the larger-scale climate conditions. There are some formal limits for determination of the thickness of the microclimatic layer concerned: in the general case, four times the height of surface features define the level where microclimatic layers disappear. However, in many cases this layer thickness should be considered as variable. There are many different classes of microclimate depending on the types of the underlying surface. Depending on land cover and land use types detalization/typization, several microclimate classes may vary from 7 up to 37 and more. Currently, the most studied and applicable types of microclimates are the “urban microclimate” (affected by street paving, buildings, industrial construction, air pollution, inhabitation, etc.), and the “vegetation/plant microclimate” (conditioned by the complex nature of interactions in the airspace over the vegetation landscapes).

At the same time, such an approach is substantially limited in numerical modeling, because the model resolution is limited and depends on the modeling step, e.g., of available input data, topology, modeling method, etc.

The main part of existing models based on the finite difference method uses non-rectangular grids with a variable resolution between 1 and 5 degrees in latitude or longitude (Collins et al. 2004). Most usable models have resolution 3.75 in longitude and 2.5 degrees in latitude (Collins et al. 2004), 1.875 degrees in longitude and 1.25 in latitude, and 1.25 in longitude x 0.83 degrees in longitude (Shaffrey et al. 2017). So, these are mainly global models to analyze macroclimatic parameters.

At the same time many important socioecological tasks, such as local community resilience, ecosystems’ vulnerability, agricultural production analysis, etc., require analysis of climatic parameter distribution with a resolution of at least 10–30 km (meso- and microclimatic analysis). In such a situation, application of varied algorithms of data regularization and downscaling is needed to obtain the regional,

subregional, and local distribution of climatic parameters instead of global.

Regularization is a formal technique for selecting the eligible level of model complexity (generalization) to give better model predictivity. This procedure is applied, in particular, for reducing topology effects and impact of non-rectangular grids. The resulting data has regular spatial and temporal distribution in a required coordinate system with the required accuracy and controlled reliability.

Downscaling is a procedure to obtain high-resolution information from low-resolution variables.

Application of regularization and downscaling procedures make it possible to analyze subregional and local climate parameters anomalies and changes. Local change is the change in a variable during a specified interval of time at a fixed point in the coordinate system in use. It is used to analyze the impact of spatially distributed processes on the local scales (microclimatic analysis).

However, as a result of the application of regularization and downscaling procedures, such regional/local distributions of parameters can be obtained that will not correspond to global models.

Summary and Conclusions

The investigated climate system is characterized by deep uncertainties, depending on the spatial and temporal scales (Caers 2011). At the same time, the existing models of the climate system are complex multicomponent, heterogeneous models based on different approaches and tools. In this context, the following should be noted.

First, regularization and downscaling procedures should not be based on simple averaging, but on complex statistical algorithms, and should take into account additional data, so as not to lose important local variations. So, in essence, local distributions and global models may be recognized as methodologically different cases.

Second, traditional approaches to the assessment of model accuracy, correctness, reliability, and usability cannot be applied to climate models and need to be modified depending on the type of model, scale, problem to be solved, and types of data used.

Understanding of the methodological and methodological context, as well as the existing limitations, will allow correctly building a hierarchy of climate models by spatial resolution and adequately solving the required applied tasks.

Cross-References

- [Coupled Modeling](#)
- [Earth Surface Processes](#)
- [Machine Learning](#)
- [Optimization in Geosciences](#)

- [Scaling and Scale Invariance](#)
- [Simulation](#)
- [Uncertainty Quantification](#)

Acknowledgments The authors are grateful to anonymous referees for constructive suggestions that resulted in important improvements to the entry, to colleagues from the American Statistical Association (ASA), and from the American Meteorological Society (AMS) for presented data and materials, as well as for their critical and constructive comments and suggestions.

Bibliography

- Arias AC (1942) The classification of climates. *Mon Weather Rev* 70(11):249–225
- Caers J (2011) *Modeling uncertainty in the Earth sciences*. Wiley, Chichester, 229pp
- Collins WD, Rasch PJ, Boville BA, Hack JJ, McCaa JR, Williamson DL, Kiehl JT, Briegleb B, Bitz C, Lin SJ, Zhang M (2004) Description of the NCAR community atmosphere model (CAM 3.0). Tech Rep NCAR/TN-464+STR. National Center for Atmospheric Research, Boulder, 226p
- Howarth RJ (2017) *Dictionary of mathematical geosciences*. Springer, Cham, 893pp
- Jursa AS (1985) *Handbook of geophysics and the space environment*, 4th edn. Air Force Geophysics Lab, Hanscom AFB
- Miller AA (2019) *Climatology*. Routledge, London, 330pp
- Sagar BSD, Cheng Q, Agterberg F (eds) (2018) *Handbook of mathematical geosciences: fifty years of IAMG*. Springer, Cham, 914pp
- Shaffrey LC, Hodson D, Robson J, Stevens DP, Hawkins E, Polo I, Stevens I, Sutton RT, Lister G, Iwi A, Smith D (2017) Decadal predictions with the HiGEM high resolution global coupled climate model: description and basic evaluation. *Clim Dyn* 48(1–2):297–311
- Trenberth KE (ed) (1992) *Climate system modeling*. Cambridge University Press, Cambridge, 788pp

Goodchild, Michael F.

B. S. Daya Sagar
Systems Science and Informatics Unit, Indian Statistical
Institute, Bangalore, Karnataka, India



Fig. 1 Michael F. Goodchild. (Courtesy of Prof. Goodchild)

Biography

Michael Frank Goodchild (born February 24, 1944), a British American, is a highly decorated academic and scientist. He is an Emeritus Professor of Geography at the University of California, Santa Barbara (UCSB). He joined UCSB in 1988 as part of the establishment of the National Center for Geographic Information and Analysis (NCGIA), which he directed for over two decades, after 19 years at the University of Western Ontario, including 3 years as chair. In 2008, he founded the UCSB Center for Spatial Studies, where he held until 2012 the Jack and Laura Dangermond Chair of Geography and oversaw the activities of the center as the director. His research and teaching interests focus mainly on issues in geographical information science (GISci), including uncertainty in geographic information, discrete global grids, and volunteered geographic information. He directed or codirected several large funded projects, including the NCGIA, the Alexandria Digital Library, and the Center for Spatially Integrated Social Science. He moved to Seattle upon superannuation in 2012 and currently holds part-time positions as Research Professor at Arizona State University and as Distinguished Chair Professor at Hong Kong Polytechnic University. He lectures widely on the popular topic of “Science of Where.”

He had his undergraduate in physics from the Downing College of the Cambridge University in 1965, followed by his Ph.D. in geography from the McMaster University, Canada, in 1969. As a doctoral student, he rediscovered 20-kilometer-long Castleguard Cave in Canada. His most powerful work has involved research on GISci (aka GIS). He is widely credited with coining “volunteered geographic information” and other models for the production of geographic information. He has been a pioneer in the areas of GISci (Goodchild and Mark 1988, Longley et al. 2011), science, and technology, and has been known the world over for setting GISci foundations in the early 1990s as the science of collecting, analyzing, digitizing, and using geographic data. He has published over 550 books and articles. The way he led research involving quantitative techniques such as fractal geometry, geostatistics, and spatial statistics for managing and analyzing the very large databases, now called “Big Data,” that result and handling uncertainty in the data is remarkable. He introduced theoretical frameworks of fractals (Goodchild 1980), geostatistics (Goodchild 1980, Goodchild and Lam 1980), and spatial statistics (Goodchild and Klinkenberg 1993, Goodchild 2008, 2009) to address the issues related to accuracy and uncertainty (Goodchild 1991) in all of these areas. His interest is in discrete global grids, a set of methods for representing variation over the curved surface of the Earth and implemented in tools such as Google Earth. Of late, he is also spearheading new ways of mass collection and analysis of geographic data with crowdsourced

mapping and citizen science projects, wayfinding, and urban mobility, including the long-term impacts of connected and autonomous vehicles.

Several books and edited volumes that Professor Goodchild published are of huge value not only for the GIS and geography communities but also for the mathematical geoscience community as most of the temporal data handled by mathematical geoscientists are spatial in nature. I am one of those large pools of researchers who benefited a lot from working through his publications associated with spatial data sciences while introducing mathematical morphology to address various challenges, encountered in (geo)spatial data science, such as information retrieval, information characterization, quantitative information reasoning, simulation, modeling, and spatiotemporal visualization.

Goodchild received many awards and honors from top-notch academic societies and associations. To mention a few, he has received the Prix Vautrin Lud (2007) – informally known as the “Nobel Prize of Geography” – and the Royal Geographical Society’s Founder’s Medal (2003), CODATA Prize (2012) from the Committee on Data for Science and Technology of the International Council for Science, and International Spatial Accuracy Research Association’s Peter A. Burrough Medal (2012). He has been elected as a member of the American Academy of Arts and Sciences (2006), the US National Academy of Sciences (2002), and as a fellow or foreign member of the Royal Society of London (2010), British Academy (2007), and Royal Society of Canada (2002). He has received honorary doctorates from many universities, including, to mention a few, Ryerson University, McMaster University, Keele University, and Université Laval.

The Goodchild’s “Science of Where” has attracted many young-generation students and researchers, and they all see him as one of the founding fathers of the GISci and a role model.

Cross-References

- [Fractal Geometry in Geosciences](#)
- [Geographical Information Science](#)
- [Geostatistics](#)
- [Spatial Analysis](#)

References

- Goodchild MF (1980) Fractals and the accuracy of geographical measures. *Math Geol* 12:85–98
- Goodchild MF, Lam N (1980) Areal interpolation: a variant of the traditional spatial problem. *Geoprocessing* 1:297–312
- Goodchild MF, Mark DM (1988) The invertibility of distance matrices. In: Coffey W (ed) *Geographical systems and systems of geography: essays in honour of William Warntz*. Department of Geography, University of Western Ontario, Ontario, pp 141–151
- Goodchild MF (1991) Issues of quality and uncertainty. In: Muller JC (ed) *Advances in cartography*. Elsevier, New York, pp 113–139
- Goodchild MF, Klinkenberg B (1993) Statistics of channel networks on fractional Brownian surfaces. In: Lam N, de Cola L (eds) *Fractals in geography*. Prentice-Hall, Englewood Cliffs, pp 122–141
- Goodchild MF (2008) Statistical perspectives on geographic information science. *Geogr Anal* 40(3):310–325
- Goodchild MF (2009) What problem? Spatial autocorrelation and geographic information science. *Geogr Anal* 41(4):411–417
- Longley PA, Goodchild MF, Maguire DJ, Rhind DW (2011) *Geographical information systems and science*, 3rd edn. Wiley, Hoboken

Gradstein, Felix

Frits Agterberg

Central Canada Division, Geomathematics Section,
Geological Survey of Canada, Ottawa, ON, Canada



Fig.1 Prof. Felix Gradstein (2021). (Courtesy of Prof. F.M. Gradstein)

Biography

Born in 1941 in Eindhoven, the Netherlands, Felix Gradstein studied paleontology and stratigraphy at the University of Utrecht under the supervision of Professor CW Drooger. This early research resulting in his PhD thesis (Gradstein 1974) was concerned with a novel biometrical approach in micropaleontology. After working 2 years for an oil company in Calgary, Canada, he joined the Geological Survey of Canada in its eastern division at the Bedford Institute of Oceanography in Nova Scotia. He was instrumental in developing a novel quantitative approach to the analysis of stratigraphic

events (observed highest or lowest occurrences in hydrocarbon wells or natural sections) ranking and scaling them, initially for the Northwest Atlantic continental margin (Gradstein and Agterberg 1982). Since then, Ranking and Scaling (RASC) has remained a key method for quantitative biostratigraphy of fossil event data. In 1992 Felix moved to Norway where he currently has an office at the University of Oslo.

Gradstein's major contributions to geoscience included his work on the Geologic Time Scale that started in 1985/1986 when he helped construct the Decade of North America geologic time scale (Kent and Gradstein 1986). This was the first scale to use the Mesozoic marine magnetic anomaly patterns for stage scaling. Soon thereafter Felix took over from Professor WB Harland as editor-in-chief of the international Geologic Time Scale (GTS) project. His first GTS (Gradstein et al. 2004) was published by Cambridge University Press, but the 2012 and 2020 versions were published by Elsevier. GTS2020 had a team of 4 editors to standardize all radiogenic isotope ages, the mathematical interpolations, and the many Periods and specialist discipline chapters contributed by over 100 specialists. For the Cenozoic, GTS2020 is entirely based on Milankovic'-based astrocylicity. For nearly all of the Mesozoic and Paleozoic, cubic-spline fitting (age versus chronostratigraphy) was used for scaling and 95% confidence intervals were estimated for the stage boundary ages.

Currently, Felix is Professor Emeritus at Oslo University, Norway, and visiting Research Fellow, University of Portsmouth, UK. From 2000 to 2008, he was chair of the International Commission on Stratigraphy. Under his leadership, major progress was made with the formal definition of chronostratigraphic units from Precambrian through Quaternary. His major contributions to the geosciences have been widely recognized. For his fundamental work concerning the Geologic Time Scale, geochronology in general, quantitative stratigraphy and micropaleontology, the European Geosciences Union awarded him in 2010 the Jean Baptiste Lamarck Medal. He is Chair of the Geologic Time Scale Foundation and teaches courses in quantitative stratigraphy and the geologic time scale. After completion of GTS2020, he has free time again and studies the early evolution of planktonic foraminifera.

Bibliography

- Gradstein FM (1974) Mediterranean Pliocene Globorotalia – a biometrical approach. PhD thesis, Utrecht Univ, published by Krips Repro, Meppel, Netherlands, 128 pp
- Gradstein FM, Agterberg FP (1982) Models of cenozoic foraminiferal stratigraphy – Northwestern Atlantic margin. In: Cubitt JM, Reymont RA (eds) Quantitative stratigraphic correlation. Wiley, New York, pp 119–173
- Gradstein FM, Agterberg FP, Brower JC, Schwarzacher WS (1985) Quantitative stratigraphy. Reidel/UNESCO, Dordrecht/Paris, p 598
- Gradstein FM, Ogg J, Smith A (2004) A geologic time scale. Cambridge University Press, Cambridge, p 589
- Gradstein FM, Ogg J, Schmitz MB, Ogg GM (eds) (2020) Geologic time scale 2020. Two volumes. Elsevier, Amsterdam, p 1357
- Kent DV, Gradstein FM (1986) A Jurassic to recent geochronology. In: Vogt PR, Tucholke BE (eds) The geology of North America vol M The Western North Atlantic region. Geological Society of America, Boulder, pp 45–50

Grain Size Analysis

Michael Klichowicz and Holger Lieberwirth
Institute of Mineral Processing Machines and Recycling
Systems Technology, Technische Universität Bergakademie
Freiberg, Freiberg, Germany

Synonyms

Crystal size analysis; Granulometry; Particle size analysis

Definition

Grains are three-dimensional objects of a size between that of dust particles and rock fragments. They can be loose, as sand or dust grains or fragments of broken stones, or may appear as embedded objects like crystals or pores in textures of stones. Typically, populations of grains are studied. Care is necessary when grains contain smaller objects, which also may be called “grains” in the same context.

This entry always uses the general term “grain,” which is specified where needed as “loose grain” or “embedded grain.” Size is a geometric characteristic of a grain. It is often a one-dimensional quantity like diameter, which may result from two- or three-dimensional measurement and some calculation, where grain shape plays an important role. There are many different methods of measurement and specialized statistics.

Overview

Dispersed materials are omnipresent both in nature and in the way in which humans use nature. In geosciences, grain size analysis is often used to characterize dispersed materials, which are powdery, granular, or lumpy material mixtures. This ranges from loose rock material such as soil or more general sediments to blasted or crushed hard rock in mining. However, grain size analysis is not limited to rock materials

but also applied for numerous other tasks like characterizing suspended sediments in water bodies or quantifying the texture of mineral material, which is often formed by a multitude of grains.

Standard references to grain size analysis are Allen (1990) and Bernhardt (1994). Both textbooks approach the topic in a relatively general way. If the size of embedded grains like crystals is in the focus, Higgins (2006) and Randolph and Larson (1988) are valuable references. Further, Rumpf and Scarlett (1990) gives a summary for natural materials, which are further processed, like crushed hard rock. The notation of the international standard ISO 9276-1 (2004) is used as far as possible in the following.

Grain Size

Grain size analysis is always bound to the *definition* of grain size, the size of the individual three-dimensional elements forming grain collectives. Unlike regular geometric objects like spheres or cubes, there is no single axiomatic parameter to describe the size of an irregularly shaped grain. In contrast, there is a multitude of size definitions, which mainly depend upon the type of grain and measurement. The measurable physical properties used to characterize the size x of a grain can be grouped as follows:

- (a) *Characteristic length parameters*: The grain size can be defined by geometric lengths, which are directly measured from the grain or from a sectioned or projected grain outline.
- (b) *Surface or projected area*: The grain size can be characterized by its surface, which can be measured with gas-adsorption methods. However, measuring the projection area is more common since it can be realized relatively simple with means of image processing.
- (c) *Volume or mass*: The volume of coarse grains can be measured by fluid displacement. A simpler method is weighting.
- (d) *Stationary sedimentation rate*: Assuming objects of the same density and similar shape, their sedimentation rate in a fluid is only affected by their sizes. Hence, this characteristic can be used to define grain size.
- (e) *Field disturbance effects*: The size of relatively small grains can often be determined by reliably measuring their interference effects with electric, electromagnetic, or fluid-dynamic fields.

Table 1 shows some common geometrical definitions of grain size. Often, a spherical reference grain is used to ease

understanding if originally higher-dimensional characteristics were measured. An example is the Stokes diameter d_{st} .

Grain Size Distributions

Usually populations of grains are considered, the sizes of which are variable. Size measurement leads to data lists, which are statistically analyzed.

The Cumulative Size Distribution Function $Q_r(x)$

The variability of grain sizes x is usually described by grain size distribution functions $Q_r(x)$ for $r = 0, 1, 2, 3$. These functions describe how many grains measured in the quantity type r (Table 2) are smaller than a given grain size x for $x_{\min} \leq x \leq x_{\max}$. Commonly, the mass distribution with $r = 3$ or the number distribution with $r = 0$ are used.

In particular

$$Q_0(x) = \frac{\text{number of grains} < x}{\text{total number of grains}} \quad \text{for } x_{\min} \leq x \leq x_{\max}, \quad (1)$$

$$Q_3(x) = \frac{\text{mass of grains} < x}{\text{mass of all grains}} \quad \text{for } x_{\min} \leq x \leq x_{\max}. \quad (2)$$

Frequently these distributions are given in histogram form according to the principle of measurement.

The cumulative size distribution $Q_r(x)$ is plotted against grain size x . In many cases it is practicable to use a logarithmic scale for $r = 3$ (Fig. 1).

The use of a plot with logarithmic x -scale is quasi-standard in some application fields, since size differences between 0.1 mm and 0.2 mm are frequently considered of the same importance as size difference between 1 mm and 2 mm.

The Probability Density Function $q_r(x)$ of Grain Size

Given that the cumulative distribution function $Q_r(x)$ is differentiable, there exists the probability density function $q_r(x)$, which is given by:

$$q_r(x) = \frac{dQ_r(x)}{dx} \quad \text{for } x_{\min} \leq x \leq x_{\max} \quad (6)$$

The density function can be estimated by means of kernel estimators; however, there are difficulties since the data are often given in histogram form. Analogously to the cumulative size distribution function, it can be useful to plot the frequency distribution in a log-normal plot.

Grain Size Analysis, Table 1 Definitions of grain sizes

Symbol	Name	Definition	
<i>Size definitions based on projections or sections</i>			
d_M	Martin's diameter	Length of the area bisector in measuring direction	
d_F	Feret's diameter	Distance between two parallel lines restricting the object perpendicular to the measuring direction	
d_C	Maximum chord	Maximum chord length in measuring direction	
d_p	Perimeter diameter	Diameter of a circle having the same circumference as the projected outline of the particle	
d_A	Projected area diameter	Diameter of a circle having the same area as the projected outline of the particle	
<i>Size definitions based on spatial characteristics</i>			
a	Length	Length of the bounding box	
b	Width	Width of the bounding box	
c	Height	Height of the bounding box	
d_v	Volume diameter	Diameter of a sphere having the same volume as the particle	
d_s	Surface diameter	Diameter of a sphere having the same surface as the particle	
d_{St}	Stokes diameter	Diameter of a sphere having the same density and in the laminar flow region the same free-falling velocity as the particle in a fluid of the same density and viscosity	

Figure 2 shows a typical histogram and the corresponding constructed probability density functions in linear and logarithmic scale. The intervals of both histogram bins are in geometric progression; hence, the histogram bars are of increasing and equal width, respectively.

Models for Distribution Functions

Besides normal, log-normal, and gamma distributions, several more specific distribution functions have proven useful in practice. The most important are:

- Rosin-Rammler-Sperling-Bennett (RRSB) distribution, also called Weibull distribution, with scale parameter x' and shape parameter n ,

$$Q_3(x) = 1 - \exp\left(-\left(\frac{x}{x'}\right)^n\right) \quad \text{for } x \geq 0, \quad (12)$$

where x' satisfies $Q_3(x') = 1 - \frac{1}{e}$ and typically $0.5 \leq n \leq 2$.

- Gates-Gaudin-Schuhmann (GGS) distribution with scale parameter $x_{\max}$ and shape parameter n :

$$Q_3(x) = \left(\frac{x}{x_{\max}}\right)^n \quad \text{for } 0 \leq x \leq x_{\max} \text{ and } n > 0. \quad (14)$$

- Gaudin-Melloy (GM) distribution with scale parameter $x_{\max}$ and shape parameter n :

$$Q_3(x) = 1 - \left(1 - \frac{x}{x_{\max}}\right)^n \quad \text{for } 0 \leq x \leq x_{\max} \text{ and } n > 0. \quad (15)$$

These distributions are frequently used, but experience shows that often the fit at the tails is of low quality. To overcome such shortcomings, tailor-made model adjustments are used, e.g., the three parameter RRSB distribution function, where an additional term $x_{\min}$ is used.

Analysis Methods

Grain sizes are measured in different approaches and methods, depending on the fineness and accessibility of the grains. The following are brief outlines of such methods; for details we refer to Allen (1990), Bernhardt (1994), and Higgins (2006).

Sieve Analysis

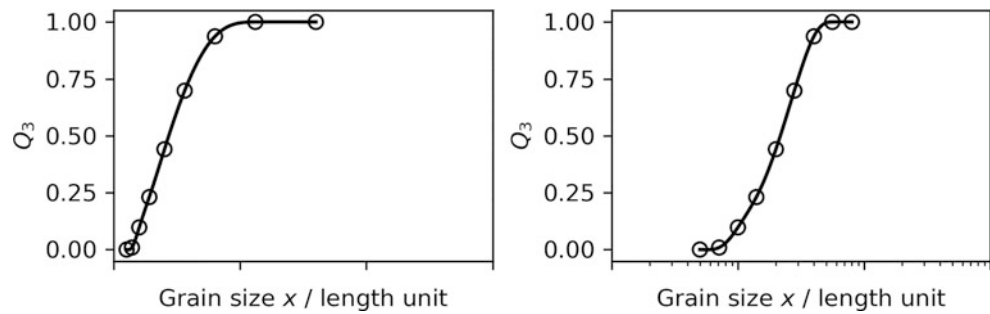
Sieve analysis is a popular measurement approach for loose grains in the range of 5–125 mm. During sieving the material is separated into fractions, the masses of which are determined and compiled as histograms. The analysis can be done consecutively with single sieves or in parallel with a set of sieves by hand or by machine with different setups, mostly depending on the material properties. The loose grains

Grain Size Analysis, Table 2 Types of quantity for grain size distributions

r	Definition
0	Number
1	Length
2	Area
3	Volume or mass

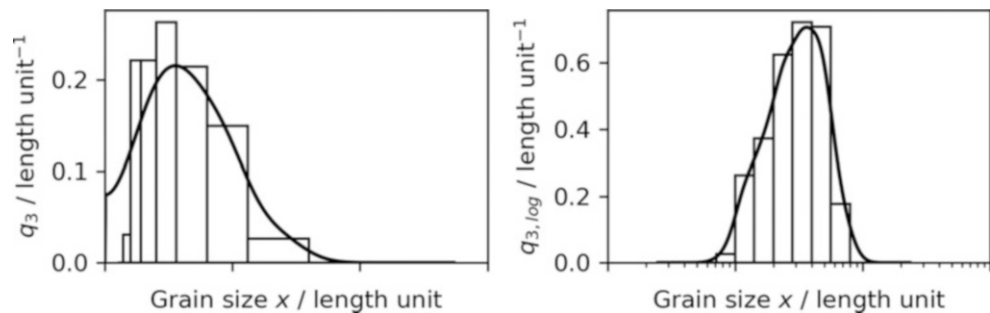
Grain Size Analysis,

Fig. 1 Cumulative grain size distribution function $Q_3(x)$ plotted against linear and logarithmic x -scale. Since the size scale can be in arbitrary dimensions, no explicit length unit was used for these example plots



Grain Size Analysis,

Fig. 2 Histograms and corresponding density functions obtained by kernel estimators in linear and logarithmic scale



are transported through the mesh openings by inertia forces, by gravity and/or flow forces, or by hand if very large grains are sieved.

Imaging Particle Analysis

This complex of analysis methods is based on images taken with cameras, microscopes, X-ray detectors, etc. The grains can be loose like fractured rock as well as embedded grains like crystals in a mineral microstructure. While microstructures are analyzed in a static or quasi static setups, for example, under a microscope or in a tomograph, images of loose grains can also be taken in a dynamic situation, like sand in a flow (Fig. 3).

Based on the kind of images it can further be distinguished between approaches for the analysis of projections, sections, or spatial images like those resulting from tomography. In theory, the size resolution of imaging particle analysis is only limited by the resolution of the imaging method.

For deriving grain sizes of projections or sections of original three-dimensional objects, size parameters can be calculated directly from measured parameters like area or perimeter of the outlines. However, those are fraught with uncertainties, since the spatial character of the measured objects is only partly taken into account. To overcome this, stereological methods for either sections or projections can be applied. In this context, stereology offers a comprehensive toolbox to calculate spatial characteristics based on specific assumptions on the grain shape and orientation. In contrast, it is possible to measure spatial characteristics directly on spatial images by means of tomography without further assumptions, which is a great benefit. However, tomography is difficult when objects with similar physical properties, like crystals of the same material, are touching each other. (Since tomography methods measure the interactions between penetrating waves and the matter of the objects and X-ray absorption and refraction are

isotropic for all materials, it is hardly possible to distinguish adjoining objects if they have similar properties (Baker et al. 2012).) Furthermore, tomography needs rather complex measurement instruments and produces huge datasets, and the evaluation of tomography data needs special statistical methods (Ohser and Schladitz 2006).

As an example, for the application of stereological methods, grain measurement for the quartz grains in Fig. 4 is discussed. For the analysis of such planar sections there exist some solutions in the mineral-related literature. One, see Higgins (2000, 2006), assumes that the grains are parallelepipeds or ellipsoids, whereas another, see Popov et al. (2014), assumes ellipsoids. In the planar section the areas of the section profiles are measured. By means of approximate stereological methods then grain shape is characterized according to the assumed geometry and some linear size characteristic, which is the diameter of volume-equivalent sphere for Popov and maximum length or major axis of equivalent ellipsoid in the Higgins approach. The Popov method is parametric and uses the log-normal distribution. Figure 4 shows the solutions for both approaches for the quartz grains.

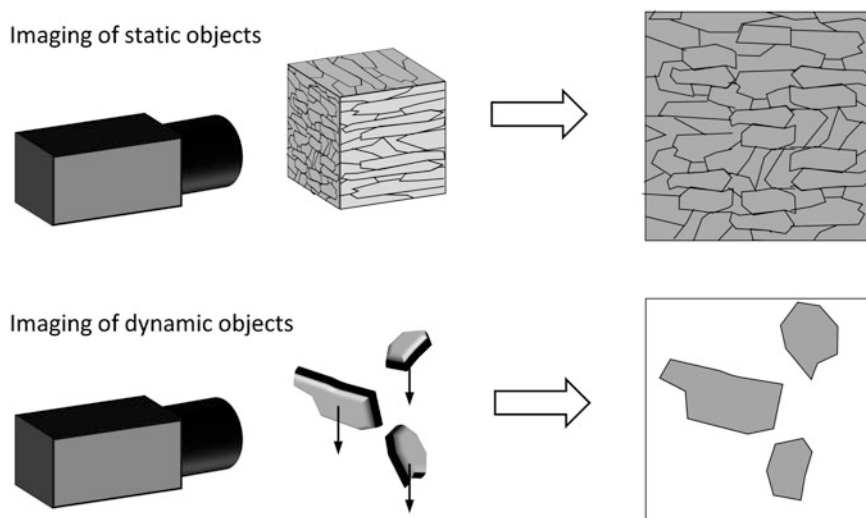
Light Scattering Methods

If a loose grain is irradiated with light, several interactions like light scattering occur. Since the intensity of the scattered light is a function of grain size, this principle can be used for grain size measurement.

In this context, there exist two main measuring principles. The first is direct measurement of scattered light, which allows measurement of loose grains in the narrow size range between 0.1 μm and 50 μm . The second is measurement of extinction caused by a loose grain in a light beam. This method of measurement has a lower limit of approximately 1 μm but is theoretically capable of measuring grains up to the

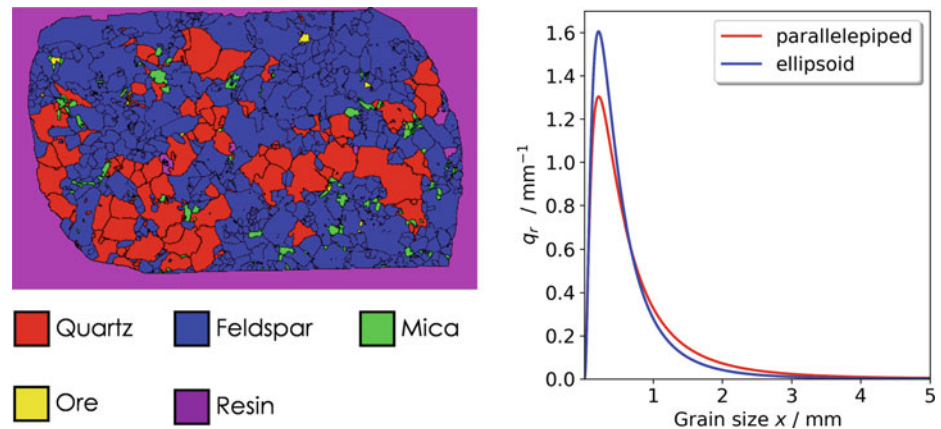
Grain Size Analysis,

Fig. 3 Imaging particle analysis of a static microstructure and of moving loose grains. For grain size analysis, either direct measures of the planar projections can be used or spatial length measures are calculated with stereological tools



Grain Size Analysis,**Fig. 4** Thin section

(35 mm × 31 mm) of a granite and corresponding empirical log-normal distributions for grain size of quartz



size of centimeters. For both measuring principles, the grains have to be dispersed into a gas or a fluid.

For measuring grain size with light scattering, a variety of solutions with different requirements and specifications are available. To measure the size of solid grains, solutions with absorbance measurement are commonly used. There, the material is dispersed in a liquid or a gas and transported through a measurement cell where the grains are measured automatically.

Sedimentation Analysis

Since the settling behavior of loose grains in a fluid largely depends on their size, the sedimentation principle provides a basis for grain size analysis. Both, gravitation and centrifugal force can be used. Generally, two main approaches are distinguished. For the incremental method, the rate of change of density or concentration is used, while the cumulative approach uses the settling rate of loose grains. While measures based on the first method are fast, an advantage of cumulative measurement is that it works also with small amounts of material, which reduces the error due to interactions between the settling grains.

Sedimentation analysis allows various measurement setups of different complexity (Allen 1990; Bernhardt 1994). It is mainly used for the analysis of fine grains in the lower micron range. If sedimentation in a laminar flow field is used, the Stokes diameter d_{St} serves as reference size:

$$d_{St} = \sqrt{\frac{18\eta}{(\rho_g - \rho_f) \cdot g}} \cdot w_s, \quad (11)$$

where η – dynamic viscosity of the fluid, ρ_g – grain density, ρ_f – fluid density, g – acceleration due to gravity or centrifugal force, w_s – settling velocity of the loose grain. This implies that the grain Reynolds number calculated for the Stokes diameter is below the value of 0.25.

Particle imaging and light scattering methods benefit from advances in computer technology and optoelectronics, whereas sedimentation analysis has lost some of its importance. However, it is still considered one of the most important methods of grain size analysis.

Summary

The size of grains is measured by various methods, which are adapted to the nature of the material. There is a tendency to use more and more computer techniques. Grain size is defined in different ways, in dependence of the material and the method of measurement. The data obtained are analyzed statistically, using various special distribution functions. The statistics is difficult for the tails of distributions, for very small or very large grains. For embedded grains stereological methods are used.

Cross-References

- Cumulative Probability Plot
- Pore Structure
- Stereology

Bibliography

- Allen T (1990) Particle size measurement. Powder technology. Springer Netherlands, Dordrecht
- Baker DR, Mancini L, Polacci M, Higgins MD, Gualda G, Hill RJ, Rivers ML (2012) An introduction to the application of X-ray microtomography to the three-dimensional study of igneous rocks. Lithos 148:262–276. <https://doi.org/10.1016/j.lithos.2012.06.008>
- Bernhardt C (1994) Particle size analysis: classification and sedimentation methods. Springer Netherlands, Dordrecht
- Higgins MD (2000) Measurement of crystal size distributions. Am Mineral 85:1105–1116. <https://doi.org/10.1515/am-2000-8-901>

- Higgins MD (2006) Quantitative textural measurements in igneous and metamorphic petrology. Cambridge University Press, Cambridge
- Ohser J, Schloditz K (2006) 3D images of materials structures. Wiley, Weinheim
- Popov O, Lieberwirth H, Folgner T (2014) Properties and system parameters – quantitative characterization of rock to predict the influence of the rock on relevant product properties and system parameters – part 1: application of quantitative microstructural analysis. *AT Miner Process* 55:76–88
- Randolph AD, Larson MA (1988) Theory of particulate processes: analysis and techniques of continuous crystallization, 2nd edn. Academic, San Diego
- Rumpf H, Scarlett B (1990) Particle technology. Springer Netherlands, Dordrecht

Graph Mathematical Morphology

Rahisha Thottolil and Uttam Kumar

Spatial Computing Laboratory, Center for Data Sciences,
International Institute of Information Technology Bangalore
(IIITB), Bangalore, Karnataka, India

Definition

Mathematical morphology provides a set of filtering and segmenting tools for image analysis. Here, an image space is considered as a graph where the connectivity of the pixels/grids is taken into account. A simple graph is a collection of vertices and edges, and it is denoted as $G = (X^*, X^\times)$, where X^* is a non-empty set of vertices and $X^\times$ (representing adjacency relation among the vertices) is composed of pairs of distinct vertices from X^* called edges. In other words, each element of X^* is called a vertex or a node point of a graph G , and each element of $X^\times$ is called an edge of a graph G . A simple graph is an undirected graph where each edge connects two different vertices and no two edges connect the same pair of vertices. Graph mathematical morphology (GMM) is a systematic theory and has been developed on these graph spaces (G). GMM extracts structural information from simple and small pre-defined structuring graphs by constructing morphological operators such as openings and closings. However, morphological transformations by GMM are restricted by their graph structure like the subsets of vertices, the subsets of edges, and the sub-graphs of graphs.

Introduction

The term morphology is widely used in the field of biology that deals with the shape and structure of plants and animals. However, in this chapter, we outline the important concept of mathematical morphology (MM) in the context of graphs that

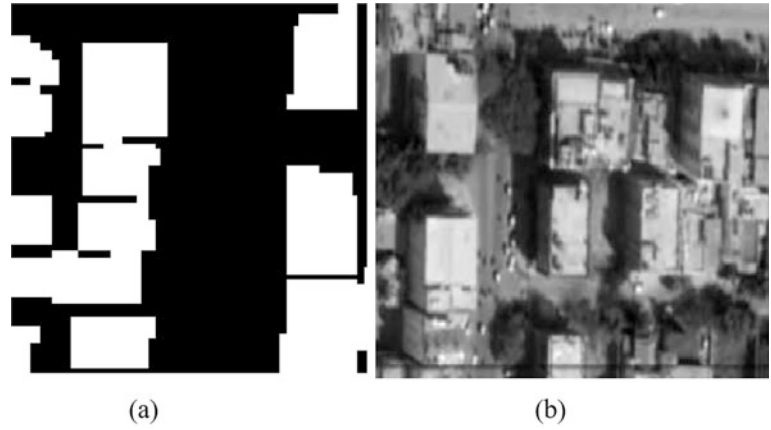
has wide applications in digital image processing. MM was introduced in mid-1960s by two researchers Georges Matheron and Jean Serra (Youkana 2017) for investigating the structures present in crystals and alloys. The theoretical basis of a morphological operator was derived from set theory and integral geometry (Heijmans and Ronse 1990). Initially, MM was applied on binary, gray-scale and color images. Luc Vincent (1988) extended the application of MM from image space to graph space. Later, MM operators were applied on multivalued data such as in remote sensing and geoscience applications, medical image analysis, and astronomy to extract spatial features. This chapter aims to present a brief introduction of classical morphology operators on two-dimensional binary and gray-scale images and discuss graph-based mathematical morphology (GMM). In the following section, important terminologies used in morphological methods are defined.

A 2D binary image is assumed to be represented in the form of a point set (vector in 2D space), where each pixel is an image point with value 0 or 1, whereas in the case of a gray-scale image, the value of an image point is an ordered set of gray levels (e.g., 8-bit image value ranges from 0 to 255). Figure 1 (a) and (b) shows examples of a discrete binary and gray-scale images, where each image pixel (point) is represented as a grid.

In practice, the points belong to an area of interest (building footprint illustrated in white pixels) in Fig. 1a and are represented as a point set X . It means, $X = \{(x_0, y_0), (x_0, y_1), \dots, (x_i, y_i)\}$, and all the points (x_i, y_i) belong to the location where the area of interest/object is present, which is a simple morphological description of an image. Similarly, the background pixels (black) can also be represented in the form of a point set, which is the complement of $X(X^c)$. In case of gray-scale, it is a multidimensional (n) image and is represented as a set of points, $X \subset E^n$, where E^n is n -dimensional Euclidian space. However, Fig. 1b illustrates the concept of E^n on a 2D gray-scale image, where each grid denotes a single pixel, then $(n - 1)$ coordinates represent the spatial domain, and n^{th} coordinate represents the value of the function.

Morphological operation on an image involves a relation of the point set X with another small structural point set (B). Generally, B is a small subset of a binary image (called structuring elements: $SE(B)$ defining the neighborhood of the points) that uses probes with well-defined shape and size to extract specific information (Heumans et al. 1992). Here, one of the pixels represent an origin (usually center pixel), and its size is based on the dimension of matrix while shape depends on the position of values ones or zeroes (Youkana 2017). It is to be noted that the selection of SE is done by considering the image geometry and prior knowledge of spatial properties of an object. Usually, flat SE are used in

Graph Mathematical Morphology, Fig. 1 Example of (a) a binary image and (b) a gray-scale image



case of 2D image analysis, and non-flat SE are used for higher dimensional images. Once the image and $SE(B)$ are represented in the form of sets, then MM operators on images are basic set operations (usually nonlinear) on the point sets. In practice, $SE(B)$ moves systematically over the input image(X), and the performance of operation depends on the type of SE s and translation operations applied. To learn more about the theoretical aspects of classical morphology, we need to define a complete lattice notion which is the fundamental source of MM operations.

Fundamental Operators: Dilation and Erosion

The fundamental morphological operators are dilation (δ) and erosion (ε) (Heijmans and Ronse 1990). The dilation operation is implemented in terms of union (the Minkowski addition) of two sets. We denote the dilation of a set X over a $SE(B)$ by $\delta(X)$ as shown in Eq. (1) in 2D space ($\mathbb{Z}$).

$$\begin{aligned}\delta(X) &= X \oplus B = \{p \in \mathbb{Z}^2 / p = x + b, x \in X, b \in B\} \\ &= \cup_{b \in B} X_b\end{aligned}\quad (1)$$

The erosion operation of a set X by $SE(B)$ - $\varepsilon(X)$ is defined in terms of intersection (the Minkowski subtraction) operation of two sets as given by Eq. (2), where $X-b$ denotes the translation of X by $-b$.

$$\begin{aligned}\varepsilon(X) &= X \ominus B = \{p \in \mathbb{Z}^2 / (p + b) \in X, b \in B\} \\ &= \cap_{b \in B} X_{-b}\end{aligned}\quad (2)$$

In general, after applying dilation operation on an image, the regions of an object area expand and noisy pixels are removed from the object region. In contrast, erosion is an inverse operation of dilation where the boundary of an object gets eroded. Interestingly, both the operators are dual by complementation and can be applied at various levels such as during pre-processing, analysis, and post-processing of an

image. Dilation of an image is always commutative with $SE(B)$ and always associative with different SE . To compensate for the distraction of object shape in image processing, we use the combination of these basic MM tools with the same $SE(B)$. The implementation of dilation and erosion on gray-scale image (F) consists of finding the maximum and minimum of original image (F) respectively, as given by Eqs. (3) and (4).

$$\delta_B(F) = \max_{b \in B} [F(x + b)] \quad (3)$$

$$\varepsilon_B(F) = \min_{b \in B^v} [F(x + b)] \quad (4)$$

In case of dilation, the algorithm finds the maximum value of the original image within the scope of the SE , whereas erosion consists of minimum value of F . Erosion and dilation operations are considered as adjunction pairs of basic morphological operators (δ, ε). Most of the morphological algorithms are derived from these two basic elementary operations.

Morphological Filtering: Opening and Closing

Two main filtering operators derived from the composition of dilation and erosion are opening ($X \circ B$) and closing ($X \bullet B$). Thus, the opening of X by B is an anti-extensive ($\gamma(x) \leq x$) filtering function. Opening and closing filters on a binary image are expressed as follows:

$$\gamma_B(X) = X \circ B = (X \ominus B) \oplus B \quad (5)$$

$$\varphi_B(X) = X \bullet B = (X \oplus B) \ominus B \quad (6)$$

The opening operation on an image (X) by $SE(B)$ is defined as the erosion of X by B followed by a dilation of the results by the same SE . Likewise, the closing operation on an image X is the dilation of X by B followed by an erosion of the result by

B , which is extensive ($x \leq \varphi(x)$). Both the filters are complementary to each other. Opening operations are useful for smoothening the contour of the objects present in an image and to remove the long narrow gaps. Further, closing fuses narrow gaps and removes small patches by filling the contour breaks. While applying a closing filter, we can remove all the internal noise present in the region, and by using an opening filter, the external noise is removed from the background region of an image.

Composing closing and opening operations with increasing size of SE is called an alternate sequential filter (ASF). These MM algorithms have practical use cases in image processing such as extracting boundary of an object, region filling, extracting connected components, retrieving convex hull information from high-level images, thinning and thickening operations, and enhancing the object structure by skeletonization. During pre-processing, applications such as filtering of salt and pepper noise, shape simplification (i.e., complex structured object broken into a number of simple structures), extracting image components such as segmentation of an image by using object shape, quantification of geometric features (area, perimeter and axis length) of an object present in an image, and object masking, etc. are possible. Efficient filters can be obtained for image processing by various compositions and iterations of basic morphological operators.

Basic Concept of Graphs

In this section, we describe how to extend the classical morphological operation from image to graph spaces. Generally, the image space is considered as a graph whose vertex set is generated from pixels, and the edges represent the local pixel adjacency relation or Euclidian distance between the pixels.

In a 4-pixel adjacency graph, the weight of an edge is the distance between the neighboring pixels, and to model color images, the edge weights represent the similarity or dissimilarity between the neighboring pixels (Danda et al. 2021). Edge weight is computed by taking the absolute Euclidian

difference between the adjacent pixels and it is called discrete gradient. These edge-weighted graph models are useful for image segmentation and filtering. The objects that appear in an image have structural information, therefore, MM operators can be applied on more complex data structures like graphs and hypergraphs. For example, in the field of transportation, road networks are represented as a planar graph. In such cases, the intersection of roads and endpoints is represented as vertices and road segments correspond to edges. The distance between the vertices is some kind of a measure and the graph representation of the road network describes its spatial pattern. The basic concept of a graph in the form of a road network is shown in Fig. 2 (a) and (b) where vertices are represented by red dots and the edges by yellow line segments.

In GMM operators, SE are extracted from the graphs using predefined probes (structuring graphs) (Heijmans et al. 1992). Similar to the SE , structuring graphs are also small as compared to the input graph. Cousty (2018) proposed a large collection of morphological operators acting on graphs such as dilation, erosion, opening and closing, granulometries, and alternate sequential filters. Here, the input operands and the result of the operators were both considered as graphs (Cousty et al. 2013).

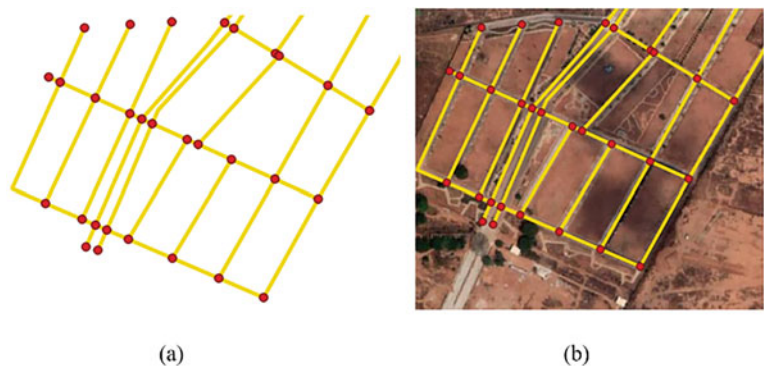
MM Operators on Graphs: Dilation and Erosion

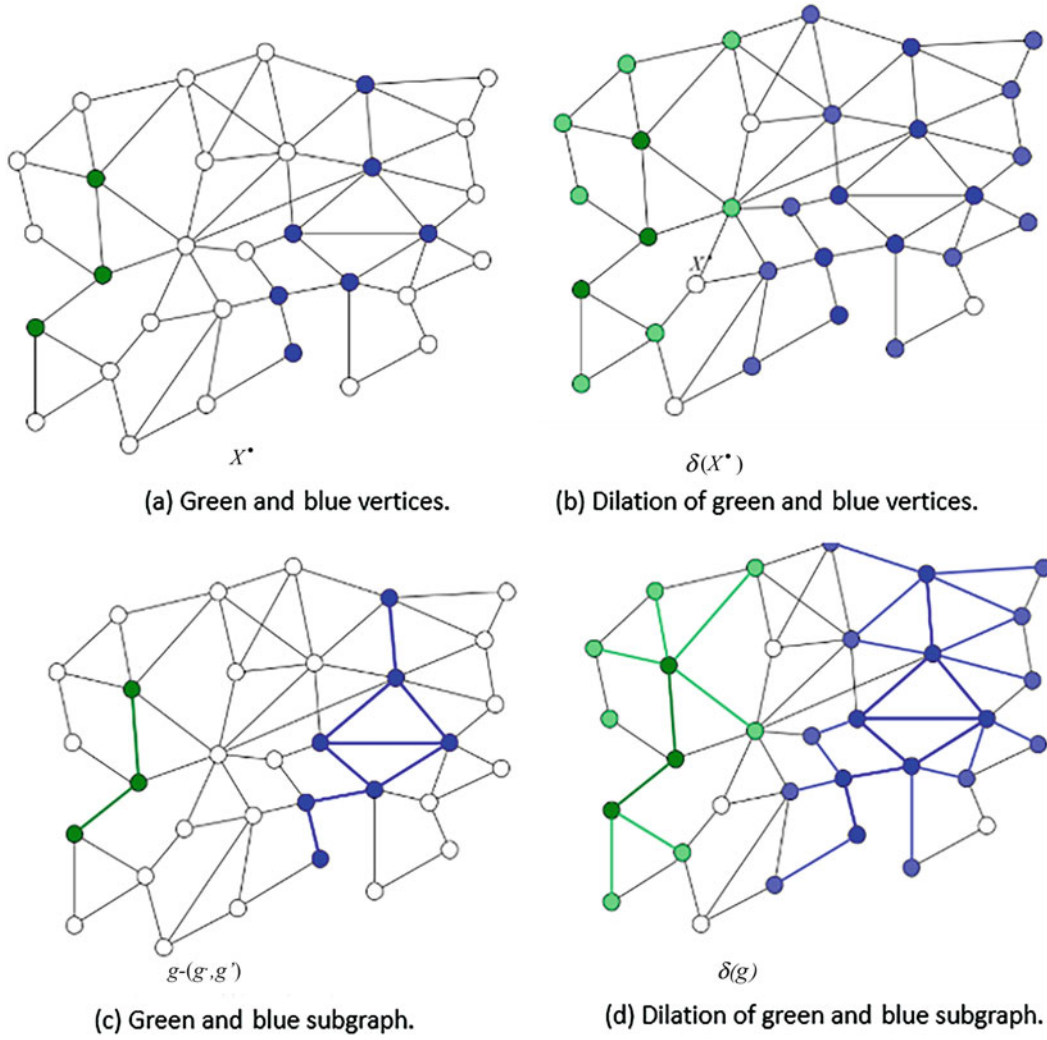
Here, operands are considered as a graph and the basic morphological operators are defined. Complex structures and relations with neighboring structures present in an image are represented by simple (unweighted) graphs.

A dilation operator on a graph with the set of vertices X^* is illustrated in Fig. 3a, b (Cousty 2018; Youkana 2017). The dilation operator from the set of vertices expands to all the neighboring vertices which are connected through edges. Dilation operator on a graph with subgraphs as shown in Fig. 3 (c) and (d) shows that dilation enrich by expanding the set of elements with the adjacent and connected vertices and edges.

Graph Mathematical Morphology, Fig. 2 (a)

Undirected road network represented as a simple graph, (b) road network overlaid on Google Earth image





Graph Mathematical Morphology, Fig. 3 Illustration of dilation operation on set of vertices (a–b) and subgraphs (c–d) (Cousty 2018)

The graph space $G = (X^*, X^\times)$, the set of x^* is the subset of X^* , $x^\times$ is the subset of $X^\times$, and $g = (g^*, g^\times)$ is the subgraph of G from the complete lattice. Four elementary building blocks on graphs are used to derive a set of edges from a set of vertices and a set of vertices from a set of edges (Cousty 2018). The dilation and erosion operators of vertices (δ^* & ε^*) map from $g^\times$ to g^* , and the operators of edges ($\delta^\times$ & $\varepsilon^\times$) map from g^* to $g^\times$ as illustrated in Fig. 3.

$$\delta^*(X^\times) = \{x \in G^* \mid \exists \{x, y\} \in X^\times\}, \text{ for any } X^\times \subseteq G^\times \quad (7)$$

$$\varepsilon^*(X^\times) = \{x \in G^* \mid \forall \{x, y\} \in G^\times, \{x, y\} \in X^\times\}, \text{ for any } X^\times \subseteq G^\times$$

(8)

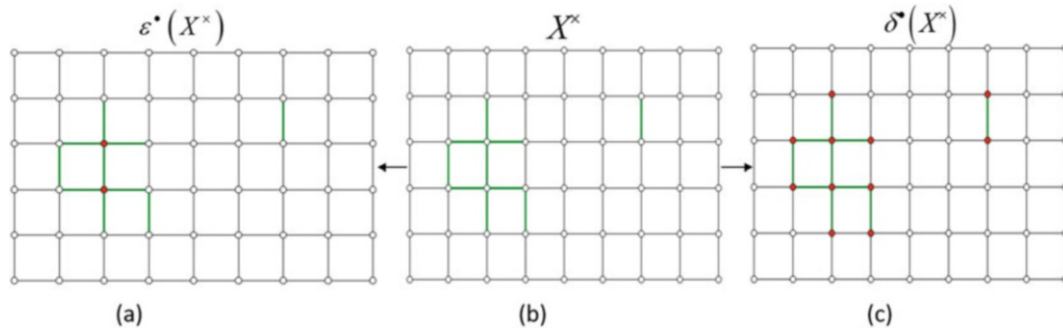
In Eq. 7, the dilation operator maps to any set of edges $X^\times$ denoted as $\delta^*(X^\times)$, which results in all the set of vertices that belong to an edge $X^\times$. The erosion operator $\varepsilon^*(X^\times)$ maps to any

set of edges $X^\times$ (Eq. 8) that results from the set of vertices that are completely covered by edges $X^\times$. Resultant graphs are shown in Fig. 4a, c.

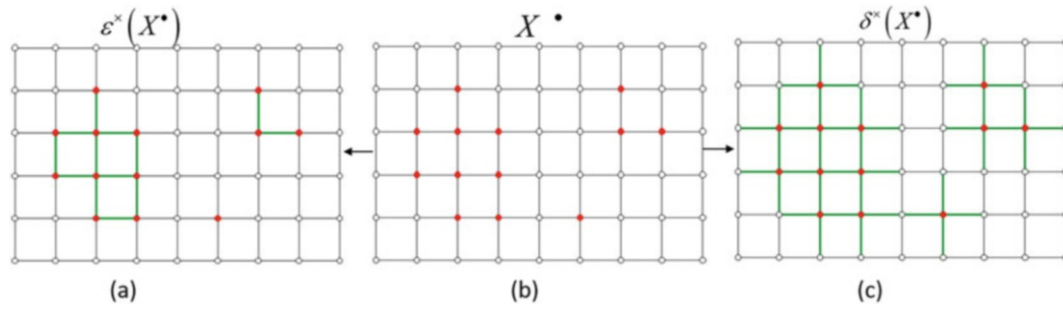
The erosion operator $\varepsilon^\times$ maps to any set of vertices X^* , the set of all edges whose two extremities are in X^* (Eq. 9). The dilation operator $\delta^\times$ maps to any set of vertices X^* , the set of all edges that have at least one extremity in X^* (Eq. 10). Resultant graphs are shown in Fig. 5a, c.

$$\varepsilon^\times(X^*) = \{\{x, y\} \in G^\times \mid x \in X^* \text{ and } y \in X^*\}, \text{ for any } X^* \subseteq G^* \quad (9)$$

$$\delta^\times(X^*) = \{\{x, y\} \in G^\times \mid x \in X^* \text{ or } y \in X^*\}, \text{ for any } X^* \subseteq G^* \quad (10)$$



Graph Mathematical Morphology, Fig. 4 Operators acting on the lattice from g^* to g^* (Cousty 2018)



Graph Mathematical Morphology, Fig. 5 Operators acting on the lattice from g^* to g^* (Cousty 2018)

The dilation operator commutes with union set operation, and the intersection operation is erosion. MM operators on binary unweighted graphs commute with set operations as follows:

$$\delta^*(X^\times \cup Y^\times) = \delta^*(X^\times) \cup \delta^*(Y^\times), \text{ for any } X^\times, Y^\times \subseteq G^\times \quad (11)$$

$$\varepsilon^*(X^\times \cap Y^\times) = \varepsilon^*(X^\times) \cap \varepsilon^*(Y^\times), \text{ for any } X^\times, Y^\times \subseteq G^\times \quad (12)$$

$$\delta^\times(X^\bullet \cup Y^\bullet) = \delta^\times(X^\bullet) \cup \delta^\times(Y^\bullet), \text{ for any } X^\bullet, Y^\bullet \subseteq G^\bullet \quad (13)$$

$$\varepsilon^\times(X^\bullet \cap Y^\bullet) = \varepsilon^\times(X^\bullet) \cap \varepsilon^\times(Y^\bullet), \text{ for any } X^\bullet, Y^\bullet \subseteq G^\bullet \quad (14)$$

where $\delta^\times$ and δ^* are called vertex-edge dilation and edge-vertex dilation. Similarly, $\varepsilon^\times$ and ε^* are called vertex-edge erosion and edge-vertex erosion, respectively. The elementary vertex dilation ($\delta(X^\bullet) = \delta = \delta^* \circ \delta^\times$) and edge dilation ($\delta(X^\times) = \delta = \delta^\times \circ \delta^*$) by composition of the vertex-edge dilation and edge-vertex dilation. Likewise, vertex erosion ($\varepsilon(X^\bullet) = \varepsilon = \varepsilon^* \circ \varepsilon^\times$) and edge erosion ($\varepsilon(X^\times) = \varepsilon = \varepsilon^\times \circ \varepsilon^*$) are obtained by the composition of the vertex-edge erosion and edge-vertex erosion. In case of morphological analysis on

gray-scale images, intensity value is also considered; therefore, MM operators are proposed on weighted graphs. The major applications of gray-scale morphology are contrast enhancement, texture extraction, edges detection, and thresholding. Definitions involved in gray-scale graph morphological operations are shown below:

Let $F^\bullet \in I(X^\bullet)$ and $F^\times \in I(X^\times)$; then MM operators are

$$\delta^*(F^\times)(x) = \max\{F^\times\{x, y\} \mid \{x, y\} \in X^\times\}, \forall x \in X^\bullet \quad (15)$$

$$\varepsilon^\times(F^\bullet)\{x, y\} = \min\{F^\bullet(x), F^\bullet(y)\}, \forall \{x, y\} \in X^\times \quad (16)$$

$$\varepsilon^*(F^\times)(x) = \{F^\times\{x, y\} \mid \{x, y\} \in X^\times\} \forall x \in X^\bullet \quad (17)$$

$$\delta^\times(F^\bullet)\{x, y\} = \max\{F^\bullet(x), F^\bullet(y)\}, \forall \{x, y\} \in X^\times \quad (18)$$

Grayscale image dilation involves assigning the maximum value over the neighborhood of the SE . In contrast, erosion involves assigning minimum value over the neighborhood of the SE (Cousty et al. 2013).

Morphological Filters on Graphs

Graph-based morphological filters such as opening and closing were obtained from the composition and iteration of dilation and erosion either by applying on set of vertices and

set of edges or subgraphs of graph G . The number of iteration operations (also called parameter size denoted as ' λ ') is important in opening and closing filters. The value of λ depends on the size of the features to be preserved or removed (Youkana 2017). Filtering operators depend on the order of basic MM operators, and their complexity corresponds to the value of λ and the size of the graphs.

We denote opening and closing on vertices by γ , ($\gamma = \delta \circ \varepsilon$) and ϕ , ($\phi = \varepsilon \circ \delta$), respectively. Similarly, opening and closing on edges are denoted by Γ , ($\Gamma = \Delta \circ E$) and Φ , ($\Phi = E \circ \Delta$), respectively. Combination of these operators gives opening ($[\gamma, \Gamma]$) (Eq. 19) and closing ($[\phi, \Phi]$) (Eq. 20) on a graph G .

$$[\gamma, \Gamma] = [\gamma(X^\bullet), \Gamma(X^\times)] \quad (19)$$

$$[\phi, \Phi] = [\phi(X^\bullet), \Phi(X^\times)] \quad (20)$$

We can deduce the operators on the subgraph of graph called half-opening and half-closing (Najman and Cousty 2014). The series of these morphological filters are called granulometries. Furthermore, an alternate sequence of filters can be derived from intermixing of opening and closing with the increasing size of structuring graph. It has been found that GMM operators are more efficient than image-based operators (Youkana 2017). However, a major drawback of graph-based image analysis is the images with a large number of pixels that have very high asymptotic complexity (Danda et al. 2021).

Application of MM and GMM

Graph-based MM operations have been used in applications involving quantitative description of spatial relationship between objects in images. Heijmans and Vincent (1992) studied the heterogeneous media at a macroscopic level based on the information on their microstructure by GMM. Furthermore, GMM filtering on preprocessing of medical images with higher accuracy among existing operations were demonstrated by Cousty (2018). The usage of GMM based ASF has proved to render minimum mean square error for noise removal (Cousty et al. 2013; Youkana 2017). GMM models have also been applied successfully for various image segmentations and filtering applications including astronomical images (Danda et al. 2021) and for segmentation of 3D images from multimodal acquisition devices (Grossiord et al. 2019). However, application of GMM in geospatial data analysis is a field of ongoing research. Hence, there is a tremendous scope of applied GMM in remote sensing data analysis.

Summary

The concept of morphological techniques discussed in this chapter constitutes a significant tool in the image-based geoscience data analysis. The main purpose of mathematical morphology operation is the quantitative description of spatial features of the objects, specifically, extracting meaningful information from noisy images. Dilation and erosion are the fundamental morphological operators, and all other morphological algorithms are derived from these primitive functions. In the first part, MM operators on binary and gray-scale image space were discussed. In the second part, we focused on the framework of MM operators on graph space. Finally, we summarized a few studies in the field of GMM.

Cross-References

- Binary Mathematical Morphology
- Grayscale Mathematical Morphology
- Mathematical Geosciences
- Mathematical Morphology

Bibliography

- Cousty J (2018) Segmentation, hierarchy, mathematical morphology filtering, and application to image analysis. Doctoral dissertation, Université Paris-Est
- Cousty J, Najman L, Dias F, Serra J (2013) Morphological filtering on graphs. *Comput Vision Image Unders* 117(4):370–385
- Danda S, Challa A, Sagar BSD, Najman L (2021) A tutorial on applications of power watershed optimization to image processing. *Eur Phys J Spec Topics* 230:1–25
- Grossiord E, Naegel B, Talbot H, Najman L, Passat N (2019) Shape-based analysis on component-graphs for multivalued image processing. *Math Morphol Theory Appl* 3(1):45–70
- Heijmans HJ, Ronse C (1990) The algebraic basis of mathematical morphology I. Dilations and erosions. *Comput Vision Graphics Image Proc* 50(3):245–295
- Heumans HJAM, Nacken P, Toet A, Vincent L (1992) Graph morphology. *J Visual Commun Image Represent* 3(1):24–38
- Heijmans HJ, Vincent L (1992) Mathematical morphology in image processing, E. Dougherty, editor. Marcel-Dekker, New York, pp. 171–203.
- Najman L, Cousty J (2014) A graph-based mathematical morphology reader. *Pattern Recog Lett* 47:3–17
- Vincent L (1988) Mathematical morphology on graphs. *Proceeding SPIE 1001, visual communications and image processing '88: third in a series*. <https://doi.org/10.1117/12.968942>.
- Youkana I (2017) Parallelization of morphological operators based on graphs. Doctoral dissertation, UNIVERSITE DE MOHAMED KHIDER BISKRA).

Grayscale Mathematical Morphology

Pravan Danda¹, Aditya Challa¹ and B. S. Daya Sagar²

¹Computer Science and Information Systems, APPCAIR, Birla Institute of Technology and Science, Pilani, Goa, India

²Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, Karnataka, India

Definition

Gray-scale morphology encompasses all operators from mathematical morphology (MM) on gray-scale images. MM operators are defined on complete lattices. Gray-scale images constitute a specific lattice. This allows the definition of the MM operators to be simplified. In this entry we shall give an overview of different gray-scale morphological operators. Detailed explanation of these operators can be found in their respective chapters and the books Najman and Talbot (2013), Soille (2004), and Dougherty and Lotufo (2003).

A gray-scale image is represented as a function, $f: E \rightarrow \{0, 1, 2, \dots, 255\}$ where E is either $\mathbb{R}^2$ or $\mathbb{Z}^2$. If the domain is $\mathbb{Z}^2$ it is referred to as a discrete image.

Two Kinds of Structuring Element

Recall that the structuring element in the gray-scale case is another gray-scale image with restricted domain. The structuring element is also referred to as structuring function in few places. There are two kinds of structuring elements based on the values the gray-scale image can take – *Flat* structuring element takes only a value 0 while *Non-flat* structuring element can take any positive values.

Basic MM Operators on Binary Images

The following are the basic operators from mathematical morphology adapted to gray-scale images.

1. *Dilation*: Given a gray-scale image f and a structuring element g , the dilation is defined as

$$\delta_g(f) = (f \oplus g)(x) = \sup_{y \in E} \{f(y) + g(x - y)\} \quad (1)$$

2. *Erosion*: Given a gray-scale image f and a structuring element g the erosion is defined as

$$\varepsilon_g(f) = (f \ominus g)(x) = \inf_{y \in E} \{f(y) - g(x - y)\} \quad (2)$$

3. *Opening*: Given a gray-scale image f and a structuring element g the opening is defined as

$$\gamma_g(X) = \delta_g \circ \varepsilon_g(f) \quad (3)$$

4. *Closing*: Given a gray-scale image f and a structuring element g the closing is defined as

$$\phi_g(X) = \varepsilon_g \circ \delta_g(X) \quad (4)$$

Filtering on Gray-Scale Images

A morphological filter on a gray-scale image is any operator that is increasing and idempotent and can be obtained using the dilation/erosion operators above. The opening/closing operators above form the basic filters. Several families of filters can be constructed from these operators.

1. *Granulometries*: The idea of a granulometry is similar to filtering out particle based on their sizes. Using the opening operator we can define it as.

$$\text{Granulometry} = \gamma_{ng} \circ \gamma_{(n-1)g} \circ \dots \circ \gamma_{2g} \circ \gamma_g \quad (5)$$

Not all structuring elements g are allowed. Usually flat structuring elements with the domain constrained to be a disk or line-segment is considered.

2. *Alternating Sequential Filters*: For gray-scale images where noise is spread across different scales and contains both bright and dark distortions, using alternate sequences of opening and closing would help.

$$\text{ASF} = \phi_{ng} \circ \gamma_{ng} \circ \phi_{(n-1)g} \circ \gamma_{(n-1)g} \circ \dots \circ \phi_g \circ \gamma_g \quad (6)$$

One can also start with closing instead of opening as in the above definition.

On the other hand one can define an *algebraic filter*, which is any operator that is increasing and idempotent. Examples of algebraic filters include – Area Opening, Attribute Openings, Annular Opening, Convex Hull Closing, etc.

Other Operators on Gray-Scale Images

A host of operators can be defined on binary images. A few are stated below.

1. *Geodesic Dilation/Erosion*: The geodesic operations force the output to belong to a mask f' . A single step of geodesic dilation is defined as

$$\text{Geo} - \text{Dilate}^{(1)}(f) = \delta_g(f) \wedge f' \quad (7)$$

One can suitably define geodesic erosion as well. These operators can be repeated to obtain higher order dilations or erosions.

2. *Thinning/Thickening*: One can define the notions of thinning/thickening by considering two structuring elements, respectively, for foreground/background pixels. See Soille (2004) for details.
3. *Skeletonization* involves reducing the binary image to a one-dimensional caricature to extract the most discerning features in the image.
4. *Segmentation* of gray-scale images involves partitioning the domain such that pixels within a component (of partition) belong to the same object. Several approaches to achieve this are proposed and the most commonly used one is that of *Watershed Transform*.

Summary

In this entry, elementary MM operators on gray-scale images, namely, gray-scale dilation, gray-scale erosion, gray-scale opening, and gray-scale closing are defined. Some of the popularly used operators that use these four elementary operations are then briefly described. These are granulometries, alternating sequential filters, geodesic dilation/ erosion, thinning/thickening, skeletonization, and watershed transform.

Cross-References

► [Mathematical Morphology](#)

Bibliography

- Dougherty ER, Lotufo RA (2003) Hands-on morphological image processing, vol 59. SPIE Press, Bellingham
- Najman L, Talbot H (eds) (2013) Mathematical morphology: from theory to applications. Wiley. <https://doi.org/10.1002/9781118600788>
- Soille P (2004) Morphological image analysis. Springer, Berlin/Heidelberg. <https://doi.org/10.1007/978-3-662-05088-0>

Griffiths, John Cedric

Donald A. Singer

U.S. Geological Survey, Cupertino, CA, USA



Fig. 1 John Cedric Griffiths

Biography

Professor John C. Griffiths was one of the founders of the International Association for Mathematical Geology. His impact on geology and other wide-ranging fields began in Llanelli, County Dyfed, Wales, in 1912. Griffiths earned a B.Sc. (1933), and Ph.D. (1937) degrees in petrology from the University of Wales, a Diploma of the Imperial College in petrology from the Royal College of Science, London, and Ph.D. degree in petrology from the University of London in 1940.

His recognition of the importance of practical experiences in teaching how to solve problems came partially from his 7 years (1940–1947) working for Trinidad Leaseholds Ltd. in the British West Indies. He directed more than 50 students during his following 40 years teaching in the Department of Mineralogy and Petrology of the Pennsylvania State

Donald A. Singer has retired.

University. Griffiths was recognized as a leader in applying statistics to the analysis of sediments. Between 1948 and 1962 he published many papers on improving analysis of sediments. This work culminated in his 1967 book (Griffiths 1967), *Scientific Method in Analysis of Sediments*. From this book one can see his increasing interest in modern scientific methods for problem solving such as decision theory, operations research, systems analysis, and search theory.

In 1970 with Ondrick (Griffiths and Ondrick 1970), he showed that randomly (uniformly distributed) points on a square surface could be fit with a Poisson distribution when sampled. If part of the square containing the same points was masked, only a negative binomial distribution could be fit to the points demonstrating a clustering of points. The points had not changed, only the ability to see some of them. This is not different from our perception of the clustering of mineral deposits – deposits known are mostly based on exposed rocks and exclude areas masked by younger cover. Insights like this led to his controversial proposal to grid-drill the United States to obtain unbiased data for national planning.

His teaching ability was so remarkable that the International Association for Mathematical Geosciences gives an annual award in his name for outstanding teaching in the application of mathematics to the geosciences. Students who expected to learn everything Griffiths knew soon recognized that this was impossible because he was continuing to learn and share new aspects of science every day. His students

learned applied statistics of spot, stratified and channel sampling schemes with pebble measurements in a partially layered gravel pit and in measures of size and shape of quartz grains and point counting of minerals in thin sections.

But each student was also exposed to a different set of classes throughout the university education. Classes varied for each student and ranged from industrial engineering, computer science, mineral economics, agronomy, chemistry, mathematical statistics, geology, to psychology. From these classes, students learned the commonality of problems with major differences being each discipline's language barrier. His teaching methods were unusual and by no means obvious to others. His students were taught how to define and solve scientific problems. The teaching award in his name by the International Association for Mathematical Geoscience was certainly appropriate.

Bibliography

- Griffiths JC (1967) *Scientific method in analysis of sediments*. McGraw-Hill, New York. 508 p
- Griffiths JC, Ondrick CW (1970) Structure by sampling in the geosciences. In: Patil GP (ed) *Random counts in physical science, geoscience, and business: the Penn State statistics series*. The Pennsylvania State University Press, University Park/London, pp 31–55

Harbaugh, John W.

Johannes Wendebourg
TotalEnergies Exploration, Paris, France



Fig. 1 John W. Harbaugh, Courtesy of family Harbaugh's private collection

Biography

John W. Harbaugh (born August 6, 1926, died July 28, 2019) was a professor of mathematical geology at Stanford University where he taught quantitative methods for 44 years from 1955–1999. He received his undergraduate degree in geology from the University of Kansas in 1949 and his PhD from the University of Wisconsin in 1955. His main contributions to the field of mathematical geology are in the area of geological computer modeling where he pioneered modern simulation techniques, and in resource estimation methods. In 1968, he cofounded the International Association of Mathematical Geology (IAMG), from which he received the Krumbein award in 1985.

As a field geologist, Harbaugh worked mainly on carbonate rocks but quickly became interested in computer applications. His earliest work was on trend surfaces but quickly he moved to computer simulation. In 1970, he published together with Graeme Bonham-Carter the seminal book on *Computer Simulation in Geology* which treats a wide variety of subjects from stratigraphic modeling to geomorphology, including numerical methods (Harbaugh and Bonham-Carter 1970). This visionary book came into fruition some 15+ years later in the form of Stanford's Geomath program that he founded and ran from the early 1980s until 1999 and that was dedicated to the computer simulation of geological processes. The software called SEDSIM that he and his coworkers developed pioneered modern geological forward modeling techniques using novel Lagrangian numerical methods (Tetzlaff and Harbaugh 1989). Several monographs resulted from this work, including simulation of processes of sediment transport and deposition, and fluid flow in sedimentary rocks.

In the area of resource estimation, Harbaugh applied various statistical methods (Harbaugh et al. 1977) and worked closely together with John C Davis which culminated in a 500 + page monograph on risk analysis in oil and gas exploration (Harbaugh et al. 1995). Harbaugh also was interested in Markov processes applied to geology, as they are not purely random, and applied them to spatial representations of geology (Lin and Harbaugh 1984).

Bibliography

- Harbaugh JW, Bonham-Carter G (1970) *Computer simulation in geology*. Wiley Interscience, New York. 575 pp
- Harbaugh JW, Doveton JH, Davis JC (1977) *Probability methods in oil exploration*. Wiley-Interscience, New York. 269 pp
- Harbaugh JW, Davis JC, Wendebourg J (1995) *Computing risk for oil prospects: principles and programs*. Pergamon Press, Oxford. 464 p
- Lin C, Harbaugh JW (1984) *Graphic display of two- and three-dimensional Markov computer models in geology*. Van Nostrand Reinhold, New York. 180 pp
- Tetzlaff DM, Harbaugh JW (1989) *Simulating clastic sedimentation*. Van Nostrand Reinhold, New York. 202 pp

Harff, Jan

Martin Meschede

Institute of Geography and Geology, University of
Greifswald, Greifswald, Germany



Fig. 1 Jan Harff, 2002, on board with Professor Albrecht Penck (Photo: Foto Manthey)

Biography

Jan Harff was born in Güstrow in Mecklenburg (East Germany). After his Abitur in 1961 his first attempt to begin a study of architecture at the Technical University of Berlin failed caused by the Berlin wall construction. After some years in the building sector, he finally began his study of geology at the Humboldt University in East Berlin in 1964. The centrally controlled GDR university reform of 1968 forced him to continue his studies at the University of Greifswald in the northeasternmost part of Germany, where he finished his diploma in 1963. After this, he wanted to specialize in marine geology trying to get a position as a doctoral student at the Institute of Marine Studies in Warnemünde. On political reasons, however, he was not allowed to participate in the offshore research project, because he did not get the admission of the former socialistic regime of the GDR to pass the maritime boundary (Fig. 1).

Mainly caused by this denial, Jan Harff started his career as a mathematical geologist, first with a PhD thesis on classification methods on geophysical borehole data. After a short time at the district office of geology in Rostock, he switched to the Central Institute of Earth Physics (ZIPE, now German Research Centre for Geosciences, GFZ), where he was responsible to set up a department of mathematical geology,

a branch of geosciences which developed mainly in the 1970s. The main target of his investigations was to optimize the exploration methods of oil and gas using a theoretical basin model of the North German-Polish basin.

In 1991, after the German reunification, Jan Harff was offered the position as head of the project group mathematical geology at the newly formed GFZ (which substituted the ZIPE). However, after only 3 months at the GFZ, he got the chance of his life to become the head of the section marine geology at the Leibniz Institute of Baltic Research (IOW) in Warnemünde. With this job, his dream of marine geology came true and he could finally do what he wanted to do in the beginning of his career.

As member of the Institute of Baltic Research, Jan Harff was able to sail on various cruises, mainly using the IOW research vessel Professor Albrecht Penck in the Baltic Sea. Moreover, participations and expedition leaderships on international cruises with different German and Russian research vessels were a major part of his research activities. He concentrated on coastal regions and modeled, for instance, changes in the coastlines of the Baltic Sea according to the post-glacial uplift of Scandinavia and tried to model their future development based on climatic change data. Additionally, he also investigated the coastal areas of Vietnam and Southern China.

After his retirement at the IOW in 2008, Jan Harff moved to the University of Szczecin in Poland and took over a professorship in marine geosciences. This position allowed him to proceed with his international research projects, in particular in the coastal region of the South China (von Storch and Dietrich, 2017). Together with three colleagues from different geomarine disciplines, he edited the *Encyclopedia of Marine Geosciences* which in 2017 gained the “Mary B. Ansari Best Geoscience Research Resource Work Award” of the Geoscience Information Society (GSIS). Jan Harff was several times awarded: in 1989 with the Friedrich-Stammer Prize for geology in the former GDR, in 1998 with the Krumbein Medal of the International Association for Mathematical Geology (IAMG), in 2009 with the Serge-von-Bubnoff Medal of the German Geological Society (DGG), and in 2013 with the Chinese National Prize.

References

- von Storch H, Dietrich R (2017) Jan Harff – zwischen Welten. <http://www.hvonstorch.de/klima/Media/interviews/jan.harff.pdf>. Last accessed 15 May 2020

High-Order Spatial Stochastic Models

Roussos Dimitrakopoulos and Lingqing Yao

COSMO – Stochastic Mine Planning Laboratory, Department of Mining and Materials Engineering, McGill University, Montreal, Canada

Definition

High-order spatial stochastic models refer to a general non-Gaussian, nonlinear geostatistical modeling framework that consistently characterizes complex spatial architectures and facilitates the spatial uncertainty quantification of pertinent attributes of the earth sciences and engineering phenomena predicted from finite measurements. High-order spatial statistics such as spatial cumulants characterize non-Gaussian spatial random fields and facilitate the consistent description of complex nonlinear directional spatial patterns and the multiple point connectivity of extreme values. The corresponding high-order stochastic simulation approaches are data-driven and built upon the random field theory, beyond both the conventional second-order and the multipoint statistics based simulation frameworks.

Introduction

Traditional geostatistical simulation methods utilize covariance/variogram functions to characterize the so-called second-order spatial statistics and fully define the related multi-Gaussian random field. These traditional simulation methods have two major limitations. Firstly, most of the natural attributes are not compliant with Gaussian random fields and their symmetric bell-shape probability distributions as well as their maximum entropy aspects that conflict with the structured behavior of actual geological spatial patterns. In addition, second-order spatial statistics only capture the correlations between pairs of points and hence are unable to characterize more complex spatial structures that are typical in geological phenomena, such as curvilinear features, overlaying geological events, complex nonlinear patterns, directional multiple point periodicity, and so on. An early attempt to address the above shortcomings and to eliminate Gaussian assumptions was sequential indicator simulation (Journel 1989; Journel and Alabert 1989). This was followed by the introduction of multiple-point statistics (MPS) by Prof. Andre Journel at his Stanford Centre for Reservoir Forecasting, Stanford University, as opposed to the previously used two-point second-order statistics; this led to the development of the multiple-point simulation framework (Guardiano and Srivastava 1993; Journel 1993, 2003, 2005). The first

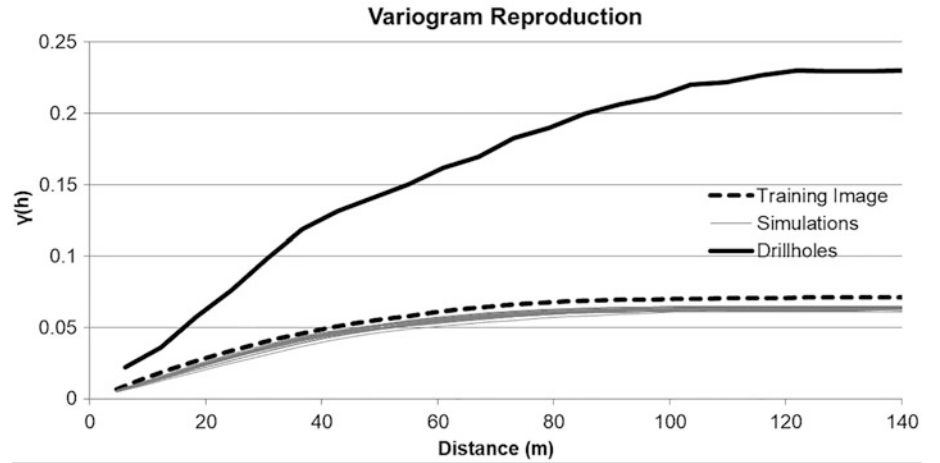
algorithm in this direction was SNESIM (Strébel 2000; Strébel 2002), a typical pixel-based MPS method, followed by several new patch-based MPS methods (Remy et al. 2009; Mariethoz et al. 2010; Mariethoz and Caers 2014; Gómez-Hernández and Srivastava 2021). Despite improvements in terms of reproducing spatial connectivity over the second-order stochastic simulation approaches, MPS methods require a training image (TI) as the source for the related inference (Guardiano and Srivastava 1993; Journel 1997, 2003, Journel and Zhang 2006) of multiple-point statistics. The term training image refers to an exhaustive image reflecting prior geological interpretations and acting as statistical analog to the underlying attributes of interest. This training-image driven nature generates a major limitation when MPS methods are applied in cases where statistical conflicts exist between the sample data available and the TI. This is a prominent issue when relatively reasonable sample data are available, as is typical in mining applications (Osterholt and Dimitrakopoulos 2007; Goodfellow et al. 2012). Figure 1 shows an example of variogram reproduction issues from a SNESIM application at a mineral deposit, where realizations reproduce the variogram of the training image but not that of the available data.

In addition, the experimentally selected multiple-point pattern (or moment) in MPS does not consider the consistency or relations between higher and lower order spatial moments over a series of orders used or facilitates data-driven connectivity of complex spatial patterns. Furthermore, while the MPS framework is more informed than those based on two-point or second-order statistics, the possibility of new approaches that are even more informed, given the wealth of information used in applications, remains to be addressed.

In recent years, a new high-order spatial modeling framework, based on measures of high-order complexity in spatial architectures, termed spatial cumulants, has been proposed. This approach not only facilitates dropping any distributional assumptions and data transformations but also develops spatial statistical models that are data-driven while capitalizing from integrating substantially more information from the underlying geological phenomena and related data. This includes complex, diverse, nonlinear, multiple-point directional or periodic patterns, advanced data-driven analytics, and the reconstruction of consistency between lower- and higher-order spatial complexity in the data used.

Cumulants are combinations of moments of statistical parameters that characterize non-Gaussian random fields. Research to date provides definitions, geological interpretations, and implementations of high-order spatial cumulants that are used, for example, in the high-dimensional space of Legendre polynomials to stochastically simulate complex, non-Gaussian, nonlinear spatial phenomena. The principal innovation, and a fundamentally challenging problem to address, is the development of mathematical models that

High-Order Spatial Stochastic Models, Fig. 1 Indicator variogram reproduction of SNESIM simulations (light gray) vs training image (dotted line) and data (solid line). (From Goodfellow et al. 2012)



move well beyond the “second-order models” to be more data-driven and consistent, as well as more informed than the previous MPS framework. The related mathematical models in turn open new avenues for the complex spatial modelling of geological uncertainty. The advantages of this work include (a) the absence of distributional assumptions and pre-/post-processing steps, e.g., data normalization or training image (TI) filtering; (b) the use of high-order spatial relations present in the available data to the simulation process (data-driven, not TI-driven), adding a wealth of additional information; and (c) the generation of complex spatial patterns to reproduce any data distribution and related high-order spatial cumulants, respecting the consistency of different high-order spatial relations. These developments define a mathematically consistent and more informed alternative to MPS, with substantial additional advantages (e.g., it is data-driven and reconstructs consistently the lower-order spatial complexity in the data used) for applications, particularly in cases where reasonably sized data sets are available. Notable is that the so-termed high-order stochastic simulation framework requires no fitting of parametric statistical models; similar to the MPS framework, high-order multiple-point relations are calculated directly from the available data sources during the simulation process, making the utilization of the related algorithms straightforward for the user. At the same time, the reproduction of related high-order spatial statistics by the resulting simulated realizations can be assessed through the cumulant maps mentioned in a subsequent section.

High-Order Spatial Cumulants

Spatial cumulants introduced in Dimitrakopoulos et al. (2010) for earth science and engineering problems are a

mathematical representation of high-order spatial statistics of a random field that quantify complex directional multiple point statistical interactions and characterize the spatial architectures of random fields. Related work in other fields includes signal processing (Nikias and Petropulu 1993; Zhang 2005; others), astrophysics (e.g. Gaztanaga et al. 2000), and image processing (Zetsche and Krieger 2001; Bouleminadjel et al. 2018).

Moments and cumulants are critical statistical concepts, given that the probability distribution of a random variable can be transformed accordingly as a moment-generating or cumulant-generating function. Let $(\Omega, \mathcal{I}, P)$ be a probability space; given a real-valued random variable, Z , as a measurable function defined on the sample space Ω with a probability density function $f_Z(z)$, the r th ($r \geq 0$) moment of Z is expressed as

$$\text{Mom}[Z, \dots, Z] = E[Z^r] = \int_{-\infty}^{+\infty} z^r f_Z(z) dz. \quad (1)$$

The moment-generating function (Rosenblatt 1985) of Z is defined by

$$M(w) = E[e^{wz}] = \int_{-\infty}^{+\infty} e^{wz} f_Z(z) dz. \quad (2)$$

The cumulant-generating function of Z is defined as the logarithm of the moment-generating function $M(w)$

$$K(w) = \ln(E[e^{wz}]). \quad (3)$$

The r th moment of Z can be obtained as the r th derivative of $M(w)$ at the origin through the Taylor expansion of the moment-generating function about the origin

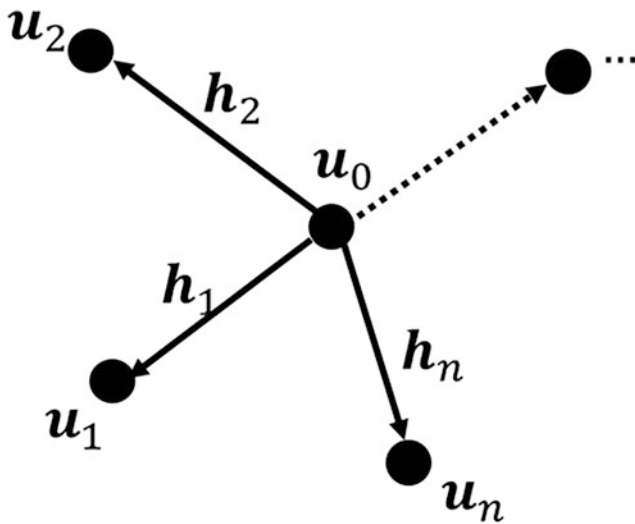
$$\begin{aligned}
 M(w) &= E[e^{wz}] = E\left[1 + wZ + \dots + \frac{w^r Z^r}{r!} + \dots\right] \\
 &= \sum_{r=0}^{\infty} \frac{w^r \text{Mom}\left[\overbrace{Z, \dots, Z}^r\right]}{r!}.
 \end{aligned} \quad (4)$$

The cumulants of Z are the coefficients in the Taylor expansion of the cumulant-generating function, $K(w)$, about the origin

$$K(w) = \ln(E[e^{wz}]) = \sum_{r=0}^{\infty} \frac{w^r \text{Cum}\left[\overbrace{Z, \dots, Z}^r\right]}{r!}. \quad (5)$$

The relationship between the cumulants and moments can be further derived by extracting the coefficients of the two Taylor series (4) and (5) through differentiation.

Spatial cumulants extend the concept of cumulants to the multivariate probability distribution of a random field endowed with more complex spatial structures, similar in a way to how the covariance function extends the second-order moments to represent two-point spatial statistics. The spatial cumulants associated with any subset of random variables of the random field $Z(\mathbf{u})$ are statistical functions of distance vectors among the locations of the related random variables. The spatial configuration of an arbitrary set of random variables $Z(\mathbf{u}_0), Z(\mathbf{u}_1), \dots, Z(\mathbf{u}_n)$, of the random field $Z(\mathbf{u})$, can be uniquely determined by the so-called spatial template with a center node noted as $Z(\mathbf{u}_0)$ and the distance vectors $\mathbf{h}_1, \dots, \mathbf{h}_n$, defined as those vectors between each node other than $Z(\mathbf{u}_0)$ to the center node (Fig. 2). The r -th spatial cumulants of random variables within a spatial template can be expressed as $c_r^z(\mathbf{h}_1, \dots, \mathbf{h}_{r-1})$. While the relations between cumulants



High-Order Spatial Stochastic Models, Fig. 2 Spatial template of $Z(\mathbf{u}_0), Z(\mathbf{u}_1), \dots, Z(\mathbf{u}_n)$

and moments can be readily established through the generalization of Eqs. (4) and (5) to multivariate probability distributions, the spatial cumulants can be written as a combination of spatial moments and vice versa (Smith 1995; Dimitrakopoulos et al. 2010). For a zero-mean stationary random field, the second-order spatial cumulant corresponds to the covariance and is given by

$$c_2^z(\mathbf{h}) = E[Z(\mathbf{u})Z(\mathbf{u} + \mathbf{h})], \quad (6)$$

the third- and fourth-order spatial cumulants are defined respectively as

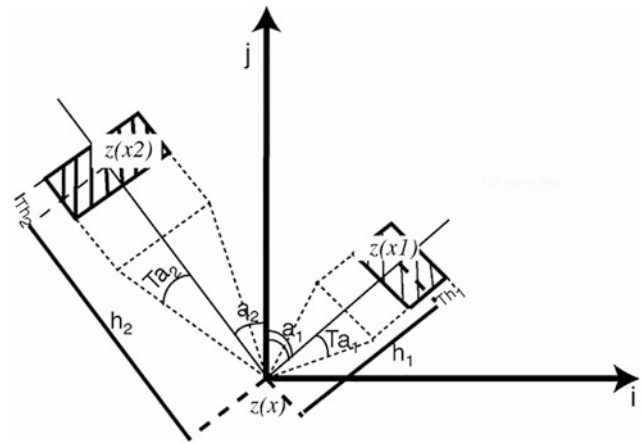
$$c_3^z(\mathbf{h}_1, \mathbf{h}_2) = E[Z(\mathbf{u})Z(\mathbf{u} + \mathbf{h}_1)Z(\mathbf{u} + \mathbf{h}_2)], \quad (7)$$

and

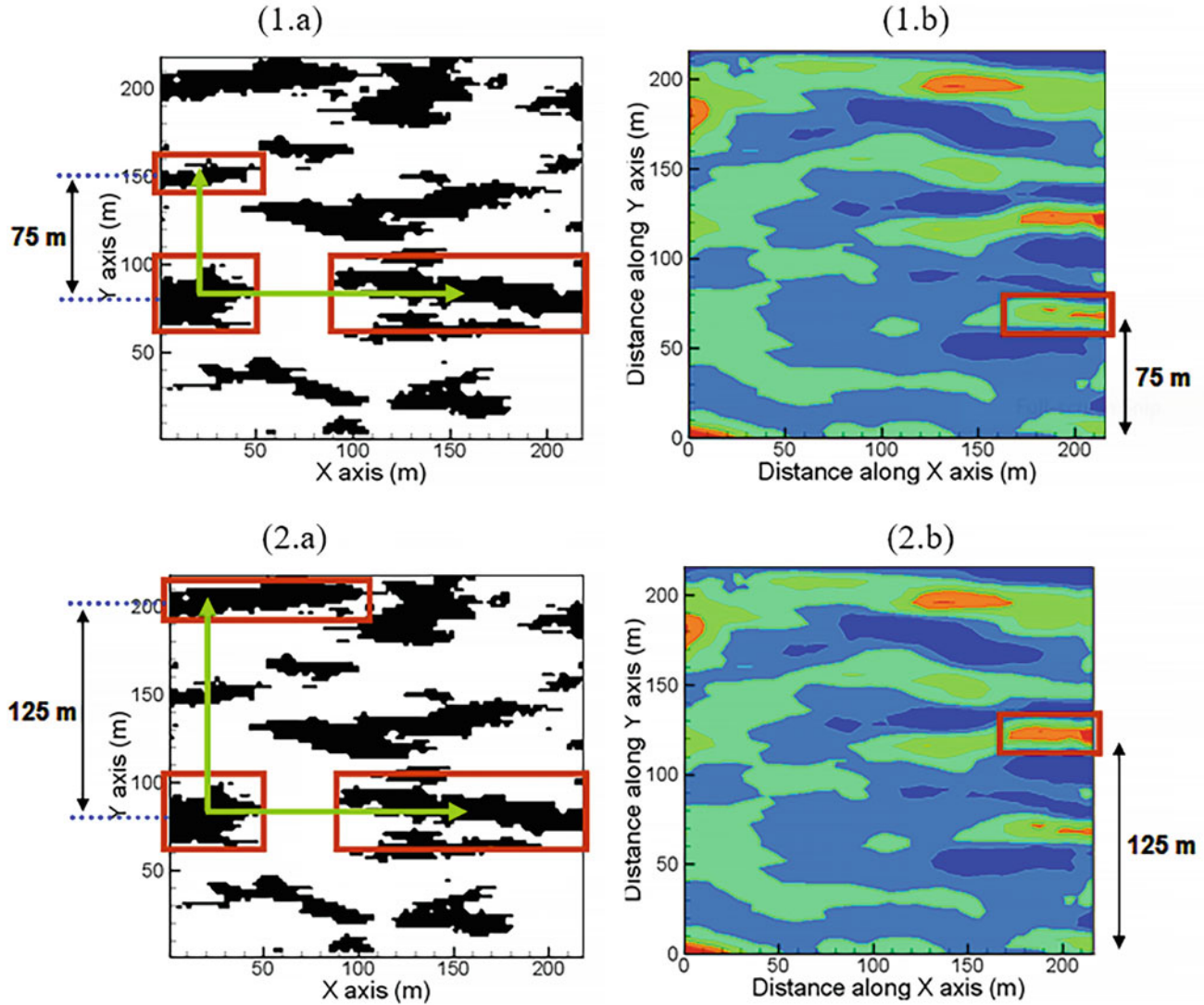
$$\begin{aligned}
 c_4^z(\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3) &= E[Z(\mathbf{u})Z(\mathbf{u} + \mathbf{h}_1)Z(\mathbf{u} + \mathbf{h}_2)Z(\mathbf{u} + \mathbf{h}_3)] \\
 &\quad - c_2^z(\mathbf{h}_1)c_2^z(\mathbf{h}_2 - \mathbf{h}_3) \\
 &\quad - c_2^z(\mathbf{h}_2)c_2^z(\mathbf{h}_3 - \mathbf{h}_1) \\
 &\quad - c_2^z(\mathbf{h}_3)c_2^z(\mathbf{h}_1 - \mathbf{h}_2).
 \end{aligned} \quad (8)$$

More specifically, spatial cumulants of order r ($r \geq 3$) consist of a combination of second-order spatial statistics and other spatial moments of order no greater than r .

The experimental computation of spatial cumulants is similar to that of experimental variograms, while the mathematical formulae are more sophisticated, as exemplified by Eqs. (7) and (8). The calculation of experimental spatial cumulants can be carried out either on an exhaustive training image or on irregularly distributed samples by allowing a certain tolerance in the corresponding spatial template (Fig. 3). A computational algorithm and related computer program for high-order spatial cumulants is available in the



High-Order Spatial Stochastic Models, Fig. 3 An example of irregular template for third-order cumulant calculation (from Dimitrakopoulos et al. 2010)



High-Order Spatial Stochastic Models, Fig. 4 Interpretations of the high positive anomalies in the third-order map. (1.a) to (2.a) explains, respectively, the high anomalies in (1.b) to (2.b) from the interactions

between the black zones in the red rectangles (from Mustapha and Dimitrakopoulos 2010a, b)

public domain (Mustapha and Dimitrakopoulos 2010a, b). This algorithm provides a tool to generate high-order cumulant maps as a typical application of spatial cumulants to delineate the spatial patterns of natural attributes. Figure 4 shows an example of third-order cumulant maps to characterize the interactions among attributes at multiple locations.

In general, high-order spatial cumulants provide a statistical entity to capture the spatial architecture of natural attributes, including spatial directionality, periodicity, homogeneity, and connectivity, in multiple directions and among multiple points. An additional property worth noting is that the spatial cumulants of the Gaussian random field vanish for orders higher than 2, indicating that high-order spatial cumulants are essential to characterize the non-Gaussian features of natural phenomena.

High-Order Simulation Methods

A general framework for generating realizations from a random field is the sequential simulation framework (Journel 1989, 1994; Deutsch and Journel 1992; Gómez-Hernández and Srivastava 2021), which has been applied to many popular stochastic simulation methods, as well as to the high-order stochastic simulation method (*hosim*) presented herein. Consider a random field $Z(\mathbf{u})$, $\mathbf{u} \in R^d$ ($d = 1, 2, 3$), consisting of a family of random variables as $Z(\mathbf{u}_1), Z(\mathbf{u}_2), \dots, Z(\mathbf{u}_N)$ on a finite discretized domain, with a multivariate probability density function denoted as $f(z_1, z_2, \dots, z_N)$. If an initial data set denoted as $\Lambda_0 = \{\zeta_1, \dots, \zeta_n\}$ is available, then the joint probability density function (PDF) can be decomposed (Johnson 1987) as

$$f(z_1, z_2, \dots, z_N | \Lambda_0) = f(z_1 | \Lambda_0) f(z_2 | \Lambda_1) \cdots f(z_N | \Lambda_{N-1}), \quad (9)$$

where $\Lambda_{i+1} = \{z_i\} \cup \Lambda_i$, $i = 0, 1, \dots, N-1$. The decomposition of joint PDF in Eq. (9) allows the generation of a realization from the related random field by sampling from a sequence of conditional probability distributions with density $f(z_i | \Lambda_{i-1})$, $i = 1, \dots, N$. In practice, the conditioning data Λ_i are often confined to a limited neighborhood considering the computational efficiency as well as the statistical irrelevancy of the farther data due to the so-called screen effect (Dimitrakopoulos and Luo 2004).

The main difference and advantage of high-order stochastic simulation from other previously developed simulation approaches, including MPS methods, is that they are not only free of distributional assumptions but also fully account for the high-order spatial statistics from the available data in a consistent, data-driven, and substantially more complete and informed manner. As the related technical literature shows (Minniakhmetov et al. 2018), the joint PDF of the random field is approximated by an orthogonal polynomial expansion series. Given a complete system of orthogonal functions, $\varphi_1(z)$, $\varphi_2(z)$, $\dots$, defined on interval $[a, b]$, then $f(z)$ can be approximated as

$$f(z) \approx \sum_{m=0}^{\omega} L_m \varphi_m(z), \quad (10)$$

where ω is the maximum order of the polynomials in the truncated expansion series. Similarly, an $(n+1)$ -variate piecewise continuous function $f(z_0, z_1, z_2, \dots, z_n)$ can be approximated as

$$f(z_0, z_1, \dots, z_n) \approx \sum_{m_0=0}^{\omega_0} \sum_{m_1=0}^{\omega_1} \cdots \sum_{m_n=0}^{\omega_n} L_{m_0, m_1, \dots, m_n} \varphi_{m_0}(z_0) \varphi_{m_1}(z_1) \cdots \varphi_{m_n}(z_n). \quad (11)$$

Especially, consider $f(z_0, z_1, \dots, z_n)$ as the joint PDF of random variables, $Z(\mathbf{u}_0)$, $Z(\mathbf{u}_1)$, $\dots$, $Z(\mathbf{u}_n)$, defined on a random field $Z(\mathbf{u})$; the coefficients $L_{m_0, m_1, \dots, m_n}$ can be derived as high-order moments according to the orthogonal property as (Minniakhmetov et al. 2018)

$$L_{m_0, m_1, \dots, m_n} \approx E[\varphi_{m_0}(z_0) \varphi_{m_1}(z_1) \cdots \varphi_{m_n}(z_n)] \\ \approx \frac{1}{N_{h_1, h_2, \dots, h_n}} \sum_{k=1}^{N_{h_1, h_2, \dots, h_n}} \varphi_{m_0}(\zeta_0^k) \varphi_{m_1}(\zeta_1^k) \cdots \varphi_{m_n}(\zeta_n^k). \quad (12)$$

The approximation in Eq. (12) also shows the experimental computation of the coefficients $L_{m_0, m_1, \dots, m_n}$, where $N_{h_1, h_2, \dots, h_n}$ is the number of replicates retrieved from the available data and $\zeta_0^k, \zeta_1^k, \dots, \zeta_n^k$ are the corresponding attribute values given a spatial template of distance vectors as $\mathbf{h}_1, \dots, \mathbf{h}_n$.

Note that the early approach in Mustapha and Dimitrakopoulos (2010b) uses Legendre polynomials as the orthogonal bases to develop the approximation of the joint PDF in Eq. (11). However, the numerical stability of approximation by Legendre polynomial series is influenced by the available data. Thus, Minniakhmetov et al. (2018) propose the use of Legendre-like splines to replace the Legendre polynomials as the orthogonal bases. This method is shown to perform better in the reproduction of spatial connectivity of extreme values, which are more likely to cause numerical instability in the expansion series.

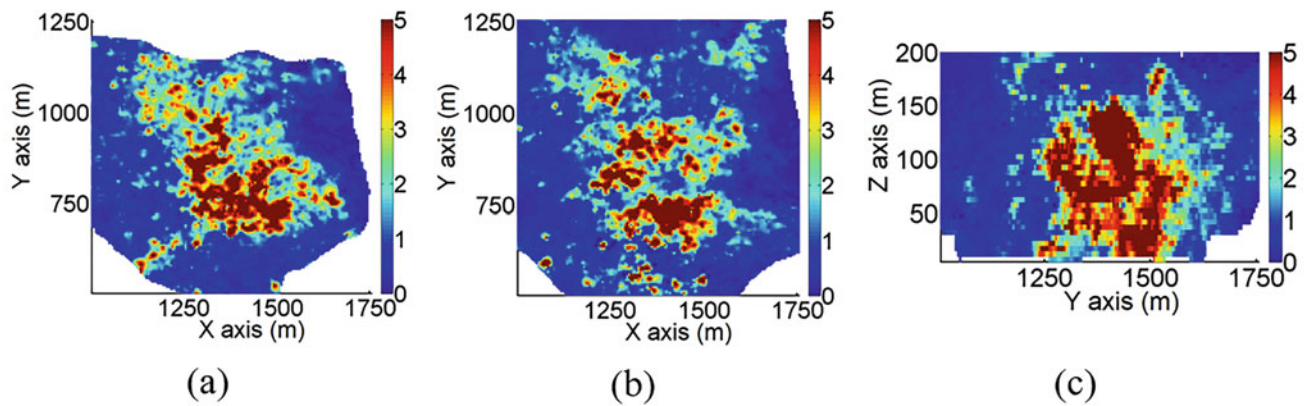
From the perspective of computational efficiency, a new computational model of high-order simulation has been proposed to replace the explicit calculation of spatial cumulants in the Legendre polynomial expansion series by a simple kernel-like function that can be computed in polynomial time (Yao et al. 2018). A statistical learning framework of high-order simulation is also developed based on a new proposed concept of spatial Legendre moment kernel so that the joint PDF of the random field is estimated in a stable solution domain through a learning algorithm, with possible extension to develop a TI-free high-order simulation (Yao et al. 2020; Yao et al. 2021a, 2021b). Other advancements in high-order simulation methods include block-support high-order simulation (de Carvalho et al. 2019), joint high-order simulation of multivariate attributes (Minniakhmetov and Dimitrakopoulos 2016), and high-order simulation of categorical data (Minniakhmetov and Dimitrakopoulos 2017, 2021).

An algorithmic description of high-order stochastic simulation can be summarized as the following:

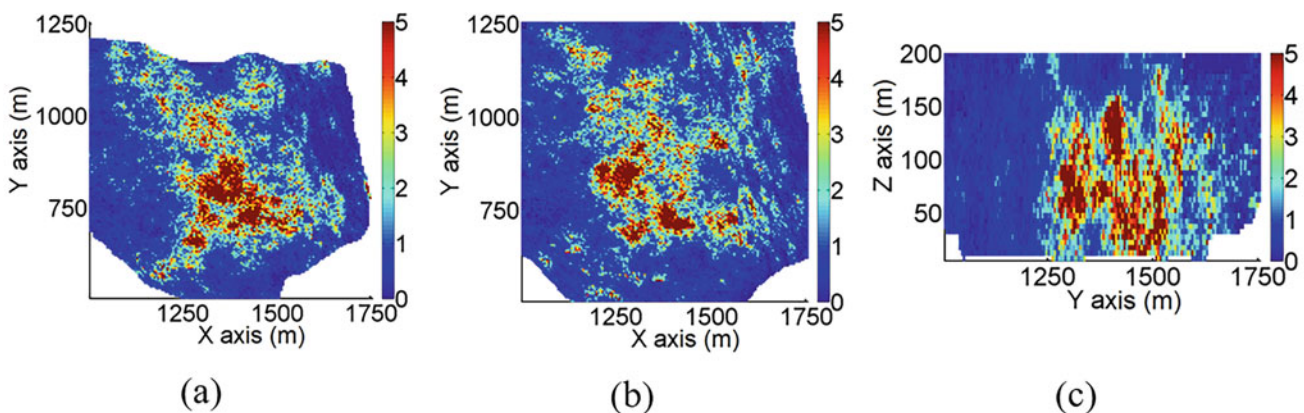
1. Draw a random path to visit all the nodes to be simulated.
2. Approximate the local conditional probability density function of each node based on Legendre polynomial expansion series in Eqs. (11) and (12) by computing the coefficients from the experimental high-order spatial statistics, which are calculated directly from the data and TI used.
3. Draw a random value from the local conditional probability density function for the simulated node and add it into the conditioning data.
4. Repeat from step (1) until all the nodes in the random path have been visited.

Examples

High-order sequential simulations (*hosim*) have been used in different applications, including for generating realizations of mineral deposit grades (e.g., Minniakhmetov and Dimitrakopoulos 2016; Minniakhmetov et al. 2018; de Carvalho et al. 2019) for mine production planning and the heterogeneity representation for subsurface flow (e.g.,



High-Order Spatial Stochastic Models, Fig. 5 Sections of a simulated realization using *hosim* from the gold deposit: (a) horizontal sect. 1; (b) horizontal sect. 2; and (c) a vertical section. Colors indicate grades in g/t. (From Minniakhmetov et al. 2018)



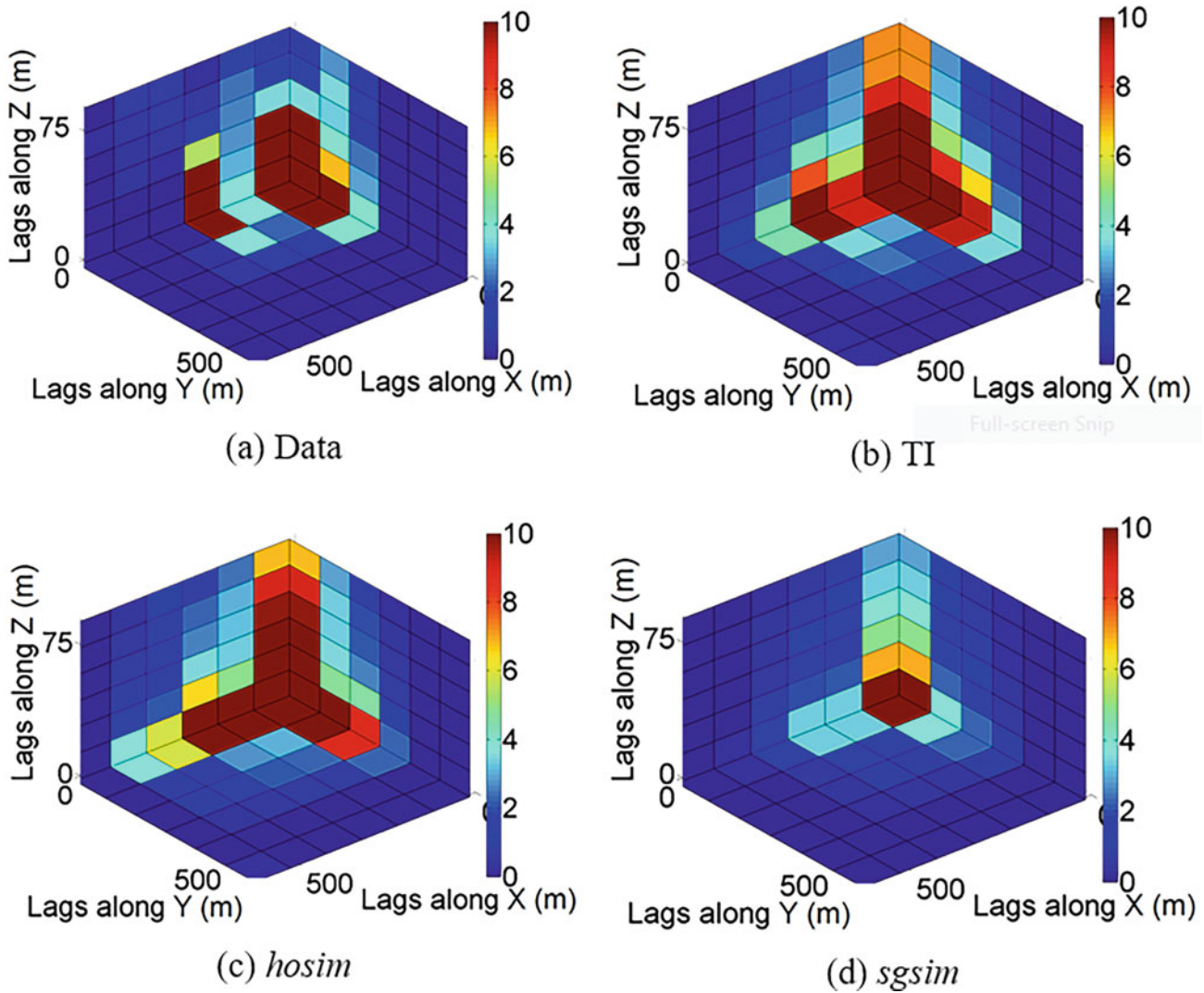
High-Order Spatial Stochastic Models, Fig. 6 The same sections as in Fig. 5 but from simulated realization of the gold deposit using *sgssim*: (a) horizontal sect. 1; (b) horizontal sect. 2; and (c) a vertical section. Colors indicate grades in g/t. (From Minniakhmetov et al. 2018)

Mustapha et al. 2011; Tamayo-Mas et al. 2016). To demonstrate the main aspects of high-order spatial statistics and simulation methods discussed in the previous sections, as well as the ability of high-order sequential simulation to capture data-driven spatial complexity, some examples from an application at a typical gold deposit (Minniakhmetov et al. 2018) are included in this section.

The data available in an operating gold mine cover an area around 4 km² and include 288 exploration drillholes; a training image is available from the blasthole samples. High-order sequential simulation is used to generate realizations of the gold grades. The high-order statistics needed are directly calculated from the available data and TI employed by the *hosim* algorithm; this is analogous to the approach used by MPS simulation algorithms; no fitting of parametric high-order statistical models is part of the related process. Figure 5 shows cross-sections from a realization generated using high-order sequential simulation, while Fig. 6 shows, for reasons of comparison, the same sections generated using the conventional sequential Gaussian simulation (*sgssim*). The differences in overall patterns are clear and as expected.

With respect to the reproduction of data histograms and variograms, realizations from both *hosim* and *sgssim* show a reasonable reproduction. As expected, high-order simulations also reproduce high-order spatial data statistics. For example, Fig. 7 shows the fourth-order cumulant maps of the data, the training image used, a *hosim* realization, and a realization from *sgssim*. Last, as seen in the visualization of the cross-sections in Figs. 5 and 6, *hosim* generates a better reproduction of the spatial connectivity of extreme values than *sgssim*, which is also shown in the connectivity measures for the 99th gold grade percentile in Fig. 8.

Note that comparisons of simulated realizations generated with high-order sequential simulation approaches to realizations generated with the MPS simulation method *filtersim* (Remy et al. 2009) also show a better performance of the high-order statistics framework with respect to integrating and reproducing complex spatial patterns (e.g., Mustapha et al. 2011; Yao et al. 2018).



High-Order Spatial Stochastic Models, Fig. 7 The fourth-order spatial cumulant maps of (a) the drill-hole data, (b) the training image (TI), (c) a simulation using *hosim*, and (d) a simulated realization using *sgsim*. (From Minniakhmetov et al. 2018)

Summary and Conclusions

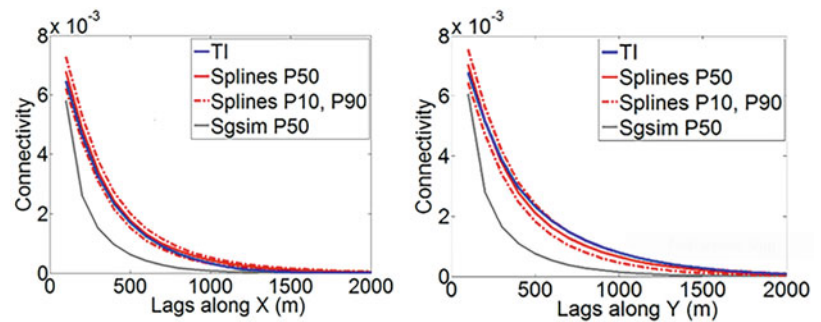
High-order stochastic simulation provides a general, data-driven sequential simulation framework free of distributional assumptions that characterize spatially distributed and varying natural phenomena with a diversity of complex spatial textures and architectures. The fundamental concepts in high-order sequential simulation are (i) the definition of high-order spatial statistics, such as spatial cumulants and (ii) the estimation of multivariate probability distributions of a random field based on mathematical models that incorporate the high-order spatial statistics.

The introduction and use of high-order spatial statistics and, more specifically, spatial cumulants is critically important and a major advancement compared to the previously developed framework of multiple point statistics and related simulation approaches. As noted in the introduction, spatial

cumulants characterize non-Gaussian spatial random fields and exhibit complex nonlinear directional spatial patterns and multiple point connectivity of extreme values, while making the subsequent simulation process consistent over a series of orders and facilitating a data-driven process. This consistency is not the case with MPS simulation methods, which are driven by a statistically less informed and arbitrarily selected pattern or moment from a training image. The data-driven nature of spatial cumulant inference facilitates (i) the development of high-order sequential stochastic simulation models, such as the ones mentioned above; (ii) the optimization of data processing and analytics, including the mitigation of the statistical conflicts among samples and other additional information such as the TI; and (iii) the utilization of advanced learning algorithms constrained to the high-order spatial statistics of the available data. It should be stressed that the high-order stochastic simulation framework is

High-Order Spatial Stochastic Models, Fig. 8

The connectivity along x (left subfigure) and y (right subfigure) for 99% percentile: the TI (blue line), P50 statistics for *hosim* simulations (solid red line), P10 and P90 statistics for *hosim* simulations (dashed red line), and P50 statistics for simulations using *sgsim* (grey line). (From Minniakhmetov et al. 2018)



nonparametric; that is, spatial cumulants or other type of high-order spatial statistics needed are calculated directly during the high-order sequential simulation process, and no parametric model fitting is introduced; thus making the use of related algorithms very practical and user friendly.

Note that computer programs for both calculating high-order cumulants as well as high-order sequential simulation with the method described herein are available in the public domain (Mustapha and Dimitrakopoulos 2010a, b; Minniakhmetov et al. 2018). The practice of high-order sequential simulation is, evidently, neither more complex nor simpler than other modern geostatistical simulation frameworks.

Cross-References

- Geostatistics
- Journel, André
- Multiple Point Statistics
- Probability Density Function
- Random Function
- Random Variable
- Sequential Gaussian Simulation
- Simulation
- Spatial Autocorrelation
- Spatial Statistics

Bibliography

- Boulemdjadjel A, Kaabache B, Kharfouchi S, Hachouf F (2018) Higher-order spatial statistics: a strong alternative. In image processing, 2018 3rd international conference on pattern analysis and intelligent systems (PAIS), pp. 1–6. <https://doi.org/10.1109/PAIS.2018.8598513>
- de Carvalho JP, Dimitrakopoulos R, Minniakhmetov I (2019) High-order block support spatial simulation method and its application at a gold deposit. *Math Geosci* 51(6):793–810. <https://doi.org/10.1007/s11004-019-09784-x>
- Deutsch CV, Journel AG (1992) *GSLIB: Geostatistical software library and user's guide*. Oxford University Press, New York
- Dimitrakopoulos R, Luo X (2004) Generalized sequential gaussian simulation on group size v and screen-effect approximations for large field simulations. *Math Geol* 36(5):567–591. <https://doi.org/10.1023/B:MATG.0000037737.11615.df>
- Dimitrakopoulos R, Mustapha H, Gloaguen E (2010) High-order statistics of spatial random fields: exploring spatial cumulants for modeling complex non-Gaussian and non-linear phenomena. *Math Geosci* 42(1):65–99. <https://doi.org/10.1007/s11004-009-9258-9>
- Gaztanaga E, Fosalba P, Elizalde E (2000) Gravitational evolution of the large-scale probability density distribution: the Edgeworth and gamma expansions. *Astrophys J* 539(2):522–531
- Gómez-Hernández JJ, Srivastava RM (2021) One step at a time: the origins of sequential simulation and beyond. *Math Geosci* 53(2): 193–209. <https://doi.org/10.1007/s11004-021-09926>
- Goodfellow R, Albor Consuegra F, Dimitrakopoulos R, Lloyd T (2012) Quantifying multi-element and volumetric uncertainty, Coleman McCreey deposit, Ontario. *Canada Computers & Geosciences* 42: 71–78. <https://doi.org/10.1016/j.cageo.2012.02.018>
- Guardiano F, Srivastava RM (1993) Multivariate geostatistics: beyond bivariate moments. In: Soares A (ed) *Geostatistics Tróia '92, Quantitative geology and geostatistics*, vol 5. Kluwer Academic, Dordrecht, pp 133–144. https://doi.org/10.1007/978-94-011-1739-5_12
- Johnson ME (1987) *Multivariate generation techniques*. In: *Multivariate statistical simulation*. Wiley, pp. 43–48. <https://doi.org/10.1002/9781118150740.ch3>
- Journel A (1989) *Fundamentals of Geostatistics in five lessons*. American Geophysical Union, Book Series 8. <https://doi.org/10.1029/SC008>
- Journel AG (1993) Geostatistics: roadblocks and challenges. In: Soares A. (eds) *Geostatistics Tróia '92. Quantitative geology and Geostatistics*, vol 5. Springer, Dordrecht. https://doi.org/10.1007/978-94-011-1739-5_18
- Journel A (1994) Modeling uncertainty: some conceptual thoughts. In: Dimitrakopoulos R (ed) *Geostatistics for the next century, Quantitative geology and Geostatistics*, vol 6. Springer, Dordrecht, pp 30–43. https://doi.org/10.1007/978-94-011-0824-9_5
- Journel AG (1997) Deterministic geostatistics: a new visit. In: Baafi E, Schofield N (eds) *Geostatistics Woolongong '96*. Kluwer, Dordrecht, pp 213–224
- Journel AG (2003) Multiple-point geostatistics: a state of the art. Report 16, Stanford Center for Reservoir Forecasting, Stanford University, Stanford Ca, USA. (available at pangea.stanford.edu)
- Journel AG (2005) Beyond covariance: The advent of multiple-point geostatistics. In *Geostatistics Banff 2004*, Springer, pp. 225–233. https://doi.org/10.1007/978-1-4020-3610-1_23
- Journel AG, Alabert F (1989) Non-Gaussian data expansion in the earth sciences. *Terra Nova* 1(2):123–134. <https://doi.org/10.1111/j.1365-3121.1989.tb00344.x>
- Journel AG, Zhang T (2006) The necessity of a multiple-point prior model. *Math Geo* 38(5):591–610
- Mariethoz G, Caers J (2014) *Multiple-point geostatistics: stochastic modeling with training images*. Wiley, Hoboken
- Mariethoz G, Renard P, Straubhaar J (2010) The direct sampling method to perform multiple-point simulation. *Water Resour Res* 46(W11536):10.1029/2008WR007621
- Minniakhmetov I, Dimitrakopoulos R (2016) Joint high-order simulation of spatially correlated variables using high-order spatial

- statistics. *Math Geosci* 49(1):39–66. <https://doi.org/10.1007/s11004-016-9662-x>
- Minniakhmetov I, Dimitrakopoulos R (2017) A high-order, data-driven framework for joint simulation of categorical variables. In: Gómez-Hernández JJ, Rodrigo-Ilarri J, Rodrigo-Clavero ME, Cassiraga E, Vargas-Guzmán JA (eds) *Geostatistics valencia 2016*. Springer International Publishing, Cham, pp 287–301. https://doi.org/10.1007/978-3-319-46819-8_19
- Minniakhmetov I, Dimitrakopoulos R (2021) High-order data-driven spatial simulation of categorical variables. *Math Geosci* 54(1): 23–45. <https://doi.org/10.1007/s11004-021-09943-z>
- Minniakhmetov I, Dimitrakopoulos R, Godoy M (2018) High-order spatial simulation using Legendre-like orthogonal splines. *Math Geosci* 50(7):753–780. <https://doi.org/10.1007/s11004-018-9741-2>
- Mustapha H, Dimitrakopoulos R (2010a) A new approach for geological pattern recognition using high-order spatial cumulants. *Comput Geosci* 36(3):313–334. <https://doi.org/10.1016/j.cageo.2009.04.015>. (code at: www.iamg.org/documents/oldftp/VOL36/v36-03-06.zip)
- Mustapha H, Dimitrakopoulos R (2010b) High-order stochastic simulation of complex spatially distributed natural phenomena. *Math Geosci* 42(5):457–485. <https://doi.org/10.1007/s11004-010-9291-8>
- Mustapha H, Dimitrakopoulos R, Chatterjee S (2011) Geologic heterogeneity representation using high-order spatial cumulants for subsurface flow and transport simulations. *Water Resour Res* 47. <https://doi.org/10.1029/2010WR009515>
- Nikias CL, Petropulu AP (1993) *Higher-order spectra analysis: a nonlinear signal processing framework*. PTR Prentice Hall, Englewood Cliffs, J
- Osterholt V, Dimitrakopoulos R (2007) Simulation of wireframes and geometric features with multiple-point techniques: application at Yandi iron ore deposit, Australia. In: *Orebody modelling and strategic mine planning*, vol 14. 2 edn. AusIMM Spectrum Series, pp 51–60
- Remy N, Boucher A, Wu J (2009) *Applied geostatistics with SGeMS: a user's guide*. Cambridge University Press, Cambridge, UK
- Rosenblatt M (1985) *Stationary sequences and random fields*. Birkhäuser, Boston
- Smith PJ (1995) A recursive formulation of the old problem of obtaining moments from cumulants and vice versa. *Am Stat* 49(2):217–218. <https://doi.org/10.1080/00031305.1995.10476146>
- Strebelle S (2002) Conditional simulation of complex geological structures using multiple-point statistics. *Math Geol* 34(1):1–21. <https://doi.org/10.1023/A:1014009426274>
- Strébelle S (2000) *Sequential simulation drawing structures from training images*. PhD thesis, Stanford University, Stanford
- Tamayo-Mas E, Mustapha H, Dimitrakopoulos R (2016) Testing geological heterogeneity representations for enhanced oil recovery techniques. *J Pet Sci Eng* 146:222–240. <https://doi.org/10.1016/j.petrol.2016.04.027>
- Yao L, Dimitrakopoulos R, Gamache M (2018) A new computational model of high-order stochastic simulation based on spatial Legendre moments. *Math Geosci* 50(8):929–960. <https://doi.org/10.1007/s11004-018-9744-z>
- Yao L, Dimitrakopoulos R, Gamache M (2020) High-order sequential simulation via statistical learning in reproducing kernel Hilbert space. *Math Geosci* 52(5):693–723. <https://doi.org/10.1007/s11004-019-09843-3>
- Yao L, Dimitrakopoulos R, Gamache M (2021a) Learning high-order spatial statistics at multiple scales: a kernel-based stochastic simulation algorithm and its implementation. *Comput Geosci*:104702. <https://doi.org/10.1016/j.cageo.2021.104702>
- Yao L, Dimitrakopoulos R, Gamache M (2021b) Training image free high-order stochastic simulation based on aggregated kernel statistics. *Math Geosci*. <https://doi.org/10.1007/s11004-021-09923-3>
- Zetsche C, Krieger G (2001) Intrinsic dimensionality: nonlinear image operators and higher-order statistics. In: Mitra SK, Sicuranza GL (eds) *Nonlinear image processing*. Academic Press, San Diego, pp 403–448. <https://doi.org/10.1016/B978-012500451-0/50014-X>
- Zhang F (2005) A high order cumulants based multivariate nonlinear blind source separation method. *Mach Learn* 61(1):105–127. <https://doi.org/10.1007/s10994-005-1506-8>

Hilbert Space

Daisy Arroyo

Department of Statistics, Universidad de Concepción, Concepcion, Chile

Definition

A Hilbert space is an inner product space that is complete with respect to the norm induced by the inner product and is commonly denoted as H . Completeness in this context means that each Cauchy sequence of elements in space converges to one element in space, in the sense that the norm of differences approaches to zero. Each Hilbert space is therefore also a Banach space (but not the reverse).

Introduction

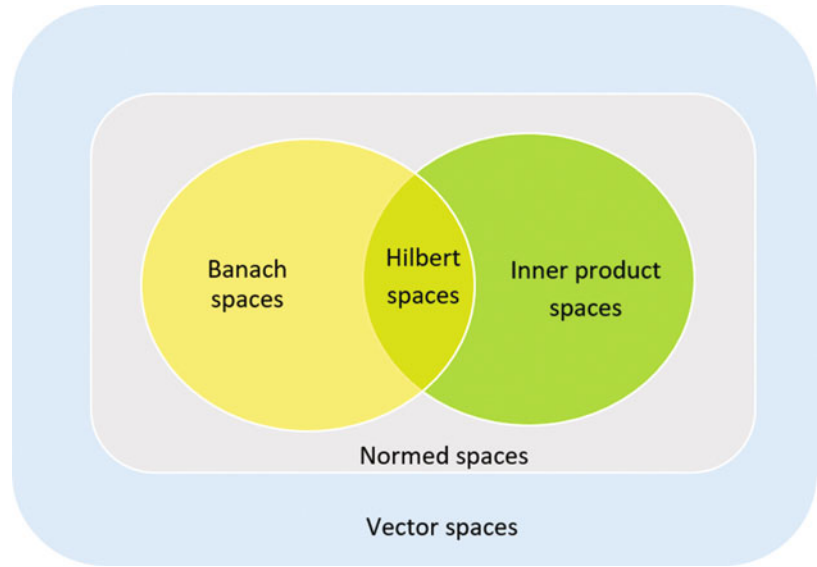
The methods of linear algebra study the concepts of spaces and apply to their transformations, operators, projectors, etc. In abstract spaces, elements are called a set of points, stochastic variables, vectors, and functions that have certain properties. In particular, the Euclidean spaces of n dimensions and the Hilbert spaces of infinite dimensions are of great importance. The Hilbert space concept is a generalization of the Euclidean space. This generalization allows algebraic and geometric notions and techniques applicable to two- and three-dimensional spaces to be extended to arbitrary-dimensional spaces, including infinite-dimensional spaces. There are no restrictions on how vectors should look in Hilbert space: they can be three-dimensional vectors, functions, or infinite sequences.

The Venn diagram below summarizes the different types of spaces and their relationship to the Hilbert space (Fig. 1).

Remembering that:

- A vector space is a set X on which two operations are defined called vector addition that takes each ordered pair (x, y) of elements of X to an element $x + y \in X$, and scalar multiplication that takes a pair (λ, x) to an element $\lambda x \in X$, for $\lambda \in \mathbb{R}$ (or $\mathbb{C}$) and $x \in X$. The vector addition must satisfy the following conditions:

Hilbert Space, Fig. 1 Venn diagram with the different types of spaces



- (a) $x + y = y + x$, for all $x, y \in X$ (commutative law).
- (b) $x + (y + z) = (x + y) + z$, for all $x, y, z \in X$ (associative law).
- (c) The set X contains an additive identity element, denoted by 0 , such that for any vector $x \in X$, $x + 0 = x$.
- (d) For each $x \in X$, there is an additive inverse $-x \in X$, such that $x + (-x) = 0$.

On the other hand, the scalar multiplication must satisfy the following conditions:

- (a) $\lambda(x + y) = \lambda x + \lambda y$, for all $x, y \in X$, and $\lambda \in \mathbb{R}$ (or $\mathbb{C}$) (distributive law)
 - (b) $(\lambda + \mu)x = \lambda x + \mu x$, for $x \in X$, and $\lambda, \mu \in \mathbb{R}$ (or $\mathbb{C}$)
 - (c) $\lambda(\mu x) = (\lambda\mu)x$, for $x \in X$, and $\lambda, \mu \in \mathbb{R}$ (or $\mathbb{C}$)
 - (d) For all $x \in X$, $1x = x$ (multiplicative identity)
- An inner product space is a vector space X along with an inner product on X , where an inner product is a function that takes each ordered pair (x, y) of elements of X to a number $\langle x, y \rangle \in \mathbb{R}$ (or $\mathbb{C}$) and has the following properties:
 - (a) $\langle x, x \rangle \geq 0$, for all $x \in X$, and $\langle x, x \rangle = 0$ if only if $x = 0$ (positive definiteness)
 - (b) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$, for all $x, y, z \in X$ (additivity)
 - (c) $\langle \lambda x, y \rangle = \lambda \langle x, y \rangle$, for all $\lambda \in \mathbb{R}$ (or $\mathbb{C}$), and all $x, y \in X$ (homogeneity)
 - (d) $\langle x, y \rangle = \overline{\langle y, x \rangle}$, for all $x, y \in X$ (conjugate symmetry), where the bar denotes the complex conjugate of a complex number

- A normed space is a vector space X endowed with a function

$$\|\cdot\| : X \rightarrow [0, \infty),$$

called the norm on X , with the following properties:

- (a) $\|\lambda x\| = |\lambda| \|x\|$, for all $x \in X$, $\lambda \in \mathbb{R}$ (homogeneity)
 - (b) $\|x + y\| \leq \|x\| + \|y\|$, for all $x, y \in X$ (triangle inequality)
 - (c) $\|x\| = 0$ if only if $x = 0$ (positive definiteness)
- A Banach space is a normed vector space X with the property that each Cauchy sequence $\{x_k\}_{k=1}^{\infty}$ in X converges toward some $x \in X$. In cases where the relevant vector space X is a subspace of a larger space, it is important to note that there are two additional requirements: each Cauchy sequence of elements in X must be convergent and the limit of the Cauchy sequence must belong to X .

Natural examples of Banach spaces are $\mathbb{R}^n$ and $\mathbb{C}^n$.

The Banach space $L^p(\mathbb{R})$, for $1 < p < \infty$, is the space of functions f for which $|f|^p$ is integrable:

$$L^p(\mathbb{R}) := \left\{ f : \mathbb{R} \rightarrow \mathbb{C} \text{ such that } \int_{-\infty}^{\infty} |f(x)|^p < \infty \right\}.$$

A Hilbert space is a complete inner product space. Every Hilbert space is a Banach space with respect to the norm.

Historical Material

In the first decade of the twentieth century, the German mathematician David Hilbert first described this space in his work on integral equations and Fourier series. Indeed, the definition of a Hilbert space was first given by von Neumann, rather than Hilbert, the origin of the designation “der abstrakte Hilbertsche Raum” in his famous work on unbounded Hermitian operators published in 1929. The name *Hilbert space* was soon adopted by others, for example, Hermann Weyl in his book *The Theory of Groups and Quantum Mechanics* published in 1931. After Hilbert’s research, the Austrian-German mathematician Ernst Fischer and the Hungarian mathematician Frigyes Riesz (1907; Fischer 1907) showed that integrable square functions could also be considered as points in a complete inner product space that is equivalent to the Hilbert space.

Examples

All finite-dimensional inner product spaces are Hilbert spaces, for example, Euclidean space with the ordinary dot product. Other examples of Hilbert spaces include spaces of square-integrable functions, spaces of sequences, Sobolev spaces consisting of generalized functions, and Hardy spaces of holomorphic functions. Infinite dimension examples are much more important in applications. Below are some of the most popular Hilbert spaces.

The Hilbert Space $L^2(\mathbb{R}^d)$

Among the $L^p(\mathbb{R}^d)$ -spaces, the case $p = 2$ is very important. The collection of square integrable functions in $\mathbb{R}^d$, $L^2(\mathbb{R}^d)$, is formed by all measurable functions of complex value f such that

$$L^2(\mathbb{R}^d) = \left\{ f : \mathbb{R}^d \rightarrow \mathbb{C} \text{ such that } \int_{\mathbb{R}^d} |f(x)|^2 dx < \infty \right\}.$$

The inner product is defined as

$$\langle f, g \rangle = \int_{\mathbb{R}^d} f(x) \overline{g(x)} dx, \forall f, g \in L^2(\mathbb{R}^d),$$

and the norm is defined as

$$\|f\| = \left(\int_{\mathbb{R}^d} |f(x)|^2 dx \right)^{1/2}.$$

Similarly, any E measurable subset of $\mathbb{R}^d$ with $m(E) > 0$ (measure of E must be positive), that is, the space of square integrable functions that are supported on E is a Hilbert space, denoted by $L^2(E)$.

The Hilbert Space $\mathbb{C}^n$

The standard inner product in this space is given by (Rudin 1986)

$$\langle x, y \rangle = \sum_{i=1}^n x_i \overline{y_i},$$

where $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$, with $x_i, y_i \in \mathbb{C}$, and the norm is defined as

$$\|x\| = \left(\sum_{i=1}^n |x_i|^2 \right)^{1/2}.$$

This space is complete, therefore is a finite-dimensional Hilbert space.

The Hilbert Space $l^2(\mathbb{Z})$

Defined as:

$$l^2(\mathbb{Z}) = \left\{ (\dots, x_{-2}, x_{-1}, x_0, x_1, \dots) : x_i \in \mathbb{C}, \sum_{n=-\infty}^{\infty} |x_n|^2 < \infty \right\}.$$

The space $l^2(\mathbb{Z})$ is a complex linear space. The inner product is given by

$$\begin{aligned} \langle x, y \rangle &= \sum_{n=-\infty}^{\infty} x_n \overline{y_n}, \text{ where} \\ x &= (\dots, x_{-2}, x_{-1}, x_0, x_1, \dots) \text{ and} \\ y &= (\dots, y_{-2}, y_{-1}, y_0, y_1, \dots), \end{aligned}$$

and the norm is defined as

$$\|x\| = \left(\sum_{n=-\infty}^{\infty} |x_n|^2 \right)^{1/2}.$$

All infinite dimensional (separable) Hilbert spaces are $l^2(\mathbb{Z})$ (Stein and Shakarchi 2005). The space $l^2(\mathbb{N})$ of square-summable sequences $\{x_n\}_{n=1}^{\infty}$ is defined in an analogous way.

The Hilbert Space $\mathbb{C}^{m \times n}$

The space of all $m \times n$ matrices with complex entries. The inner product on $\mathbb{C}^{m \times n}$ is defined as

$$\langle A, B \rangle = \text{tr}(A^* B) = \sum_{i=1}^m \sum_{j=1}^n a_{ij} \overline{b_{ij}},$$

where $A = (a_{ij})$, $B = (b_{ij})$ are matrices, tr denotes the trace, and $*$ denotes the Hermitian conjugate of a matrix (the complex-conjugate transpose). The corresponding norm

$$\|A\| = \left(\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2 \right)^{1/2}$$

is called the Hilbert-Schmidt norm.

The properties of orthogonality and projection in Hilbert spaces, which are of great interest in applications, are briefly described below.

Orthogonality

The inner product structure of a Hilbert space allows introducing the concept of orthogonality (Rudin 1986), which makes it possible to visualize vectors and linear subspaces of Hilbert space in a geometric way.

If x, y are vectors in a Hilbert space H , then x and y are orthogonal (denoted as $x \perp y$) if $\langle x, y \rangle = 0$. The subsets A and B are orthogonal (denoted as $A \perp B$) if $x \perp y$ for all $x \in A$ and for all $y \in B$. The orthogonal complement (denoted as $A^\perp$) of a subset A is the set of vectors orthogonal to A :

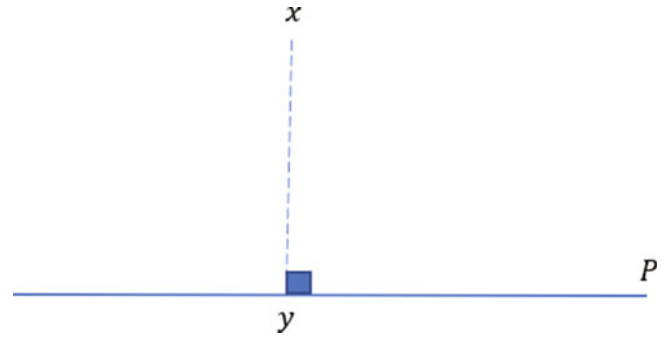
$$A^\perp = \{x \in H \mid x \perp y \text{ for all } y \in A\}.$$

The orthogonal complement of a subset of a Hilbert space is a closed linear subspace.

Projection

Let P be a closed linear subspace of a Hilbert space H , then (Fig. 2):

- (a) For each $x \in H$, there is a unique closest point $y \in P$ such that



Hilbert Space, Fig. 2 y is the point in P closest to x

$$\|x - y\| = \min_{z \in P} \|x - z\|.$$

- (b) The point $y \in P$ closest to $x \in H$ is the unique element of P with the property that $(x - y) \perp P$.

Applications

Hilbert spaces are the most important tools in the theories of partial differential equations, quantum mechanics, Fourier analysis, and ergodicity. Hilbert spaces serve to clarify and generalize the concept of Fourier expansion and certain linear transformations such as the Fourier transform. The Fourier series are a useful tool for the representation and approximation of periodic functions using trigonometric functions.

Each of following Hilbert space has a type of Fourier analysis associated with it:

- $L^2([a, b]) \rightarrow$ Fourier series
- $l^2([0, n - 1]) \rightarrow$ discrete Fourier transform
- $L^2(\mathbb{R}) \rightarrow$ Fourier transform
- $l^2(\mathbb{Z}) \rightarrow$ discrete time Fourier transform

The Hilbert space $L^2(\mathbb{R})$ plays a central role in many areas of science; for example, in signal processing, many time-varying signals are interpreted as functions in $L^2(\mathbb{R})$, and in quantum mechanics, the term “Hilbert space” simply means $L^2(\mathbb{R})$. For the case of $L^2([-\pi, \pi])$, the Fourier series is an infinite sum of functions $\sin(n\pi)$ and $\cos(n\pi)$, that is, oscillations with frequencies $\frac{n}{2\pi}$, with $n = 0, 1, 2, \dots$

In the 1920s von Neumann realized that Hilbert spaces are the natural setting for quantum mechanics, where a particle is confined to move in a straight line between two parallel walls. This particle at each instant in time t is described by an element $\varphi(\cdot, t) \in L^2([0, M])$, that is, a vector in the Hilbert

space of square-integrable, complex-valued function on the interval $[0, M]$, where M is the distance between the walls and the function φ is called the wave function of the particle.

In probability theory, Hilbert spaces arise naturally. A random experiment is modeled mathematically by a space Ω (sample space) and a probability measure P on Ω . Each point $\omega \in \Omega$ corresponds to a possible outcome of the experiment. An event A is a measurable subset of Ω . The probability measure P associates each event A with a probability $P(A)$, where $0 \leq P(A) \leq 1$ and $P(\Omega) = 1$. A random variable X is a measurable function $X : \Omega \rightarrow \mathbb{C}$, which maps each outcome in ω to a complex number (or real number). The expected value $\mathbb{E}[X]$ of a random variable X is the mean, or integral, of the random variable X with respect to the probability measure P , that is

$$\mathbb{E}[X] = \int_{\Omega} X(\omega) dP(\omega).$$

On the other hand, a random variable X is said to be second order if $\mathbb{E}[X^2] < \infty$. The set of second-order random variables forms a Hilbert space with respect to the inner product

$$\langle X, Y \rangle = \mathbb{E}[X\bar{Y}].$$

The space of second-order random variables may be identified with the space $L^2(\Omega, P)$ of square-integrable functions on (Ω, P) , with the inner product

$$\langle X, Y \rangle = \int_{\Omega} X(\omega) \overline{Y(\omega)} dP(\omega).$$

In geostatistics, one can find the Hilbert spaces of random fields (Chilès and Delfiner 2012). Consider a family of complex-valued random variables X defined on a probability space (Ω, A, P) and having finite second-order moments (Chilès and Delfiner 2012)

$$\mathbb{E}[|X|^2] = \int |X(\omega)|^2 P(d\omega) < \infty.$$

These random variables constitute a Hilbert space denoted $L^2(\Omega, A, P)$, with its inner product $\langle X, Y \rangle = \mathbb{E}[X\bar{Y}]$ (the upper bar denotes complex conjugation) and its norm $\|X\| = \sqrt{\mathbb{E}[|X|^2]}$. In this Hilbert space, two random variables are orthogonal when they are uncorrelated ($\langle X, Y \rangle = \mathbb{E}[X\bar{Y}] = 0$).

In a Hilbert space, the orthogonal projection of X onto a closed linear subspace K can be defined as the unique point X_0 in the subspace closest to X (Chilès and Delfiner 2012; Halmos 1951):

$$X_0 = \arg \min_{Y \in K} \|X - Y\| \Leftrightarrow \langle X - X_0, Y \rangle = 0, \text{ for all } Y \in K.$$

Since $X_0 \in K$, it satisfies $\langle X - X_0, X_0 \rangle = 0$, considering the triangular inequality we have

$$\|X - X_0 + X_0\|^2 = \|X - X_0\|^2 + \|X_0\|^2.$$

Or equivalently

$$\|X - X_0\|^2 = \|X\|^2 - \|X_0\|^2.$$

This result is used in kriging theory.

Conclusions

The Hilbert spaces are Banach spaces with a norm that is derived from an inner product, which is used to introduce the notion of orthogonality. The Hilbert spaces have several applications which arise naturally, mainly in physics and mathematics.

Cross-References

- [Fast Fourier Transform](#)
- [Geostatistics](#)
- [Kriging](#)

Bibliography

- Chilès JP, Delfiner P (2012) Geostatistics: modeling spatial uncertainty. Wiley Series in Probability and Statistics, 2nd edn, John Wiley & Sons, Inc., Hoboken, New Jersey, 734pp
- Fischer E (1907) Sur la convergence en moyenne. C R Acad Sci 144: 1022–1024
- Halmos PR (1951) Introduction to Hilbert space and the theory of spectral multiplicity. Chelsea, New York, 118pp
- Riesz F (1907) Sur les systèmes orthogonaux de fonctions. C R Acad Sci 144:615–619
- Rudin W (1986) Real and complex analysis, 3rd edn. McGraw-Hill Education, McGraw-Hill Book Co., Singapore, 483pp
- Stein EM, Shakarchi R (2005) Real analysis: measure theory, integration, and Hilbert spaces. Princeton University Press, Princeton, 424pp

Horton, Robert Elmer

Keith Beven

Lancaster University, Lancaster, UK



Fig. 1 Robert Elmer Horton (from Merrill Bernard 1945), from Bulletin American Meteorological Society

Biography

Robert Horton is recognized as one of the foremost hydrologists of the twentieth century. He was born in 1875 in Parma, Michigan and obtained a BSc degree from Albion College, Michigan in 1897 (which later also awarded him an honorary Ph.D. in 1932). Horton's professional work began under the direction of his uncle, George Rafter, a prominent civil engineer who had earlier worked on the Erie Canal. For better streamflow measurements, Rafter had commissioned laboratory weir studies at Cornell. Horton analyzed and summarized the results. This initial work was then considerably extended when Horton joined the US Geological Survey in 1900. The resulting publications became the standard US work on the subject.

This led on to a program of extensive stream gaging on New York streams and work on baseflow which he recognized was dominated by groundwater. He also started to focus on the role of infiltration in controlling both recharge to groundwater and surface runoff. He subsequently came to view the soil surface as a "separating surface" between water that either infiltrated and would later become baseflow or return to the atmosphere as evapotranspiration, and water in excess of the infiltration capacity of the soil surface that would become surface runoff and could contribute quickly to the stream hydrograph (Horton 1933). He proposed an equation for infiltration rates at the surface that would then allow

the calculation of rates of surface runoff given rainfall inputs (Horton 1939). This was used by Horton in his later work as a hydrologic consultant, based at the Horton Hydrological Laboratory near Voorheesville, New York. It was rapidly adopted by many others as a useful tool in flood prediction to the extent that the concept became known as "Hortonian overland flow."

Horton's view of infiltration was, however, much more sophisticated than has been commonly appreciated. He argued that infiltration capacities were controlled by surface rather than profile controls, and that these would change over time both within an event due to "extinction phenomena," such as the displacement of soil particles blocking larger pores, and between events. He also recognized the role of air pressure, cracks and macropores, on infiltration and the variability of soil characteristics in runoff production. Interestingly, analyses suggest that he might not have seen widespread surface runoff on his own experimental catchment at the Horton Laboratory very often (Beven 2004). This did not stop the concepts getting widely used (even to the present day).

Throughout his career, Horton was interested in flood generation and prediction, including publishing on the meteorological aspects of intense storms. He was one of the first to try to estimate "maximum possible rainfalls" for different places, which when combined with the infiltration theory might allow a maximum flood to be estimated. He showed how some observed events had approached these limits.

In terms of mathematical geoscience, Horton published a number of interesting analytical solutions for hydrological processes during his career, but his most significant contribution appeared in the Bulletin of the Geological Society of America just 1 month before his death in 1945 (Horton 1945). This is an extensive treatise of 95 pages that deals with the implications of his surface runoff theory on the erosional development of hillslopes and drainage basins. It is a remarkable seminal achievement that later stimulated other hydrologists and geomorphologists in the "quantitative revolution" in Geography that started in the late 1960s and the analysis of landscapes as fractal surfaces in the 1980s.

Horton's importance to both hydrology and meteorology is recognized by both the Horton Medal of the American Geophysical Union and the Horton Lecture of the American Meteorological Society.

Bibliography

- Bernard M (1945) Robert E. Horton. Bull Am Meteorol Soc 26:242
- Beven KJ (2004) Surface runoff at the Horton hydrologic laboratory (or not?). J Hydrol 293:219–234

- Horton RE (1933) The role of infiltration in the hydrologic cycle. *Trans Am Geophys Union* 14:446–460
- Horton RE (1939) Analysis of runoff-plat experiments with varying infiltration-capacity. *Trans Am Geophys Soc* 20:693–711
- Horton RE (1945) Erosional development of streams and their drainage basins: hydrophysical approach to quantitative morphology. *Bull Geol Soc Am* 56:275–370

Howarth, Richard J.

J. M. McArthur

Earth Sciences, UCL, London, WC1E 6BT, UK



Fig. 1 Richard J. Howarth, courtesy of Prof. Howarth

Biography

Richard J. Howarth was born in Southport, UK, in 1941. After gaining a B.Sc. in geology in 1963 from Bristol University he stayed on to undertake a Ph.D. (completed 1966) on the Irish extension of the Cryogenian Port Askaig Tillite formation. The research involved statistical calibrations for the first XRF spectrometer to be installed in a British geology department, and the use of factor analysis for the interpretation of geochemical and other data. The period 1966–1968 was spent in

The Hague, The Netherlands, with Bataafse Petroleum Maatschappij (Shell), where he undertook statistical analysis relating the geology of the World's sedimentary basins to their hydrocarbon production.

Returning to the UK in 1968 for family reasons, he joined John Webb's Applied Geochemistry Research Group, in the Department of Geology at Imperial College of Science & Technology, London, where he worked until 1985. His research at Imperial College was initially concerned with analytical methodologies and with mapping and interpreting regional geochemical survey data for mineral exploration, geological, and epidemiological purposes. He worked with chemist Michael Thompson to develop statistical techniques and procedures, which have now become standard, for quality assurance of analytical data used for regional geochemical mapping (e.g., Thompson and Howarth 1973). The software he developed while at Imperial underpinned production of pioneering regional geochemical atlases of Northern Ireland, of England and Wales (Webb et al. 1978), and elsewhere. He also worked with a team under George Koch Jr., of the University of Georgia, USA (1978–84), on the statistical interpretation of data from the US National Uranium Resource Evaluation Program. Much of this accumulated experience was summarized in Howarth (1983).

In 1985 he joined the British Petroleum group as a senior consultant in BP Research, Sunbury, where his research and advisory duties encompassed biostratigraphy, chemostratigraphy, estimation of hydrocarbon reserves, petrophysical data analysis, laboratory quality-control, well-inflow performance monitoring, and numerous other matters. In the course of this, he collaborated with P.C. Smalley and A. Higgins and others to develop Sr-isotope stratigraphy, work later continued with others at University College London over the next 23 years (e.g., McArthur et al. 2001).

Since 1993, he has held the honorary position of Professor in Mathematical Geology in the Department of Earth Sciences, University College London. During this time his interests broadened to include writings on the history of geology and geophysics (e.g., Howarth 2017). His wide reading of the literature of applied statistics enabled him to cross discipline boundaries and introduce techniques which were new to geology, for example, empirical discriminant function, non-linear mapping, power transform, and various estimators of uncertainty.

He has been awarded the Murchison Fund of The Geological Society of London (1987), William Christian Krumbein Medal of the International Association for Mathematical Geology (2000) for contributions to mathematical geology, and the Sue Tyler Friedman Medal of the Geological Society of London (2016) for contributions to the history of geology.

Bibliography

- Howarth RJ (ed) (1983) Handbook of exploration geochemistry. volume 2. Statistics and data analysis in geochemical prospecting. Elsevier, Amsterdam. 425 p
- Howarth RJ (2017) Dictionary of mathematical geosciences with historical notes. Springer International Publishing, Cham. 893 p
- McArthur JM, Howarth RJ, Bailey TR (2001) Strontium isotope stratigraphy: LOWESS version 3: best fit to the marine Sr-isotope curve for 0–590 Ma and accompanying look-up table for deriving numerical age. *J Geol* 109:155–170
- Thompson M, Howarth RJ (1973) The rapid estimation and control of precision by duplicate determinations. *Analyst* 98:153–160
- Webb JS, Thornton I, Thompson M, Howarth RJ, Lowenstein PL (1978) The Wolfson geochemical atlas of England and Wales. Clarendon Press, Oxford/London. 74 p

Hurst Exponent

Sid-Ali Ouadfeul¹ and Leila Aliouane²

¹Algerian Petroleum Institute, Sonatrach, Boumerdes, Algeria

²LABOPHT, Faculty of Hydrocarbons and Chemistry, University of Boumerdes, Boumerdes, Algeria

Definition of the Hölder Exponent

The singularity exponent at a point x_0 can be expressed by an exponent called the Hölder exponent, and this exponent is defined as follows:

The Hölder exponent of the distribution f at point (x_0) is the largest h , whose f is Lipschitzian with exponent h at point (x_0) . It means, there exists a constant C and a polynomial of $P_n(x)$ of n order such that for all x belonging to the neighborhood of x_0 we have:

$$|f(x) - P_n(x - x_0)| \leq C|x - x_0|^h \quad (1)$$

If $h(x_0) \in]n, n + 1[$ we can easily demonstrate that at point x_0 the polynomial $P_n(x)$ corresponds to the Taylor series of f at point x_0 , $h(x_0)$ measures how the distribution f is irregular at x_0 , it means:

$h(x_0)$ is very high $\Leftrightarrow f$ is regular.

In the most cases $h(x_0) = h + 1$ for the primitive of f , and $h(x_0) = h - 1$ for the derivative.

Definition of the Hurst Exponent

If the Hölder exponent of a function f is constant per interval ($h_i = H_j$ for the interval $[a_j, b_j]$, $i = 0, 1, 2, \dots, N_j$, $j = 1, 2, 3, \dots, M$), H_j are the Hurst exponents.

H_j measures the global regularity per interval.

Estimated Hurst exponent H of the function f is able to say that f is an anti-correlated random walk ($H < 1/2$: anti-persistent random walk), or positively correlated ($H > 1/2$: persistent random walk). $H = 1/2$ corresponds to classical Brownian motion (Audit et al. 2004).

The Multifractal Brownian Motion

Let $H \in]0, 1[$, the fractional Brownian motion (fBm) is defined as a centered Gaussian process $(B_t^H)_{t \geq 0}$ with covariance function $RH(t, s) = E(B_t^H B_s^H) = 1/2(t^{2H} + s^{2H} - |t - s|^{2H})$. The index H is called the Hurst exponent and it determines the main properties of the process B^H , such as a self-similarity and a regularity of the sample paths and a long memory. Figure 1 shows an example of fractional Brownian motion realization with 1024 samples and a Hurst exponent $H = 0.60$.

Methods of Estimation of the Hurst Exponent

Spectral Density Method

The Fourier transform of a fBm function $f(t)$ is given by:

$$FT[f(t)] = \int_{-\infty}^{+\infty} f(t)e^{-2\pi ift} dt = F(f) = A(f) + jB(f) \quad (2)$$

$$F(f) = |F(f)|e^{j\varnothing(f)} \quad (3)$$

$|F(f)|$ and $\varnothing(f)$ are the spectra of amplitude and phase, these spectra are given by:

$$|F(f)| = \sqrt{A(f)^2 + B(f)^2} \quad (4)$$

$$\varnothing(f) = \begin{cases} \arctg\left(\frac{B(f)}{A(f)}\right) & \text{if } A(f) \neq 0 \\ \frac{\pi}{2} & \text{if } A(f) = 0 \end{cases} \quad (5)$$

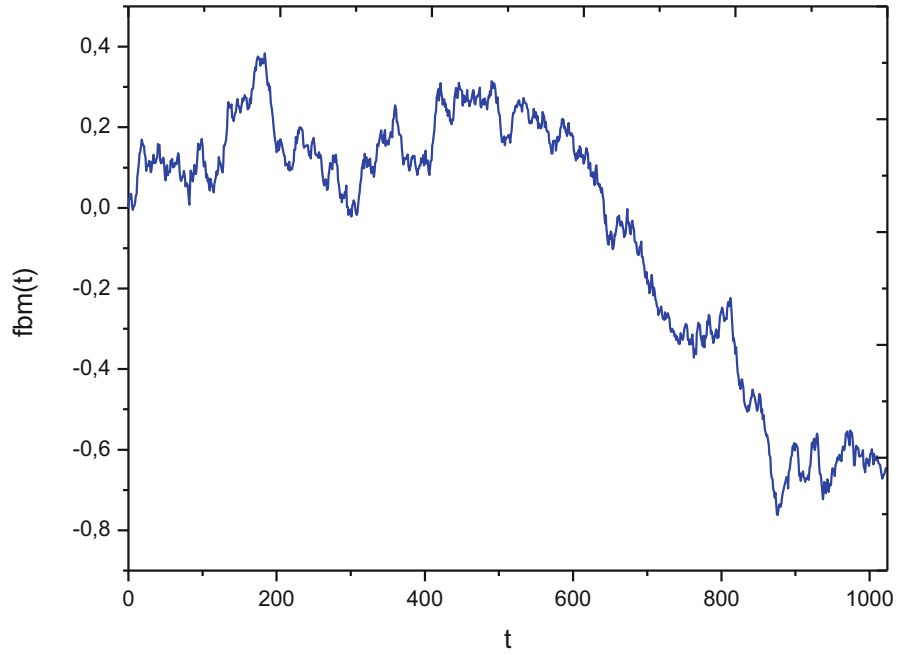
$|F(f)|^2$ is the spectral density

$|F(f)|^2 = \frac{1}{f^\beta}$ where β is the spectral exponent, it is related to the Hurst exponent by a linear relationship $\beta = 2H + 1$

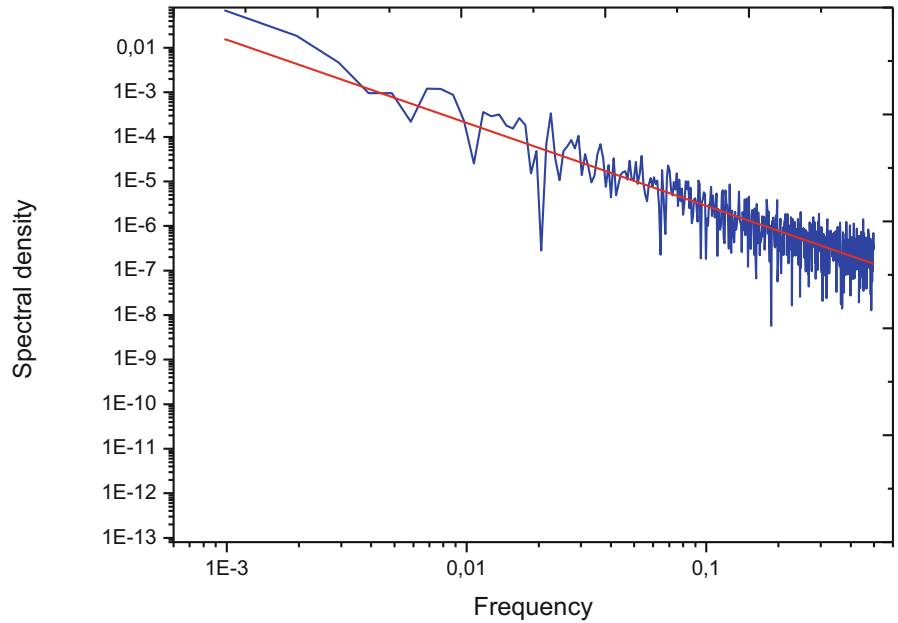
To estimate the spectral exponent, we present the spectral density versus the frequency in the log-log scale, a linear regression will provide an estimated value of β by consequence of H .

Figure 2 shows the graph of the spectral density versus the frequency in the log-log scale, the linear regression (red line)

Hurst Exponent, Fig. 1 An example of a Fractional Brownian motion realization with 1024 samples and $H = 0.60$



Hurst Exponent, Fig. 2 The spectral density of the fBm signal presented in Fig. 1 versus the frequency in the log-log scale, the red line is the linear regression



in Fig. 2 gives a slope $\beta = 2.2 \pm 0.05$ by consequence $H = 0.60 \pm 0.025$.

The Variogram

A variogram is a geostatistical method of comparing similarity of a data value to neighboring values within a field of data (Deutsch 2003). The variogram is calculated by:

$$\gamma(h) = \frac{1}{N} \sum_{i=1}^N [Z(x_i) - Z(x_{i+h})]^2 \quad (6)$$

where N is the number of neighboring data points within the specified lag distance being compared.

$Z(x_i)$: is the physical property parameter value of the initial point.

$Z(x_i + h)$: is the parameter value of the neighboring point.

The value of logarithm $\gamma(h)$ is then plotted versus logarithm the distance between the initial point and the compared points. The Hurst exponent is the slope of straight line resulting of the linear regression of this curve (Ouafeul and Aliouane 2011). The estimated Hurst exponent of the fBm signal presented in Fig. 1 is $H = 0.59 \pm 0.012$ which is very close to the actual Hurst exponent.

The Detrended Fluctuation Analysis

The method for quantifying the correlation propriety in non-stationary time series is based on the computation of a scaling exponent H by means of a modified root mean square analysis of a random walk (Peng et al. 1994).

To compute H from a time-series $x(i)$ [$i = 1, \dots, N$], like the interval tachogram, the time series is first integrated:

$$y(k) = \sum_{i=1}^k [x(i) - M] \quad (7)$$

where M is the average value of the series $x(i)$, and k ranges between 1 and N .

Next, the integrated series $y(k)$ is divided into boxes of equal length n and the least-square line fitting the data in each box, $y_n(k)$, is calculated. The integrated time series is detrended by subtracting the local trend $y_n(k)$, and the root-mean square fluctuation of the detrended series, $F(n)$, is computed as (Peng et al. 1995):

$$F(n) = \sqrt{\frac{1}{N} \sum_{i=1}^N [y(k) - y_n(k)]^2} \quad (8)$$

$F(n)$ is computed for all time-scales n .

Typically, $F(n)$ increases with n , the “box-size.” If $\log F(n)$ increases linearly with $\log n$, then the slope of the line relating $F(n)$ and n in a log-log scale gives the scaling exponent α .

where $\alpha = H + 1$

The estimated Hurst exponent of the fBm signal presented in Fig. 1 using the DFA method (see Fig. 3) is $H = 0.46 \pm 0.01$, which is different to theoretical Hurst exponent, since the DFA estimator is recommended for short time series (Ouafeul and Aliouane 2011).

The Wavelet Transform Modulus Maxima Lines

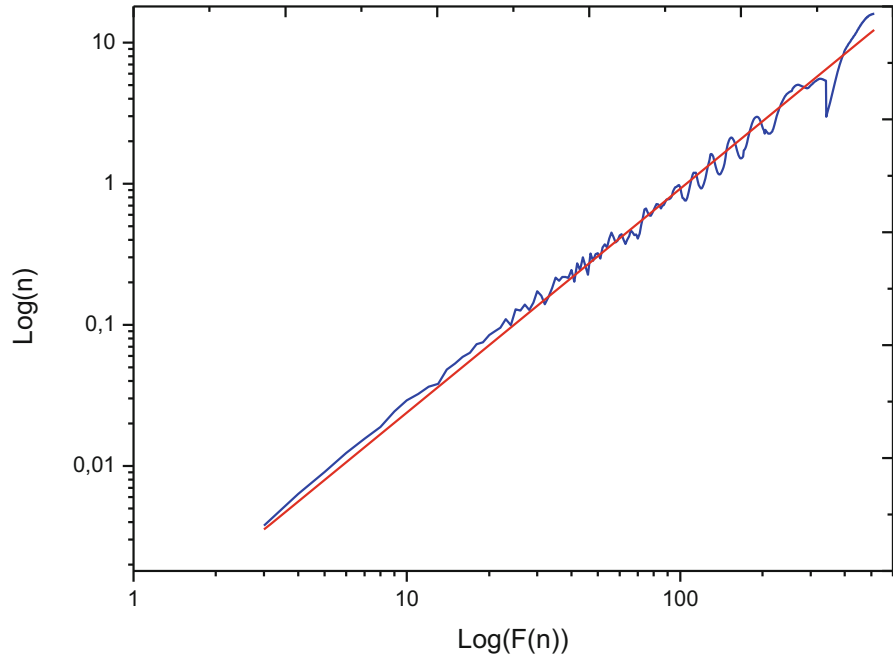
The wavelet transform modulus lines (WTMM) is a multifractal formalism introduced by Frish and Parisi (1985) revisited by the continuous wavelet transform introduced Grossmann and Morlet in 1985. The WTMM method was first introduced by Malat and Huang in 1992 and used for image processing.

The continuous wavelet transform is a decomposition of a given signal $S(t)$ into a dilated and translated wavelets $\varphi^*\left(\frac{t-b}{a}\right)$ obtained from a mother $\varphi(t)$ that must have n vanishing moments;

$$\int_{-\infty}^{+\infty} t^n \varphi(t) dt = 0, n < +\infty \quad (9)$$

$$CWT(a, b) = \int_{-\infty}^{+\infty} S(t) \varphi^*\left(\frac{t-b}{a}\right) dt \quad (10)$$

Hurst Exponent, Fig. 3 DFA estimator of the Hurst exponent of the fBm signal presented in Fig. 1, the red line is the linear fit



where $a \in \mathbb{R}^{*+}$, $b \in \mathbb{R}$

The first step of WTMM method is to calculate the continuous wavelet transform (CWT) of a given signal and the modulus of CWT, the next step is maxima of the continuous wavelet transform. Determination of local maxima is performed using the computation of the first and second derivative of the wavelet coefficients. CWT (a, b) admits a maximum at point b_0 if it satisfies the following two conditions (Ouafeul 2020):

$$\begin{aligned} \frac{\partial CWT}{\partial b} &= 0 & \text{if } b &= b_0 \\ \frac{\partial^2 CWT}{\partial b^2} &< 0 & \text{if } b &= b_0 \end{aligned} \quad (11)$$

The function of partition $Z(q, a)$ is a summation of the modulus of the CWT at local maxima b_i with a q power $\in \mathbb{R}$:

$$Z(q, a) = \sum_{b_i} |CWT(a, b)|^q \quad (12)$$

The spectrum of exponents $\tau(q)$ is related the function of partition $Z(q, a)$ by:

$$Z(q, a) = a^{\tau(q)} \quad (13)$$

So the spectrum of exponents $\tau(q)$ is obtained by a simple linear regression of $\log(Z(q, a))$ versus $\log(q)$

The spectrum of singularities is the Legendre transform of the spectrum of singularizes (Ouafeul 2020):

$$D(h) = \min_q (qh - \tau(q)) \quad (14)$$

For fractional Brownian motions signals which have monofractal character, the spectrum of exponents has a linear behavior with equation:

$$\tau(q) = qH - 1 \quad (15)$$

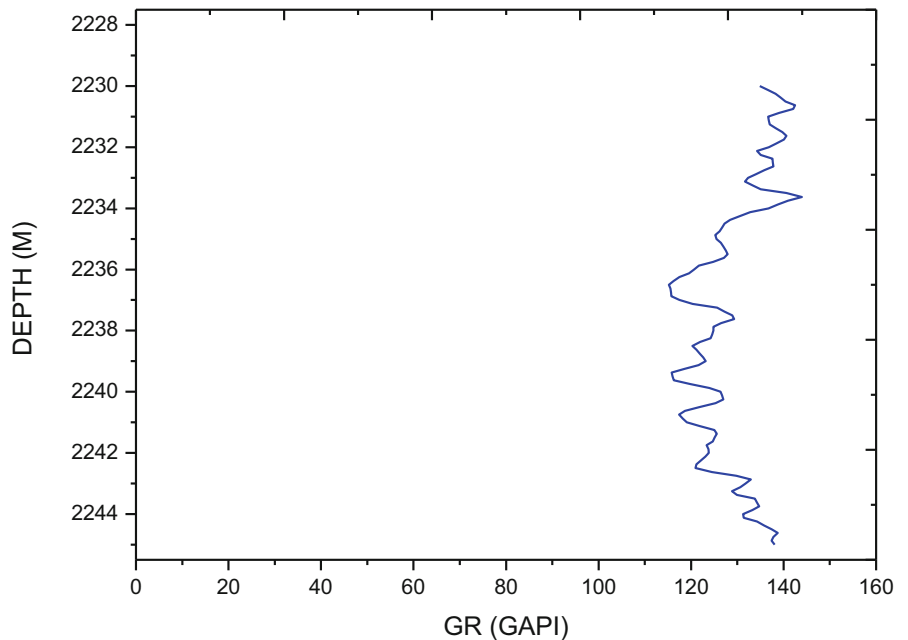
So the Hurst exponents are obtained by a simple linear regression of $\log(\tau(q))$ versus $\log(q)$.

The WTMM method applied to the fBm signal presented in Fig. 1 gives the following Hurst exponent $H = 0.56 \pm 0.002$.

Application to Well-Logs Data

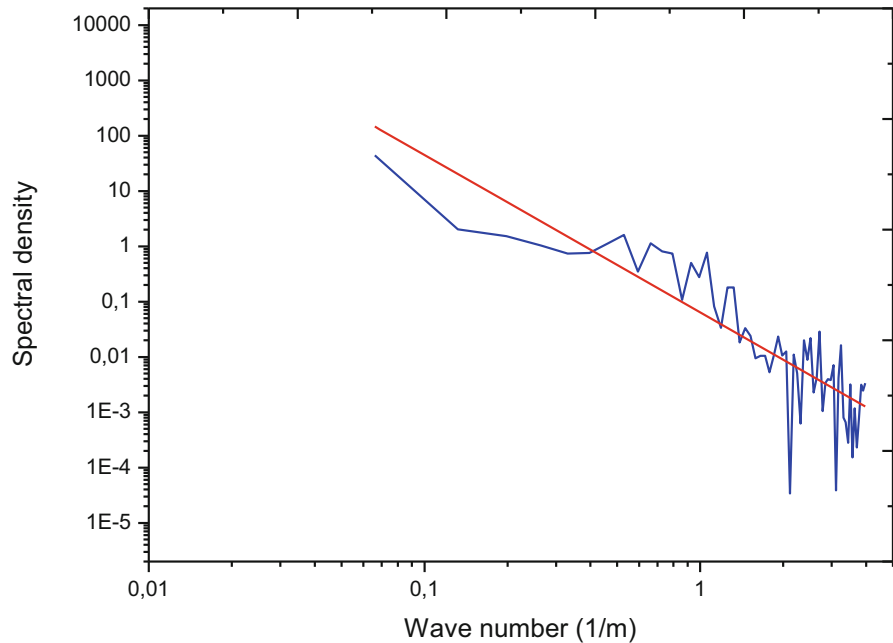
As an example, we will calculate the Hurst exponent of the gamma ray log recorded in a vertical borehole located in the Algerian Sahara, only the depth interval [2230–2245 m] will be processed. Figure 4 shows the fluctuations of the gamma ray log versus the depth. Figure 5 shows the spectral density versus the frequency in the log-log scale, the scaling behavior of the spectral density is clear. This demonstrates the fractal behavior of the gamma ray well-log. A linear regression of logarithm the spectral density versus logarithm the frequency is done. Figure 5 shows the curve of this line (red line). The slope of this line is $\beta = 2.84 \pm 0.02$ so the Hurst exponent is equal to $H = 0.92 \pm 0.01$. The calculated Hurst exponent using the DFA method is $H = 0.60 \pm 0.04$, while the

Hurst Exponent, Fig. 4 Gamma ray log fluctuations for the depth interval [2230–2245 m]



Hurst Exponent,

Fig. 5 Spectral density versus the frequency in log-log scale of the gamma ray log presented in Fig. 4



calculated exponent using the variogram method is $H = 0.93 \pm 0.02$ and the estimated Hurst exponent using the wavelet transform modulus maxima lines is $H = 0.79 \pm 0.004$.

The DFA estimator provides an estimated Hurst exponent which is different to the spectral density, variogram, and WTMM methods, since the DFA is recommended for signals with short length (low number of samples).

A comparison between the detrended fluctuations analysis and variogram on many fBm realizations shows that for signals of high number of samples (more than 1024) variogram estimated are better than the Hurst exponents compared to the DFA. However, for signals of low number of samples we distinguish two cases (Ouafeul and Aliouane 2011):

- (a) For signals with low Hurst exponent, variogram estimates better than the Hurst exponent compared the DFA.
- (b) For signals of High Hurst exponent, DFA estimates better than Hurst exponent.

Conclusions

The Hurst exponent is an important parameter to characterize monofractal signals. It can be used to investigate the long range correlation and it is a measurement of the global regularity. Many methods have been proposed in literature to estimate the Hurst exponent. Each method estimates better the Hurst exponent for special category signals.

An example of a gamma ray well-log recorded in a vertical well located in the Algerian Sahara has been processed and

the Hurst exponent has been estimated, and there is no exact estimated value. Different methods should be compared and the signal characteristics such as length and the range of the Hurst exponent (high/low) should be taken into consideration. The Hurst exponent is very useful in the other branches of geosciences such as soil, geomagnetism, seismic, volcanology, gravity, and oceanography.

Bibliography

- Arneodo A, Bacry E, Muzy J-F (1995) The thermodynamics of fractals revisited with wavelets. *Physica A* 213:232–275
- Audit B, Bacry E, Muzy J-F, Arneodo A (2002) Wavelet-based estimators of scaling behavior. *IEEE Trans Inf Theory* 48(11):2938–2954
- Audit B, Vaillant C, Arnéodo A, D’aubenton-Carafa Y, Thermes C (2004) Wavelet analysis of DNA bending profiles reveals structural constraints on the evolution of genomic sequences. *J Biol Phys* 30: 33–81
- Deutsch V (2003) Geostatistics. In: *Encyclopedia of physical science and technology*, 3rd edn. Academic, pp 697–707. ISBN 9780122274107. <https://doi.org/10.1016/B0-12-227410-5/00869-3>
- Grossmann A, Morlet J-F (1985) Decomposition of functions into wavelets of constant shape, and related transforms. In: Streit L (ed) *Mathematics and physics, lecture on recent results*. World Scientific Publishing, Singapore
- Ouafeul S (2020) Multifractal behavior of SARS-CoV-2 COVID-19 pandemic spread, case of: Algeria, Russia, USA and Italy. <https://doi.org/10.1101/2020.09.16.20196188>
- Ouafeul S, Aliouane L (2011) Multifractal analysis revisited by the continuous wavelet transform applied in lithofacies segmentation from well-logs data. *Int J Appl Phys Math (IJAPM)*. <https://doi.org/10.7763/ijapm.2011.v1.3>
- Parisi G, Frisch U (1985) Turbulence and predictability in geophysical fluid dynamics. In: Ghil M, Benzi R, Parisi G (eds) *Proceedings of the international school of physics “E. Fermi”*. North-Holland, pp 84–87

- Peng C-K, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL (1994) Mosaic organization of DNA nucleotides. *Phys Rev E* 49: 1685–1689
- Peng C-K, Havlin S, Stanley HE, Goldberger AL (1995) Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos* 5:82–87

Hyperspectral Remote Sensing

Daniele Marinelli¹, Francesca Bovolo² and Lorenzo Bruzzone¹

¹Department of Information Engineering and Computer Science, University of Trento, Trento, Italy

²Center for Digital Society, Fondazione Bruno Kessler, Trento, Italy

Definition

Hyperspectral (HS) imaging is a subcategory of spectral imaging which is the process of acquiring spatially distributed and contiguous data (images) measuring the power reflected within each resolution cell (i.e., pixel) as a function of a wavelength (i.e., performing spectral measurements to analyze how light interacts with the target). While commercial digital cameras acquire images in three spectral channels corresponding to the red, green, and blue wavelengths (i.e., RGB imaging), spectral imagers such as multispectral (MS) and HS sensors perform the acquisition in multiple bands across the electromagnetic spectrum (in the order of tens or hundreds of spectral channels). MS sensors usually acquire data in a small number of spectral channels with a broad spectral resolution whereas an HS sensor acquires a much larger number of spectral bands (typically in the order of the hundreds) with a very high spectral resolution. Therefore, HS imagers perform a very dense sampling of the spectral signature, which can be seen as a fingerprint of the measured target. Due to these characteristics, HS imaging is used in a wide variety of fields including art conservation, food security, geology, and Earth observation. In Earth observation, HS images are acquired from both airborne (e.g., UAV and planes) and spaceborne platforms and are used in several applications such as precision agriculture, forestry, and urban area monitoring.

Hyperspectral Sensors

HS imagers are passive sensors using an external illumination source to illuminate the target and measure the reflected radiation as a function of the wavelength, splitting the at-sensor radiation into multiple spectral channels where

each spectral channel, or band, represents a narrow wavelength range. This can be equated to perform a spectroscopy measurement thus acquiring and sampling the spectral signature of the surface portion included in each spatial resolution cell (a pixel in the image). The reflected radiance is measured in a given wavelength range, which typically can extend from the visible to the thermal infrared.

In HS imagers, the incident radiation is split and measured for dozens or hundreds of spectral channels requiring ad hoc optical components and detectors. Few HS sensors are capable of performing full frame acquisitions (i.e., acquire the whole scene in one shot), but to this date, their use is limited mainly to astronomy. Most of the HS imagers do not acquire the whole spatial scene in one shot, as for typical RGB cameras, but perform multiple 2-D acquisitions by means of spatial or spectral scanning and aggregate them to generate the final image. This allows for the use of imagers with simpler optical components and less detectors, compared with a full frame sensor with the same spatial and spectral resolutions. In spectral scanning, each 2-D acquisition represents the whole scene at a given wavelength. The spectral scan is performed by sequentially acquiring the whole scene using tunable passband filters that filter out all the wavelengths except the range of one spectral channel. This requires no relative movements between the sensor and the target to avoid spatial misregistration among different spectral channels. In spatial scanning, the sensor, which is moving with respect to the target, acquires at each shot a portion of the scene (one or more pixels) collecting the whole spectrum all at once for each pixel imaged along a scan line. Let along-track be the direction of relative movement of the sensor with respect to the scene, and across-track be the normal direction to the along-track. To acquire the whole 2-D scene, spatial scanners exploit the relative movement between the sensor and the target to acquire in the along-track direction while they use different techniques to scan in the across-track direction. The three main technologies for such scanning are the line, whiskbroom, and pushbroom scanners. Line and whiskbroom scanners use a rotating mirror to scan in the across-track direction while acquiring one or few pixels at a time, respectively. While such systems have been used for HS imaging (e.g., the whiskbroom scanner of the AVIRIS sensor (Green et al. 1998)), pushbroom scanners have become the most common given that they do not require a rotating mirror. Indeed, pushbroom scanners scan the complete field of view in the across-track direction using a linear array of detectors with a slit aperture to acquire one line of the scene. In order to acquire the whole spectrum for each pixel, spatial scanners use an optical dispersive element (e.g., a prism or a diffraction grating) to disperse the incident broadband spectrum onto a 2-D array of detectors. Each detector measures the dispersed light in given wavelength range for a given spatial position. Therefore, for each line acquisition, a 2-D data matrix is

generated where one dimension represents the across-track (spatial dimension) whereas the other corresponds to the spectral channels (spectral dimension). Accordingly, each column (or line, depending on the sensor orientation) of detectors collects the dispersed light for one spatial pixel. The aggregated measurements of one column (or line) of detectors represent an approximation of the spectral signature of one spatial pixel. The resulting HS image can be seen as a 3-D cube defined by the triplet of coordinates (x, y, λ) , where (x, y) are the spatial coordinates, and $\lambda \in [1, B]$ is the spectral coordinate, where typically the number of spectral channels B is in the order of the hundreds. Each 2-D array of a spectral channel represents the acquired scene in one of the spectral channels (i.e., fixed λ). Figure 1 shows a simplified scheme of a pushbroom HS scanner. Spatial scanners are widely used in Earth observation since they take advantage of the motion of the platform (e.g., aircraft or satellite) with respect to the ground to perform the scanning. These scanners are used also in the opposite configuration with the sensor being fixed while the target moves (e.g., an object on a conveyor belt). This is the case of industrial applications.

Two important parameters of scanner are the spectral and spatial resolution. Let us analyze these concepts for spatial scanners mounted on moving platforms such as aircraft. However, they hold also when the scanner is fixed and the target is moving. A critical parameter of any dispersive element is the angular dispersion which corresponds to the

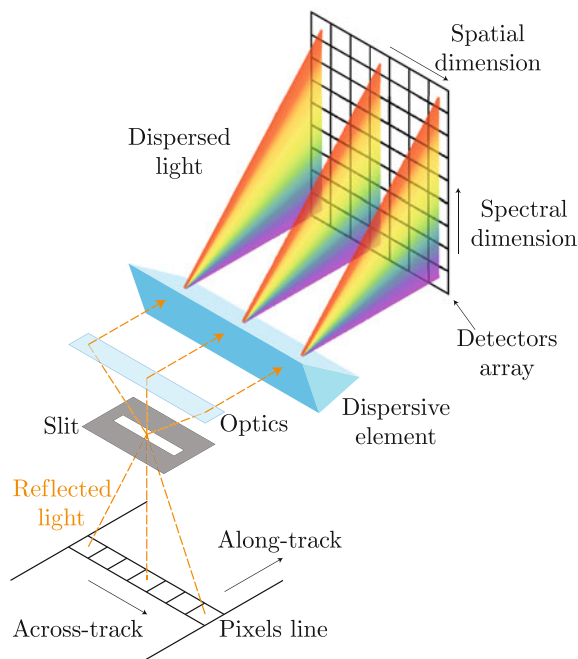
angular range of the incident radiation dispersion. The angular dispersion and the detector density (i.e., the number of detectors per surface unit) define the spectral resolution, which is the minimum wavelength range that can be distinguished by the sensor. Indeed, given a fixed angular dispersion, the detector density determines in how many spectral channels the dispersed light is separated. The spatial resolution is defined by the detector characteristic (width w and focal length f), the acquisition settings (acquisition rate r), the sensor-target distance, and flight parameters (altitude h with respect to the ground and platform speed s). The scanner resolution in the across-track direction is commonly referred as the instantaneous field of view (IFOV) and is defined as

$$\text{IFOV} \approx 2 \arctan\left(\frac{\omega}{2f}\right) \approx \frac{\omega}{f}. \quad (1)$$

To obtain the resolution at the target, i.e., the ground instantaneous field of view (GIFOV), the sensor-target distance must be considered:

$$\text{GIFOV} \approx 2h \tan\left(\frac{\omega}{2f}\right) \approx \frac{\omega h}{f}. \quad (2)$$

The along-track spatial resolution can be approximated as s/r . If feasible, the platform speed and/or the acquisition rate are set so that the along-track resolution is equal to the across-track to obtain square pixels. These resolutions cannot be increased indefinitely since an increase of spatial or spectral resolution leads to less energy reaching each detector. This requires to take into account the radiometric resolution of the detector: the minimum amount of energy that can be detected with an acceptable signal to noise ratio. The radiometric resolution requires a trade-off between spatial and spectral resolution. Therefore, given a specific detector, to increase spectral resolution, the spatial one has to be reduced to satisfy the detector radiometric constraint. Equation (2) shows that a way to improve the spatial resolution is to decrease the distance of the sensor from the scene which is one of the main reasons why HS sensors have been mounted mainly on airborne platform such as for the AVIRIS sensor (Green et al. 1998) rather than satellites. To guarantee both high spatial and spectral resolution on spaceborne sensors (where h is large) would require either very small detectors (thus violating the radiometric resolution constraints) or a large focal length obtaining a very large optical part of the sensor, making it unfeasible to install on a satellite. Indeed, the spatial resolution of existing and planned spaceborne HS sensors is typically in the order of the tenths of meters such as the 30 m of the recently launched (2019) PRISMA mission (Candela et al. 2016). In contrast, the spatial resolution of sensors mounted



Hyperspectral Remote Sensing, Fig. 1 Illustration of the measure concept implemented by an HS pushbroom sensor with 7 across-track pixels and 9 spectral channels. The three orange dashed lines represent the measured reflected light for the leftmost, center, and rightmost pixels of the scan line

on airborne platforms is in the order of the meters or even centimeters due to the recent new UAV-mounted sensors. However, spaceborne platforms have significant advantages which are mainly related to the repeat-pass nature of satellites that guarantees a continuous coverage in time at the global scale at no extra cost while airborne platforms require expensive ad hoc flights for each new scan.

Comparison with Multispectral Sensors

While HS sensors belong to the same family of imagers as multispectral (MS) ones, i.e., spectral imagers, there are significant differences between the two types of sensors. The distinction resides in the number of spectral channels (typically in the order of the hundreds) but most importantly in the properties of the spectral channels. Given a wavelength range, an HS imager performs an almost continuous and uniform sampling of the spectrum using uniformly spaced and narrow spectral channels across the whole range. As an example, the PRISMA sensor samples the 400–2500 nm range with 240 spectral channels having roughly the same spectral width of 10 nm. Note that some HS sensors have different spectral width for different spectral ranges such as for the future EnMAP instrument which will have a spectral sampling 6.5 nm for the visible and near infrared (VNIR) and 10 nm for the short-wave infrared (SWIR). The bands of an MS sensor are much wider (in the order of tens or hundreds of nm), have usually significantly different widths, and are non-uniformly distributed across the spectral range. Being a small number, they are positioned to sample specific portions of the spectrum. This is the case of the Sentinel-2 MultiSpectral Instrument (MSI) which has 13 spectral channels in the 400–2500 nm range with different widths and mostly distributed in the 400–1000 nm range.

Main Disturbances and Processing Challenges

HS imagers generate data cubes with a large informative content due to the fine sampling of the spectrum. While this provides significant advantages compared to MS data in terms of information content and potential applications, it comes at the cost of an increased amount of data volume and complexity that should be properly taken into account. The main challenges are related to:

- (i) The low-level preprocessing to model and possibly mitigate or remove distortions and disturbances introduced during the acquisition
- (ii) The high-level definition of a processing chain that takes into account the peculiar properties of HS data to effectively extract information

In the following, we will focus on HS imagers mounted on airborne or spaceborne platforms.

Distortions and disturbances in HS imaging are generated by both internal sources (e.g., instrument noise and optical distortions) and external sources (e.g., atmospheric effects). In addition, also geometrical disturbances (common also in MS scanners) are introduced during acquisition. These are mainly related to:

- (i) Disturbances of the platform motion (especially for UAVs)
- (ii) Acquisition geometry
- (iii) Surface morphology

These can be mitigated if a precise knowledge of position and attitude of the platform during the acquisition (acquired by an IMU/GNSS system) and of the surface morphology (i.e., Digital Terrain Model) are available.

Regarding internal sources, it is possible to distinguish between radiometric disturbances introduced by the electronics and distortions introduced by the optical components. Radiometric noise and disturbances are caused by the non-ideal behavior of the electronics and the acquisition conditions (i.e., temperature). One of the most relevant issues is the uneven responsivity of the detectors across the 2-D array leading to disturbances of the resulting HS image such as striping effects. These disturbances can be mitigated during preprocessing if the responsivity of each detector is known. Such information can be recorded and stored, before the actual data acquisition, in a responsivity matrix that can be acquired, for example, by positioning the spectrometer in an integrating sphere to provide a spatially uniform white scene of which the absolute radiance is known. Another important sensor characterization is the background noise level matrix which provides the background noise introduced by each detector. Since this noise depends on the acquisition conditions (e.g., temperature), some HS sensors are equipped to characterize the background noise level before each scan performing one or more acquisitions with closed shutter (i.e., acquisition in complete darkness) so that only noise is recorded. Such information is stored with each scan and used during the preprocessing to mitigate the effect of the noise source.

Among the distortions introduced by optical components and their misalignment, two of the most critical issues are the smile and keystone effects which are common in HS pushbroom scanners. The smile effect (spectral misregistration) is a shift of the spectral channels wavelengths as a function of the spatial position due to distortions caused by the dispersive elements or other optical components. If smile is present, a spectral line across the whole field of view is seen as a curve bent into the shape of a smile, hence the name. The keystone effect (spatial misregistration) is the

variation of the spatial position as a function of wavelength, and it is caused by a wavelength-dependent magnification of the radiation onto the detector array. This makes the same object to appear in different positions of the image depending on the wavelength. Such distortions can significantly affect the spectral/spatial information in HS data leading to a decrease of performance in the processing chain. Indeed, keystone can strongly change the spectral signatures especially on the boundaries between different materials (i.e., in areas with strong spatial variability) mixing spectra in an unrealistic way. These effects can be corrected, or at least mitigated, during preprocessing by characterizing them across the whole detector matrix. Such characterization can be obtained by directly analyzing the HS data cube as in Neville et al. (2004) where the spatial correlation of spectral channels is used to model the keystone. Alternatively, precise measurements can be conducted in a controlled environment to produce smile and keystone characterization matrices to be used in the preprocessing to reduce the impact of these distortions. This can be done using an integrating sphere to illuminate the imager with narrow band (for smile) or broadband (for keystone) light sources to measure the magnitude of the distortions across the whole detector matrix.

Regarding effects of the atmosphere, depending on the acquisition conditions such as altitude, water vapor column, or the presence of aerosols, the radiance measured at the sensor (i.e., at-sensor radiance) can be altered with respect to the at-surface radiance, leading to measurements where the spectral signatures are noticeably different with respect to the actual reflectance. Algorithms that take as input several parameters such as the time and location of the acquisition can be used to model and remove atmospheric disturbances. They include the ATmospheric REMoval (ATREM) (Gao and Davis 1997) and the Fast Line-of-sight Atmospheric Analysis of Spectral Hypercubes (FLAASH) (Cooley et al. 2002) algorithms.

Moving to the processing, one of the challenges is related to the high number of spectral channels that results in a high dimensional feature space where data should be analyzed. The analysis of data in high dimensional spaces is affected by several problems that are grouped under the name curse of dimensionality. Indeed, since the number of samples (pixels in the case of HS images) is finite, as the number of features (i.e., spectral channels) increases, the volume of the feature space increases quickly with the data becoming more and more sparse. This leads to a decrease of effectiveness of techniques developed for low dimensional spaces such as for distance metrics where in high dimensional spaces, the distances between different pairs of coordinates tend to become all similar (Bellman 2015). Another problem of high dimensional spaces like those defined by HS data is related to supervised data analysis techniques related to the Hughes phenomenon (Melgani and Bruzzone 2004). This

phenomenon describes the unintuitive decrease of accuracy as the number of features given as input to a classifier goes above a certain value for a fixed number of training samples. It is caused by the small ratio between the number of training samples and the number of features that prevents a reliable modelling of the statistical properties of the data. The high dimensionality of HS data introduces also technical challenges related to the data volume given by the large number of values to be stored for each pixel typically with high radiometric resolutions (i.e., large number of bits representing pixels value) that may lead to problems related to both storage, computational load, and memory needs.

Main Methodologies and Applications in Earth Observation

HS images can be used in a wide variety of applications where a detailed knowledge of the spectral properties of a scene or object is required. Some of them are:

- (i) Earth observation
- (ii) Food security
- (iii) Cultural heritage preservation
- (iv) Civil engineering
- (v) Medical imaging

In Earth observation, HS data are used in multiple applications including agriculture, forestry, mining, water, and cryosphere monitoring. HS data are analyzed by different methodological approaches having different objectives, among which the most common are classification (Ghamisi et al. 2017), target detection (Nasrabadi 2014), spectral unmixing (Bioucas-Dias et al. 2012), and in recent years, change detection (Liu et al. 2019). Classification exploits the detailed spectral information of HS data to discriminate classes showing very similar yet different spectral signatures that result inseparable at the spectral resolution of MS data. Therefore, HS images allows for the generation of detailed classification maps with a wide variety of classes. Classification of HS data is used in different applications and environments including:

- (i) Urban mapping
- (ii) Mineral mapping
- (iii) Agriculture for the automatic identification of crops
- (iv) Forestry for the generation of tree species maps

If more images are available for the same area, it is possible to exploit HS data to monitor the dynamics of the land cover at a fine spectral resolution. Indeed, while methods for change detection in MS data focuses mainly on abrupt changes that correspond to a strong intensity variation over a large range of

the spectrum (e.g., transition from vegetation to concrete), using HS multitemporal data is possible to detect changes associated to variations that affect only limited portions of the spectrum and are invisible to MS sensors. Similar considerations hold in the case of multiple change detection (i.e., identification and discrimination among change types) where HS data allows us to separate changes that differ only in limited ranges of wavelengths. Change detection methods for HS data can be used for several applications such as to monitor open-pit mines, track the crops evolution in agricultural areas, and forestry monitoring.

Another set of methodologies for HS data processing is spectral unmixing. Since especially for spaceborne data the spatial resolution is limited, it is common that pixels contain more than one type of material resulting in a spectrum which is a mix of the spectra of the different materials. The aim of unmixing techniques is to automatically separate the mixed spectra to identify the materials (i.e., endmembers) and their abundance in the pixel. A critical aspect of unmixing is the need of formulating a model for the mixture. This can be linear or nonlinear with the first being the most common due to the easier implementation. The linear model describes the mixture as a weighted sum of the individual spectra where the weights correspond to the abundance value. In this case, the unmixing can be performed using, for example, least square methods given that the spectra of the endmembers are known. These can be done by extracting them from the pure pixels in the image or using existing spectral libraries.

It is worth noting that while these are the most common methodologies and applications, Earth observation HS data have several other uses. In the case of agriculture, forestry, or vegetation monitoring, HS data are used to monitor specific aspects. They include the monitoring of several plants health factors such as water stress and the diffusion of diseases. HS data are also used for water quality analysis generating different products such as the concentration of chlorophyll and total suspended matter. Another application field is related to the monitoring of glaciers and snow both in terms of classification and the extraction of specific parameters such snow grain size, wetness, and surface contamination. Such parameters provide valuable information as they can be used to study the status of glacier and predict the melting rate.

Summary and Conclusions

Hyperspectral images provide a very detailed representation of the spectral reflectance of the target surface. Such data have been used in several applications such as land cover maps generation and agricultural monitoring and methodologies including classification and change detection where the dense sampling of the spectrum can be used to separate classes or changes associated to very similar, yet semantically

different, spectral signatures which are undistinguishable in MS images. Indeed, compared to MS, sensors which sample spectrum at few broad bands, HS imagers perform a dense sampling of the spectrum across the whole spectral range of interest. However, the analysis of HS images requires to address several challenges related both to disturbances and distortion introduced during the acquisition and the processing of high dimensional data.

While to this date HS images are still relatively less used than MS images, this will significantly change in the next years due to the current new sensors, future HS spaceborne missions, and the new generation of compact HS imagers designed specifically for UAV systems. Accordingly, HS data will become more and more available and thus their use in operational scenarios will significantly increase.

Cross-References

- [Geostatistics](#)
- [Machine Learning](#)
- [Remote Sensing](#)
- [Spectral Analysis](#)

Bibliography

- Bellman RE (2015) Adaptive control processes: a guided tour. Princeton university press
- Bioucas-Dias JM, Plaza A, Dobigeon N, Parente M, Du Q, Gader P, Chanussot J (2012) Hyperspectral unmixing overview: geometrical, statistical, and sparse regression-based approaches. *IEEE J Selected Topics Appl Earth Obs Remote Sens* 5(2):354–379. <https://doi.org/10.1109/JSTARS.2012.2194696>
- Candela L, Formaro R, Guarini R, Loizzo R, Longo F, Varacalli G (2016) The prisma mission. In: 2016 IEEE international geoscience and remote sensing symposium (IGARSS), pp 253–256. <https://doi.org/10.1109/IGARSS.2016.7729057>
- Cooley T, Anderson GP, Felde GW, Hoke ML, Ratkowski AJ, Chetwynd JH, Gardner JA, Adler-Golden SM, Matthew MW, Berk A, Bernstein LS, Acharya PK, Miller D, Lewis P (2002) Flaash, a modtran4-based atmospheric correction algorithm, its application and validation. In: IEEE international geoscience and remote sensing symposium, vol 3, pp 1414–1418. <https://doi.org/10.1109/IGARSS.2002.1026134>
- Gao BC, Davis CO (1997) Development of a line-by-line-based atmosphere removal algorithm for airborne and spaceborne imaging spectrometers. In: Descour MR, Shen SS (eds) *Imaging spectrometry III*, international society for optics and photonics, vol 3118. SPIE, pp 132–141. <https://doi.org/10.1117/12.283822>
- Ghamisi P, Plaza J, Chen Y, Li J, Plaza AJ (2017) Advanced spectral classifiers for hyperspectral images: a review. *IEEE Geosci Remote Sens Mag* 5(1):8–32. <https://doi.org/10.1109/MGRS.2016.2616418>
- Green RO, Eastwood ML, Sarture CM, Chrien TG, Aronsson M, Chipendale BJ, Faust JA, Pavri BE, Chovit CJ, Solis M, Olah MR, Williams O (1998) Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (aviris). *Remote Sens Environ* 65(3): 227–248. [https://doi.org/10.1016/S0034-4257\(98\)00064-9](https://doi.org/10.1016/S0034-4257(98)00064-9)
- Liu S, Marinelli D, Bruzzone L, Bovolo F (2019) A review of change detection in multitemporal hyperspectral images: current techniques,

- applications, and challenges. *IEEE Geosci Remote Sens Mag* 7(2): 140–158. <https://doi.org/10.1109/MGRS.2019.2898520>
- Melgani F, Bruzzone L (2004) Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans Geosci Remote Sens* 42(8):1778–1790. <https://doi.org/10.1109/TGRS.2004.831865>
- Nasrabadi NM (2014) Hyperspectral target detection : an overview of current and future challenges. *IEEE Signal Process Mag* 31(1): 34–44. <https://doi.org/10.1109/MSP.2013.2278992>
- Neville RA, Sun L, Staenz K (2004) Detection of keystone in imaging spectrometer data. In: Shen SS, Lewis PE (eds) *Algorithms and technologies for multispectral, hyperspectral, and ultra spectral imagery X*, international society for optics and photonics, vol 5425. SPIE, pp 208–217. <https://doi.org/10.1117/12.542806>

Hypothesis Testing

Rogério G. Negri
São Paulo State University (UNESP), Institute of Science and Technology (ICT), São José dos Campos, São Paulo, Brazil

Synonyms

Significance testing; Statistical hypothesis testing; Statistical test

Definition

A hypothesis test comprises a procedure aimed to prove, or not, a statistical statement or conjecture.

A statistical hypothesis $\mathcal{H}$ comprises a statement or conjecture made about one or several statistical elements like random variables, populations, distributions, or even parameters. Therefore, a hypothesis test stands for a procedure performed to decide whether $\mathcal{H}$ should be rejected or not.

Generically, let us suppose a data space $\mathcal{X}$ where it is observed a sample $\{X_i : i = 1, \dots, n\}$ whose elements are realizations of a random variable $\mathbf{X}$. Such sample is observed aiming to infer about particular propriety stated by $\mathcal{H}$. A critical region $C_\tau \subset \mathcal{X}$ represents extreme situations stated according to a statistic of test τ . Whether verified that $\{X_1, \dots, X_n\}$ comes from the critical region, the hypothesis $\mathcal{H}$ is rejected.

Usually, $\mathcal{H}$ is unfolded into a null ($\mathcal{H}_0$) and an alternative ($\mathcal{H}_1$) hypothesis. While $\mathcal{H}_0$ states that there is no difference between the statistical elements under comparison (i.e., the difference is null), the hypothesis $\mathcal{H}_1$ contradicts the previous one. In the light of these discussions, if $\mathcal{H}_0$ is assumed as false, then $\mathcal{H}_1$ becomes valid and $\mathcal{H}_0$ is rejected; in turn, if $\mathcal{H}_0$ is considered true, then $\mathcal{H}_1$ is false and $\mathcal{H}_0$ should not be rejected. Still, when $\mathcal{H}_1$ only establishes that there is a

difference between the compared elements, the hypothesis test is bilateral (or two-sided). On the other hand, when the difference has a direction (i.e., positive or negative), the hypothesis test is unilateral (or one-sided).

Regarding the process to decide about the rejection of $\mathcal{H}_0$, it may occur a type I error, when it rejects $\mathcal{H}_0$ although it is a valid hypothesis; or a type II error, when it does not reject $\mathcal{H}_0$ although it is not valid. In other words, a type I error (also called false positive error) means that the conclusions point to reject $\mathcal{H}_0$, but in fact, such rejection occurred by chance. The type II error (also called false negative error) implies no statistical difference, although such difference is a reality.

The occurrence probability of type I and type II errors are usually denoted by α and β , respectively. Furthermore, α is called as the “significance of the test” and the probability $(1 - \beta)$ is the “power of the test.” β is computed through a theoretical probability distribution that expresses the values from τ that leads to the rejection of $\mathcal{H}_0$.

Without loss of generalization, suppose one tests the hypothesis that a parameter θ of a population differs from θ_0 obtained through an observed sample. A hypothesis test may be structured according to the following steps (Mood et al. 1974):

- (i) Define the hypothesis:

$$\mathcal{H}_0 : \theta = \theta_0$$

$$\mathcal{H}_1 : \theta \neq \theta_0 \text{ if bilateral; } \theta \leq \theta_0 \text{ (if unilateral)}$$

- (ii) Identify the appropriate statistic of test τ .
- (iii) Specify the significance of test α and define the critical region C_τ to reject $\mathcal{H}_0$.
- (iv) Compute τ and decide whether $\mathcal{H}_0$ is rejected.

Additionally, after selecting τ and computing the value $\hat{\tau}$ with basis on $\{X_1, \dots, X_n\}$, it is also possible to determine the probability p of occurring a value equal or more extreme than $\hat{\tau}$. Such probability is called p -value and leads to the rejection of $\mathcal{H}_0$ when $p < \alpha$.

Brief Overview

One of the first statistical inference techniques, the so-called “parametric” tests, is usually based on assumptions regarding the distribution of the population. Posterior, “non-parametric” techniques rise as an alternative approach that is independent of previous suppositions. Regardless of its parametric or non-parametric basis, the hypothesis tests in the literature may be categorized as, but not limited to, tests for a single sample; two paired or independent samples; several paired or independent samples; correlation; concordance; and time-series.

While tests for a single sample focus on checking if a set of observations is adherent to a specific distribution or has an expected parameter, the tests for two or several samples compare if such samples are, or not, similar. The denomination “paired” implies that the samples are element-wise compared. “Independent” means that the comparisons do not consider the order of the elements or suppose a relation among them. Correlation and concordance tests check the significance of correlation values computed through two or multiple variables. Lastly, tests for time series allow proving the existence of trends in the data over time.

There are several hypothesis tests in the literature, making it unfeasible to cover this topic herein. Therefore, the following discussions embrace just some most popular tests. More details are given for tests of a single sample as well as for two samples.

Tests for a Single Sample

The T -test is a parametric test to check if a sample differs from the population's mean value. This test is conditioned to the suppositions that the population follows a Gaussian distribution; the sample is small ($n \leq 30$); the observations were randomly and independently drawn; and the population standard deviation is unknown. The T -test statistic is expressed by:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \quad (1)$$

where μ and $\bar{X}$ stands for the mean value from the population and the estimated mean from the sample, respectively; s and n represent the standard deviation and the number of observations from the sample, respectively. Once it follows a Student's T distribution with $(n - 1)$ freedom degrees, the rejection of $\mathcal{H}_0$ occurs when t lies the critical region defined according to the adopted significance α . Tabulated values or computed through the integral of Student's T distribution function may be employed to obtain the critical region.

When the population standard deviation σ is known and the sample is not small ($n > 30$), the Z -test is the alternative to T -test. The statistic z follows a standard Gaussian distribution and is defined replacing s by σ at Eq. 1.

Tests for Two Samples

Given the samples $X = \{X_1, \dots, X_n\}$ and $Y = \{Y_1, \dots, Y_m\}$, a variant of T -test allows comparing the mean values from two samples. The statistic t for this case is defined

as:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{s_X^2/n + s_Y^2/m}} \quad (2)$$

where $\bar{X}, \bar{Y}$, s_X , and s_Y are the means and standard deviations computed through the samples X and Y , which are composed by n and m observations, respectively. The same suppositions made for the single sample version still kept.

The statistic t follows a Student's T distribution with $(n + m - 2)$ freedom degrees. The critical region definition and rejection of $\mathcal{H}_0$ are similar to its single sample version.

In analogy, knowing the standard deviations σ_X and σ_Y regarding the populations where X and Y are extracted, the comparison of its mean values is carried through the Z -test for two samples. Again, the statistic of this test follows a standard Gaussian distribution and is defined replacing s_X and s_Y by σ_X and σ_Y in Eq. 2.

When the focus is to check the similarity between the variances of two populations, with supposed Gaussian distributions, the F -test may be applied. Its statistic of test f , defined in Eq. 3, follows a Snedecor F -distribution with parameters $(n - 1)$ and $(m - 1)$ freedom degrees.

$$f = \frac{(m - 1) \sum_{i=1}^n (X_i - \bar{X})^2}{(n - 1) \sum_{j=1}^m (Y_j - \bar{Y})^2} \quad (3)$$

A non-parametric alternative to the T -test is given by the U Mann-Whitney test. In this case, supposing that the samples X and Y has at least 20 observations, the u statistic expresses the Mann-Whitney's test:

$$u = \frac{U - nm/2}{\sqrt{(nm(n + m + 1))/12}} \quad (4)$$

with $U = \min \{U_X, U_Y\}$ where $U_X = nm + \frac{n(n+1)}{2} - R_X$, $U_Y = nm + \frac{m(m+1)}{2} - R_Y$; R_X is obtained by the sum of ranks assigned to the observations in X after a joint regarding both observations from X and Y ; R_Y is computed in analogy to R_X . If the samples does not attain the minimum required size, distinct procedures, as discussed in Siegel and Castellan (1988) and Sheskin (2011), should be adopted. The statistic u follows a standard Gaussian distribution, and then such distribution rules the critical region definition and $\mathcal{H}_0$ rejection.

In the case of paired comparison, the Wilcoxon test is a non-parametric procedure that can be employed to determine whether or not two samples X and Y represent the same population. When the analysis involves small samples, this test embraces: (i) calculate the difference d_i relative to the

scores of the pair (X_i, Y_i) ; (ii) assign a rank to each d_i regardless of its sign; (iii) include the sign to the ranks; and (iv) calculate the statistic r that corresponds to the sum of the ranks of the same sign. Finally, the probability of occurrence of r is obtained through tabulated distribution (Siegel and Castellan 1988; Sheskin 2011). However, when the number of pairs is greater than 25, the statistic of test w , which comes from the sum of the ranks follows a standard Gaussian distribution, is expressed by:

$$w = \frac{r - (n(n+1))/4}{\sqrt{(n(n+1)(2n+1))/24}} \quad (5)$$

where n is the number of observations in both samples X and Y . Once again, the decision to reject $\mathcal{H}_0$ is still in the same mood as previous discussions.

Tests for Several Samples, Correlation, Concordance, and Time Series

Hypothesis tests for comparisons involving several samples are not performed essentially by combining tests for two samples applied to each pair of samples. On the assumption of normal distribution of k samples, whose variances are equal, the analysis of variance (ANOVA) allows comparing the means of these samples. In this test, $\mathcal{H}_0$ is characterized as a simultaneous equivalence among the means of all samples (i.e., $\mathcal{H}_0 : \mu_1 = \mu_2 = \dots = \mu_k$), and $\mathcal{H}_1$ establishes that there is difference at least between a pair of samples. This technique has different versions and may be applied on paired and independent samples. The statistic of the test resulting from the ANOVA follows Snedecor's F-distribution. For this situation, when the data does not respect the assumptions required by the parametric approaches, the Friedmann and Kruskal-Wallis tests are the alternatives to deal with paired and independent samples, respectively.

Correlation coefficients are applied to measure the relationship between variables to explain how a variable behaves in a scenario where the other changes. The Pearson and Spearman coefficients are, respectively, the parametric (it assumes that a bivariate Gaussian distribution expresses a joint distribution of the variables) and non-parametric ways of measuring the correlation between variables. These correlation measures range in $[-1, +1]$, where the lower and upper bounds indicate perfect inverse and direct correlations, respectively. When these measures approach zero, it implies that there is no correlation between the variables. In addition to calculating and interpreting the correlation values, a hypothesis test can be applied to verify that such value is significant to the null coefficient (i.e., no correlation).

Correlation coefficients involving more than two variables can be computed using Kendall's coefficient.

Agreement coefficients extracted from contingency tables (confusion matrices) that express the relation between expected and observed quantities according to different classes are also subject to hypothesis tests. Specifically, values of coefficients such as kappa (κ) and tau (τ) from different confusion matrices may have their significance verified through statistic of tests in the form $\frac{c_1 - c_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$ where c_1 and c_2 represent the agreement coefficients (of common type) and σ_1^2 and σ_2^2 of the variances associated with the respective coefficients. The presented statistic follows standard Gaussian distribution.

Finally, regarding the analysis of time series, identifying increasing or decreasing trends is of great importance. In this context, the Mann-Kendall test allows this verification where the $\mathcal{H}_0$ and $\mathcal{H}_1$ hypotheses imply the absence and existence of a trend in the observed period, respectively.

Applications in Geosciences

A hypothesis test is an essential tool in data analysis processes. Regarding the geosciences, several studies employ this concept.

In Luo and Yang (2017), hypothesis tests are used to verify the effectiveness of a sensor network developed to detect the concentration of water pollution. In this case, the T -test is applied to identify deviations from expected behavior and then reveal the presence of the pollutant source. In Eberhard et al. (2012), the Wilcoxon test is used to evaluate the accuracy of earthquake forecasting models. Similarly, Jena et al. (2015) employs the Z -test, among others, to evaluate the precipitation and climate change models. Finally, Nascimento et al. (2019) develop a new hypothesis test to perform temporal change detection in the Earth's surface using polarimetric images obtained by remote sensing.

Summary

A hypothesis test is a fundamental statistical procedure designed to check statistical elements like random variables, populations, distributions, or parameters that behave according to a statistical hypothesis.

Cross-References

- [Probability Density Function](#)
- [Random Variable](#)

Bibliography

- Eberhard DAJ, Zecher JD, Wiemer S (2012) A prospective earthquake forecast experiment in the western pacific. *Geophys J Int* 190(3): 1579–1592. <https://doi.org/10.1111/j.1365-246X.2012.05548.x>
- Jena P, Azad S, Rajeevan MN (2015) Statistical selection of the optimum models in the CMIP5 dataset for climate change projections of Indian monsoon rainfall. *Climate* 3(4):858–875. <https://doi.org/10.3390/cli3040858>
- Luo X, Yang J (2017) Water pollution detection based on hypothesis testing in sensor networks. *J Sens* 2017:3829894. <https://doi.org/10.1155/2017/3829894>
- Mood AM, Boes DC, Graybill FA (1974) Introduction to the theory of statistics, 3rd edn. McGraw-Hill
- Nascimento ADC, Frery AC, Cintra RJ (2019) Detecting changes in fully polarimetric SAR imagery with statistical information theory. *IEEE Trans Geosci Remote Sens* 57(3):1380–1392. <https://doi.org/10.1109/TGRS.2018.2866367>
- Sheskin DJ (2011) Handbook of parametric and nonparametric statistical procedures, 5th edn. Chapman & Hall/CRC
- Siegel S, Castellan N (1988) Nonparametric statistics for the Behavioral sciences. McGraw-Hill international editions. Statistics series. McGraw-Hill

Hypsometry

Sebastiano Trevisani¹ and Lorenzo Marchi²

¹Università Iuav di Venezia, Venice, Italy

²Consiglio Nazionale delle Ricerche – Istituto di Ricerca per la Protezione Idrogeologica, Padova, Italy

Definition

In very general terms, hypsometry is the measurement of land elevation. In geomorphology, hypsometry refers to the analysis of the cumulative distribution of the elevation, absolute or relative, in a given region. The distribution of elevation is a basic topographic feature of a territory, and its representation is useful for interpreting and characterizing geomorphic processes. Hypsometry can be computed in spatial domains of any shape and extent, including the whole earth surface; however, often the studies of hypsometry focus on drainage basins.

Hypsometric Parameters

Hypsometric curves, which represent the fraction of area above a given elevation, are a widely adopted tool for the graphical representation of hypsometry. The hypsometric curve is the cumulative distribution of elevation in a defined spatial domain of analysis (Hypsometry, Fig. 1). Hypsometric curves are often presented in nondimensional form

(Hypsometry, Fig. 2), which enables the comparison between different regions or drainage basins. Nondimensional hypsometric curves are plots of the relative height (h/H) versus the relative area (a/A) of the region under investigation (Hypsometry, Fig. 1). The relative height is the ratio of the height above the lowest point in the basin at a given contour to the total relief ($H=H_{max}-H_{min}$), and the relative area is the ratio of horizontal cross-sectional area at that elevation to the total area (i.e., the frequency of values above a given threshold of elevation h).

Hypsometry requires the assessment of the areas at different elevation belts, which was done in the past through manual measurements on topographic maps with contour lines, and is now easily implemented through the analysis of digital terrain models (DTMs) or, more in general, digital elevation models (DEMs).

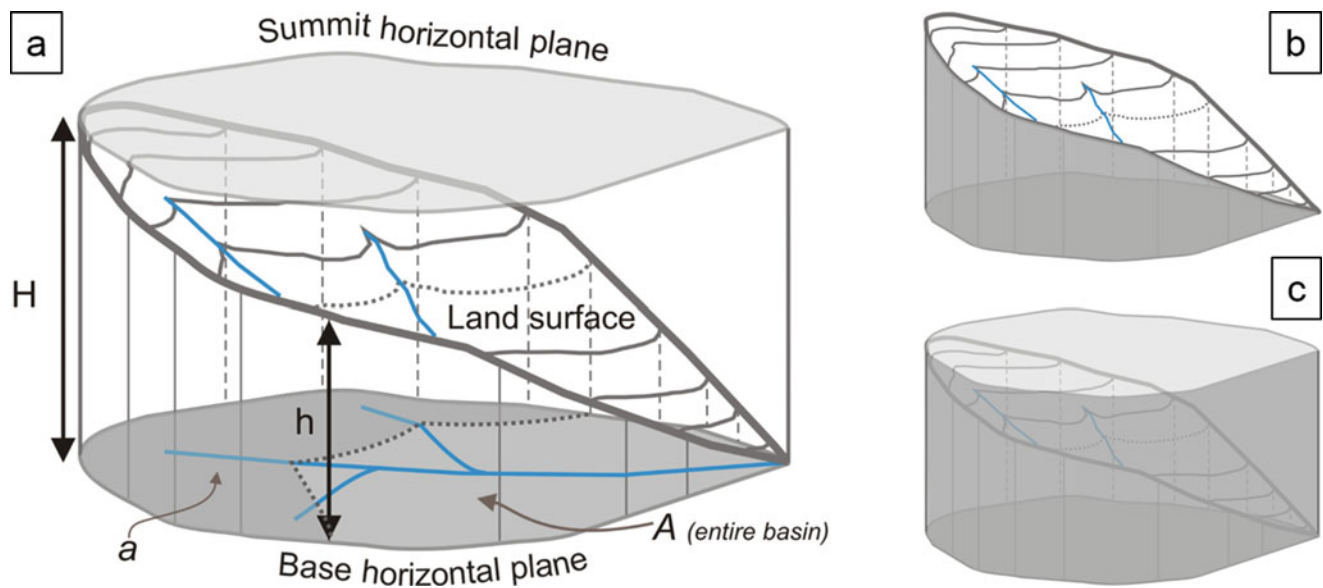
From the hypsometric curve, multiple parameters describing the statistical distribution of elevation can be derived. One of the most popular parameter is the hypsometric integral (HI), which corresponds to the area below the nondimensional hypsometric curve, a numerical index introduced by Strahler (1952) to summarize the hypsometry of fluvial basins. From the geometrical point of view, the hypsometric integral is the ratio between the volume contained between the land surface and the base plane (Hypsometry, Fig. 1b) and the volume obtained extruding the base plane for the basin relief (Hypsometry, Fig. 1c).

A simple equation (Eq. 1) proposed by Pike and Wilson (1971) permits the computation of the hypsometric integral (HI) using the mean (H_m), minimum (H_{min}), and maximum (H_{max}) elevation in the studied area:

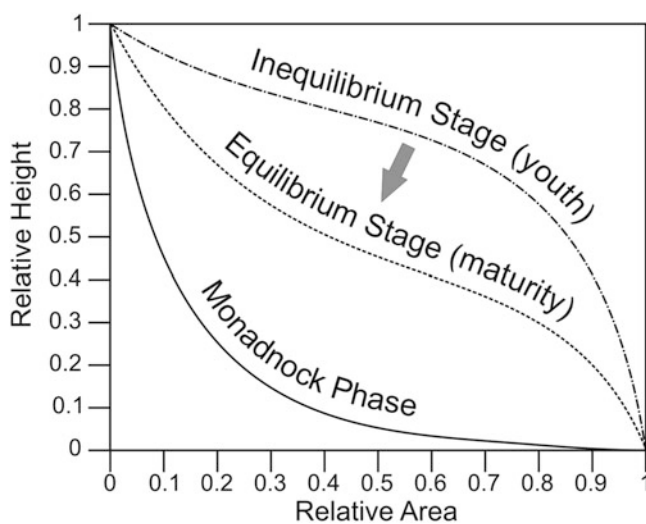
$$HI = \frac{H_m - H_{min}}{H_{max} - H_{min}} \quad (1)$$

The hypsometric curves and integrals are robust against different estimation methods (Singh et al. 2008). In case of derivation from a DTM, HI shows low sensitivity to variations in DTM resolution (Hurtrez et al. 1999). These features enable the comparison of hypsometric parameters computed with different methods and using topographic data of different resolutions. However, a relevant limitation of the hypsometric integral is that different hypsometric curves, which imply different distribution of areas with elevation, could have similar hypsometric integrals.

In earth sciences and ecological studies, multiple indices derived from hypsometry are considered, including statistical moments of the distribution of elevation and the possibility to consider the probability density distribution. In this regard, the modeling of the experimental hypsometric curve with a theoretical function (Strahler 1952) facilitates the derivation of descriptive indices. For example, third degree polynomial



Hypsometry, Fig. 1 (a) sketch with the main geometric elements of hypsometry for a drainage basin (modified from Strahler 1952); (b and c) representative volumes in hypsometric integral calculation



Hypsometry, Fig. 2 Nondimensional hypsometric curves for three stages of drainage basin evolution (modified from Strahler 1952)

curves (Eq. 2) have been widely applied to fit the hypsometric curves in a quite wide range of morphological settings.

$$h/H = c_1 + c_2 \cdot a/A + c_3 \cdot (a/A)^2 + c_4 \cdot (a/A)^3 \quad (2)$$

where c_1 , c_2 , c_3 , and c_4 are coefficients determined by curve fitting. However, in presence of complex hypsometric curves, related, for example, to lithological and geomechanical heterogeneities in the basin, as in the monadnock phase, these polynomials are not sufficiently flexible (Vanderwaal and Ssegane 2013).

Relevance in Geomorphology

In his pioneering studies on hypsometry, Strahler (1952, 1957), considering also the work of Langbein et al. (1947), related the shape of the hypsometric curves and the values of hypsometric integral to the different stages of evolution of drainage basins (Hypsometry, Fig. 2). A young stage features high hypsometric integrals and is characterized by inequilibrium conditions that lead to a mature equilibrium stage. The monadnock phase with very low hypsometric integral, when it does occur, can be followed by a return to equilibrium conditions after the removal of isolated elevated areas of resistant rock.

Several studies (e.g., Willgoose and Hancock 1998; Hurtrez et al. 1999; Korup et al. 2005) have provided evidence of the scale dependency of hypsometric parameters of drainage basins. The transition from diffusive to fluvial dominance (Willgoose and Hancock 1998), which affects the relative importance of river processes versus hillslopes processes (Hurtrez et al. 1999), determines the dependence of hypsometry on basin scale. The shape of the hypsometric curves (e.g., Lifton and Chase 1992; Pérez-Peña et al. 2009; Marchi et al. 2015) is directly related to multiple geomorphic processes and factors, such as geostructural setting, lithological heterogeneity, climatic factors, and biotic processes which control the morphological evolution of a basin.

Given the relationships between hypsometric characteristics and geomorphic processes and factors, it is not surprising the ample literature related to the exploitation of hypsometric-based indices in multiple research contexts, not limited to the analysis of drainage basins (e.g., morphology of glaciers, planetary morphology, etc.). Moreover, hypsometry-based indices are also considered in geomorphometry as input

parameters for geocomputational approaches based on supervised (e.g., landslide susceptibility models) and/or unsupervised learning (e.g., landscape classification based on geomorphometric features).

Application in Hydrology

In hydrology, hypsometry helps computing the spatial average precipitation in regions where the orographic effects are important. In the traditional approach, the combined use of a relationship linking precipitation and elevation and the hypsometric curve permits assessing the precipitation at different elevation belts and the mean average precipitation in the studied area (Dingman 2002). Like in other sectors of quantitative terrain analysis, also in the assessment of mean areal precipitation, the use of gridded elevation data permits taking into account the effect of elevation without using graphical tools such as the hypsometric curve.

Conclusions

Hypsometry describes the statistical distribution of elevations in a defined spatial domain of analysis, often represented by a drainage basin, through parameters that provide a spatial-statistical representation of a landscape. Hypsometry can thus be considered a precursor of modern geomorphometry.

Hypsometry has found many applications in geomorphology: it was historically considered as an indicator of basins and landforms evolution and was then related to the geomorphic processes and their controlling factors.

Care should be paid when using hypsometric curves and derived parameters, such as the hypsometric integral, for comparative purposes. The dependency of hypsometry on multiple geomorphic processes and factors, and, in some conditions, on spatial scale should be always taken into account.

Cross-References

- Digital Elevation Model
- Digital Geological Mapping
- Earth Surface Processes

- Horton, Robert Elmer
- LiDAR
- Machine Learning
- Morphometry
- Quantitative Geomorphology
- Scaling and Scale Invariance

Bibliography

- Dingman SL (2002) Physical hydrology. Prentice Hall, Upper Saddle River
- Hurtrez JE, Sol C, Lucazeau F (1999) Effect of drainage area on hypsometry from an analysis of small-scale drainage basins in the Siwalik Hills (central Nepal). *Earth Surf Process Landf* 24: 799–808
- Korup O, Schmidt J, McSaveney MJ (2005) Regional relief characteristics and denudation pattern of the western Southern Alps, New Zealand. *Geomorphology* 71:402–423. <https://doi.org/10.1016/j.geomorph.2005.04.013>
- Langbein WB et al (1947) Topographic characteristics of drainage basins. United States Geological Survey, Water-Supply Paper 968-C, pp 125–158
- Lifton NA, Chase CG (1992) Tectonic, climatic and lithologic influences on landscape fractal dimension and hypsometry: implications for landscape evolution in the San Gabriel Mountains, California. *Geomorphology* 5:77–114
- Marchi L, Cavalli M, Trevisani S (2015) Hypsometric analysis of headwater rock basins in the Dolomites (Eastern Alps) using high-resolution topography. *Geogr Ann Ser B* 97(2):317–335. <https://doi.org/10.1111/geoa.12067>
- Pérez-Peña JV, Azañón JM, Booth-Rea G, Azor A, Delgado J (2009) Differentiating geology and tectonics using a spatial autocorrelation technique for the hypsometric integral. *J Geophys Res* 114:F02018. <https://doi.org/10.1029/2008JF001092>
- Pike RJ, Wilson SE (1971) Elevation-relief ratio, hypsometric integral, and geomorphic area – altitude analysis. *Geol Soc Am Bull* 82: 1079–1084
- Singh O, Sarangi A, Sharma MC (2008) Hypsometric integral estimation methods and its relevance on erosion status of North-western Lesser Himalayan watersheds. *Water Resour Manag* 22:1545–1560. <https://doi.org/10.1007/s11269-008-9242-z>
- Strahler AN (1952) Hypsometric (area-altitude) analysis of erosional topography. *Bull Geol Soc Am* 68:1117–1142
- Strahler AN (1957) Quantitative analysis of watershed geomorphology. *Trans Am Geophys Union* 38(6):913–920
- Vanderwaal JA, Ssegane H (2013) Do polynomials adequately describe the hypsometry of monadnock phase watersheds? *J Am Water Resour Assoc* 49(6):1485–1495
- Willgoose G, Hancock G (1998) Revisiting the hypsometric curve as an indicator of form and process in transport-limited catchment. *Earth Surf Process Landf* 23:611–623

Imputation

Javier Palarea-Albaladejo
Biomathematics and Statistics Scotland, Edinburgh, UK

Definition

In data analysis and statistical modelling, imputation refers to a procedure by which missing or defective entries in a data set are replaced by plausible estimates to generate a complete data set that is usable for further processing.

Missing Data and Nondetects in Geoscientific Studies

Modern datasets generated in the geosciences ordinarily contain collections of variables referring to varied characteristics of the medium under study. The observed data are a realization of the underlying physical and chemical processes being investigated and multivariate statistical methods are well suited for their analysis and interpretation. Thus, along with the specialized methods of geostatistics, which are mostly focused on modelling spatiotemporal phenomena, multivariate techniques for data reduction, classification, and prediction are part of the basic quantitative toolbox of the geoscientist. Regardless of how cautiously a researcher designs and executes a study, one of the most prevalent problems in empirical work relates to the presence of missing values in multivariate data sets. Missing data in general can be due to many reasons, e.g., circumstances affecting data collection, failures of the instruments, or data handling issues. These are commonly indicated in a data matrix by empty cells or using a special code or character. A prevalent type of missing data in the geosciences refers to left-censored data. For example, in spatial exploratory studies, there are frequently geographic locations where indicators of the presence

of an element of interest are found, but its precise quantities cannot be determined. This commonly corresponds to trace element concentrations falling below the limit of detection of the measuring instrument and is a form of left-censored data: the actual value is unknown, but we do know that it must be somewhere in the lower tail of the data distribution below the detection limit. Thus, they are often called less-thans or non-detects. A detection limit (DL) is considered a threshold below which measured values cannot be distinguished from a blank signal or background noise, at a specified level of confidence.

Once it has been deemed impossible to recover the actual values, the practical problem is how to deal with missing data, nondetects, or often both simultaneously, to be able to carry on with data analysis and modelling. Statistical analysis based on only the available data, or after a naïve substitution of the unobserved values, can introduce important bias and achieve reduced statistical power, particularly as the number of cases increases. These solutions are unfortunately common default options in statistical software packages.

Missingness Mechanisms

Frameworks for missingness mechanisms have been developed that can assist in determining the approach to deal with missing data. Understanding the occurrence of missing elements in a data matrix $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{mis})$ as a probabilistic phenomenon, with $\mathbf{X}_{obs}$ and $\mathbf{X}_{mis}$ referring to the observed and missing partitions, respectively; we can define a random missingness matrix $\mathbf{M}$ with elements 0 or 1 corresponding to observed or missing values in $\mathbf{X}$ respectively. Depending on the conditional distribution of $\mathbf{M}$ given $\mathbf{X}$, and a set of unknown parameters ξ , the missingness mechanisms are generally classified as:

- Missing completely at random (MCAR), when the probability of missingness is the same for all elements of the data

matrix, i.e., $P[\mathbf{M} | \mathbf{X}_{obs}, \mathbf{X}_{mis}; \xi] = P[\mathbf{M}; \xi]$. This is the simplest case, implying that missing values are a random sample of the data matrix. Basic approaches like data analysis using only the observed data can be sensible in this setting, however this is generally unrealistic.

- Missing at random (MAR), when the probability a value is missing depends only on the available information, i.e., $P[\mathbf{M} | \mathbf{X}_{obs}, \mathbf{X}_{mis}; \xi] = P[\mathbf{M} | \mathbf{X}_{obs}; \xi]$. This is the situation most commonly assumed in practice. The missing part can be ignored as long as the variables that affect the probability of a missing value are accounted for. Note, however, that if $\mathbf{M}$ depended on other variables that are not explicitly included in $\mathbf{X}_{obs}$, then MAR would no longer be satisfied.
- Missing not at random (MNAR), when the probability a value is missing depends on the value itself and hence both $\mathbf{X}_{obs}$ and $\mathbf{X}_{mis}$ are involved. This is the most complicated situation and implies a systematic difference between the observed and missing data. The case of nondetects as discussed above falls into this category.

Given its practical relevance, the treatment of missing data is an active statistical research area with ramifications throughout varied applied disciplines (Little and Rubin 2002; Molenberghs et al. 2014). Approaches like complete-case analysis (list-wise deletion) or available-case analysis (pair-wise deletion) simplify the problem by discarding data. The first leaves out all samples including any missing values. The second intends to maximize the use of the information available in each case. Thus, e.g., the calculation of means can be based on different samples depending on the missingness pattern of each variable and, analogously, correlations use data simultaneously available for each particular pair of variables. Note that the latter can lead to the technical issue of non-definite positiveness of correlation matrices, preventing the use of most multivariate techniques. As stressed above, these strategies must be used with great caution, particularly when missing data are not MCAR and there are meaningful associations between variables. Nonetheless, it is not feasible to define universal rules of thumb about when to use these methods or not, or actually any others. A fundamental consideration is of course the rate of missing data, but there are also important factors that are context specific. Hence, some testing and exploration blended with expert knowledge would better assist in making informed decisions in practice.

More sophisticated methods that commonly assume a MAR mechanism include different forms of regression modelling, likelihood-based estimation methods, Bayesian data augmentation, and single and multiple imputation procedures. For the MNAR case of nondetects, a popular approach among practitioners in the earth and environmental sciences is simply substituting them by a fraction of the detection limit, typically $DL/2$, or simply using DL itself.

Advantages of this simple substitution are simplicity and the fact that no model assumptions are made. However, different studies have shown the distortion and undesirable consequences that such practice can have and advocate for methods that exploit the statistical properties of the data. Specialized maximum likelihood, robust, and nonparametric methods have been introduced for the unbiased estimation of common univariate statistics (means, standard deviations, quantiles, and so on) from left-censored data (Helsel 2012). However, these are not well suited for complex missingness patterns typical of multivariate data sets, possibly in high dimensions, which often involve varied interrelations between variables, for example spatial or temporal dependencies. Thus, prediction to recover the regional structure of an analyte or mineral resource and estimation of the overall spatial variability are common tasks in spatial modelling. These may benefit from exploiting the spatial correlation to deal with any censored data (Schelin and Sjöstedt-de Luna 2014).

Multivariate Data Imputation

Data imputation provides an intuitive and powerful tool for the analysis of multivariate data with complex missing data patterns. The basic idea is straightforward: replacing missing values across variables by plausible values, ideally by statistically exploiting the observed data and the associations between variables, using different approaches depending on the nature of the data (van Buuren 2018; Schafer 1997). The result is a completed version of the data set that allows to carry on with the subsequent data analysis and modelling. Note that imputation can be also useful to deal with other forms of irregular data such as outliers, particularly when these are regarded as erroneous or inaccurate measurements. There exists a wide range of imputation methods of varied levels of complexity. The actual impact of any procedure on the results will depend on the characteristics of the data set and the focus of the subsequent analysis. We refer the reader to the general references provided above for an overview. One of the simplest methods is mean imputation, i.e., replacing missing values in a variable by the mean of the observed values. Note that this will tend to underestimate standard deviations and distort the correlations between variables. Model-based imputation will be generally preferred for moderate to high fractions of missing data in regular multivariate data sets, even when the imputation model only weakly agrees with the model assumed for the complete data (Schafer 1997). Unlike single imputation, multiple imputation is meant to create several versions of the imputed data set and combine estimates from them. This principally aims to incorporate the extra uncertainty derived from the own imputation process to the estimation of standard errors and, hence, to, e.g., the determination of rejection regions in statistical hypothesis

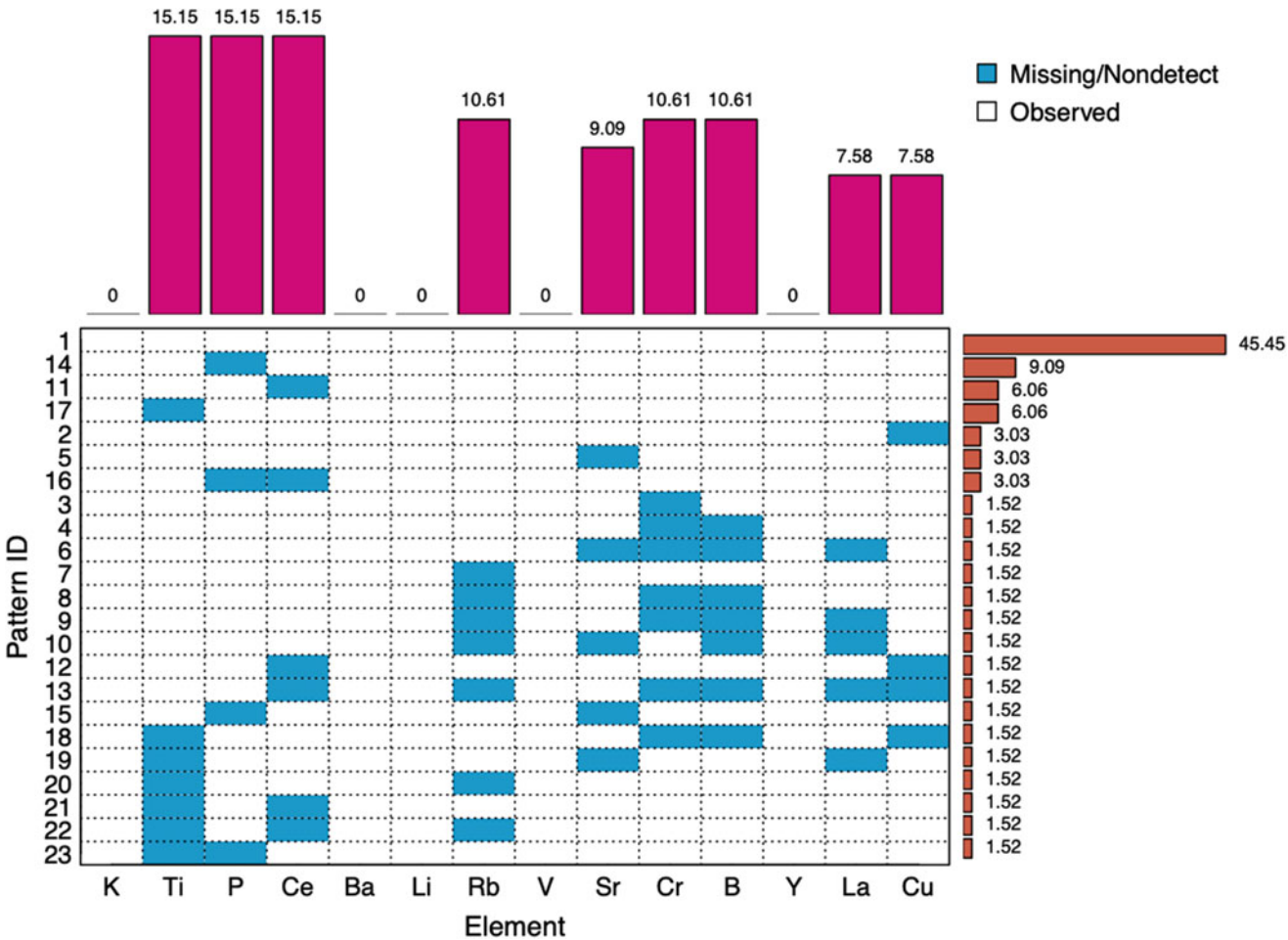
testing. Methods within the sphere of machine learning such as random forests or k-nearest neighbors (KNN) algorithms have been also formulated to serve as imputation devices. Other imputation methods particularly relevant in computationally intensive high-dimensional contexts, e.g., in high-throughput experiments, rely on low-rank matrix completion. Specialized computer routines can nowadays be found in the most popular statistical computing platforms. Within the open-source arena, we refer the reader to The Comprehensive R Archive Network (CRAN) Task View on missing data (<https://cran.r-project.org/web/views/MissingData.html>), which keeps a detailed account of related packages available on the R statistical computing environment.

The Case of Compositional Data

Compositional data (CoDa) are observations referring to fractions or parts of a whole, often represented in units such as proportions, ppm, weight %, mg/kg, and similar.

Compositional data are ubiquitous in different areas of the geosciences, including geochemistry, soil science, mineralogy, sedimentology, among others (Buccianti et al. 2006; Tolosana-Delgado et al. 2019). Compositions are defined on a sample space equipped with a special geometry that departs from the ordinary real space assumed by most statistical methods. Hence, statistical analysis and modelling of CoDa present some challenges and particularities that are not well addressed by conventional methods, which can lead to misleading results and interpretations. The mainstream approach to CoDa analysis involves focusing on (log-)ratios between the parts of the composition. These log-ratios carry the relevant relative information and crucially do not depend on an arbitrary scale of measurement.

Note that zero values do not fit into the above conceptualization, since they are incompatible with the ratio and logarithm operations. However, they commonly occur in experimental studies involving compositions. Hence, this has been a particularly important issue in practical CoDa analysis. Zero parts in our context often concur with the



Imputation, Fig. 1 Patterns of missingness generated by simulation in illustrative geochemical data set and relative frequencies of cases

concept of nondetects as described above, i.e., small values that have been rounded off or have fallen below a limit of detection and then registered as zeros. The problem is more complex here than in standard multivariate analysis due to the fact that a single unobserved part renders the whole composition undefined. A basic property of any method to deal with zeros, nondetects, or missing data in general within this context, is that they should preserve the relative variation structure of the data. A number of specialized imputation methods have been developed in CoDa analysis following different methodological flavors. The R package *zCompositions* (Palarea-Albaladejo and Martín-Fernández 2015) provides an integrated framework for compositional data imputation (applicable to zeros, nondetects, missing data, or combinations of them) following the principles of the log-ratio approach. This includes a consistent treatment of closed and non-closed compositions, unique or varying detection limits, parametric and nonparametric imputation, single and multiple imputation, maximum likelihood and robust estimation, and other related tools.

For illustration, we used 66 complete samples of a 14-part geochemical composition (in $\mu\text{g/g}$) from La Paloma stream (Venezuela) including the elements K, Ti, P, Ce, Ba, Li, Rb, V, Sr, Cr, B, Y, La, and Cu (the entire data set is available in *zCompositions*; the elements have been sorted here by decreasing mean relative abundance). The first five were considered major elements, whereas the remaining nine were assumed to be minor elements. We simulated around 15% of missing values in each of a random selection of 60% of the major elements (MCAR; 3.25% missing overall) and around 10% of nondetects in each of a random selection of 70% of the minor elements (using their 10% quantiles as DLs; 4% nondetects overall). This generated an artificial data set including 30 complete samples (45.45%); along with 18 (27.27%), 11 (16.67%), and 7 (10.61%) samples having respectively either missing, nondetect, or both types simultaneously. Figure 1 summarizes the patterns of missingness in the data using the *zPatterns* function in *zCompositions* (combining missing and nondetects and sorting patterns by decreasing frequency). The bars on the sides indicate the percentage cases, either by element (columns) or samples per pattern (rows).

Imputation was conducted using the *lrEMplus* function in *zCompositions* (missing and nondetect values were coded as NA and 0, respectively, for this), with the 10% quantiles used as vector of DLs for the nondetects. This function implements an iterative model-based regression-type imputation method (more details and references can be found in the documentation accompanying the package). Basic compositional summary statistics of central tendency and variability were computed from the originally complete data and the imputed data to briefly assess the results. In particular, Table 1 shows

Imputation, Table 1 Compositional summary statistics from complete and imputed data sets: closed geometric mean (in percent units), total variance, and individual contributions to total variance (raw and percent values)

	Complete data		Imputed data	
	CGM (in %)	Contrib. var. (%)	CGM (in %)	Contrib. var. (%)
K	79.883	0.181 (9.44)	79.271	0.183 (9.71)
Ti	16.547	0.068 (3.57)	17.144	0.078 (4.15)
P	1.005	0.277 (14.45)	1.034	0.275 (14.57)
Ce	0.806	0.097 (5.07)	0.804	0.099 (5.24)
Ba	0.373	0.259 (13.52)	0.370	0.261 (13.86)
Li	0.337	0.177 (9.24)	0.335	0.174 (9.24)
Rb	0.326	0.140 (7.33)	0.320	0.147 (7.81)
V	0.233	0.040 (2.08)	0.231	0.039 (2.09)
Sr	0.154	0.135 (7.04)	0.154	0.127 (6.75)
Cr	0.106	0.082 (4.28)	0.107	0.074 (3.93)
B	0.083	0.052 (2.70)	0.083	0.054 (2.84)
Y	0.071	0.114 (5.95)	0.070	0.112 (5.94)
La	0.055	0.136 (7.10)	0.056	0.092 (4.87)
Cu	0.020	0.157 (8.23)	0.020	0.170 (9.00)
Total variance		1.915 (100)		1.885 (100)

the closed geometric mean (CGM; a.k.a. compositional center); that is, the vector of geometric means of the individual elements rescaled (closed) to be expressed in percentage units in this case. Moreover, the total variance of the data set and the relative contributions of the elements to it are reported (computed from ordinary variances of the centered log-ratio transformation of the composition; raw and percent values shown). It can be observed that the imputation method introduced negligible distortion regarding the center of the data set. The total variation was slightly underestimated, with main differences in contribution being related with the minor elements La and Cu and the major element Ti.

Figure 2 shows the pair-wise log-ratio variances (variances of the log-ratios between each pair of elements; arranged by rows and columns) from the complete and imputed data (displayed in upper and lower triangle respectively). Lower values correspond to higher associations between elements in terms of proportionality. A color scale was used to facilitate

Imputation, Fig. 2 Pair-wise log-ratio variances from complete and imputed data sets (upper and lower triangle respectively)

	K	Ti	P	Ce	Ba	Li	Rb	V	Sr	Cr	B	Y	La	Cu
K		0.25	0.61	0.36	0.47	0.28	0.15	0.23	0.40	0.34	0.25	0.32	0.38	0.40
Ti	0.27		0.43	0.21	0.29	0.32	0.24	0.13	0.25	0.13	0.07	0.14	0.23	0.18
P	0.64	0.29		0.36	0.67	0.49	0.56	0.31	0.34	0.36	0.41	0.32	0.46	0.45
Ce	0.33	0.19	0.44		0.37	0.35	0.30	0.14	0.20	0.22	0.19	0.19	0.11	0.29
Ba	0.47	0.34	0.70	0.36		0.55	0.54	0.37	0.38	0.32	0.31	0.36	0.45	0.45
Li	0.28	0.38	0.47	0.35	0.55		0.14	0.17	0.37	0.30	0.26	0.37	0.43	0.36
Rb	0.16	0.30	0.57	0.29	0.54	0.14		0.14	0.37	0.28	0.19	0.34	0.27	0.36
V	0.23	0.15	0.31	0.14	0.37	0.17	0.14		0.22	0.10	0.09	0.21	0.15	0.21
Sr	0.39	0.24	0.35	0.18	0.37	0.35	0.39	0.20		0.24	0.21	0.25	0.27	0.31
Cr	0.35	0.17	0.34	0.21	0.31	0.28	0.26	0.09	0.21		0.07	0.26	0.24	0.21
B	0.25	0.13	0.42	0.19	0.31	0.26	0.19	0.09	0.20	0.06		0.20	0.20	0.19
Y	0.32	0.10	0.33	0.19	0.36	0.37	0.35	0.21	0.25	0.26	0.19		0.24	0.31
La	0.33	0.19	0.42	0.07	0.38	0.35	0.22	0.11	0.22	0.19	0.15	0.20		0.39
Cu	0.42	0.23	0.46	0.32	0.48	0.38	0.39	0.23	0.31	0.20	0.20	0.32	0.33	

interpretation, ranging from redder for stronger associations to greener for weaker associations.

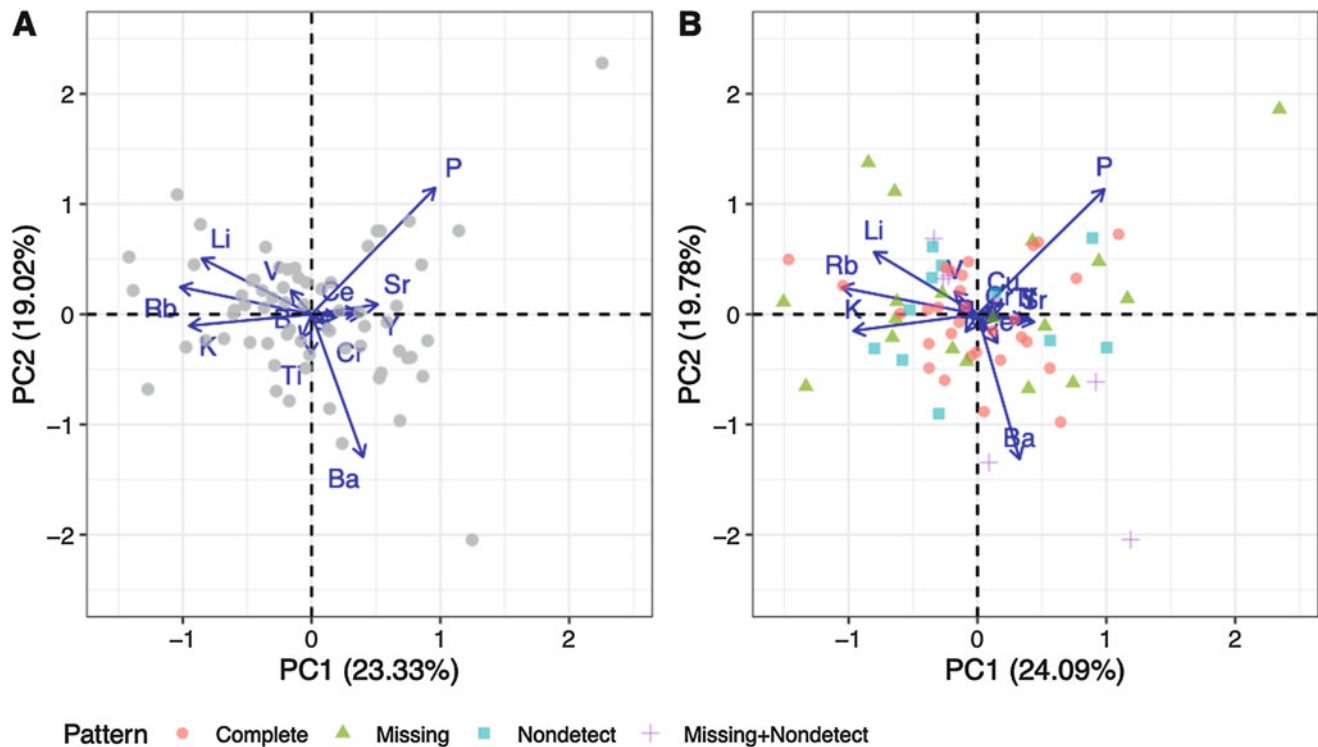
In line with the results above, the patterns of associations were very much comparable between complete and imputed data. On the one hand, the strongest associations were generally found among the minor elements, with B being often involved. Major elements Ti and Ce appeared as the most associated with minor elements. On the other hand, the weakest associations were generally found among the major elements, with the highest log-ratio variance being that between P and Ba.

Finally, complete and imputed data sets were visualized in two dimensions using a compositional principal component analysis (PCA) biplot (Fig. 3). The first two principal components (PC1 and PC2) explained similarly moderate 42.35% and 43.87% of the total variability of the respective data sets. Colors and symbols were used to distinguish samples that were complete from those including either nondetects, missing values, or both simultaneously. The overall structure of the data was well preserved after imputation. The length of the links between the arrowheads corresponding to the different elements is related with the log-ratio variances between them. Thus, in agreement with the results showed in Fig. 2, the most independent elements appeared to be P and Ba, particularly with respect to K, Rb, and Li. Obviously, a main factor in

determining the distortion introduced by any imputation method in practice will be the gravity of the missingness problem. We refer the reader to the specialized literature for more detailed discussions about performance in different scenarios.

Summary

Missing data and nondetects are ubiquitous in experimental studies. Ignoring them or treating them carelessly can seriously distort the results and scientific conclusions, particularly where there is a systematic difference between observed and unobserved values. In this sense, it is convenient to investigate why they occurred and make an informed decision about how they could be best handled. Important factors to consider include the severity of the problem, the size of the data set, and the distributional characteristics and nature of the variables involved. Data imputation is a popular approach, and there is a wide range of methods available under different flavors and assumptions. This includes specialized methods for data of compositional nature; that is, data representing relative amounts as frequently found across the geological and environmental sciences.



Imputation, Fig. 3 Compositional PCA biplot of original complete data set (a) and imputed data set (b) after simulating missing values and nondetects in illustrative geochemical composition data

Cross-References

- [Compositional Data](#)
- [Geostatistics](#)
- [Multivariate Analysis](#)
- [Statistical Computing](#)

Tolosana-Delgado R, Mueller U, van den Boogaart KG (2019) Geostatistics for compositional data: an overview. *Math Geosci* 51:485–526

van Buuren S (2018) Flexible imputation of missing data. Chapman & Hall/CRC Press, Boca Raton, Florida, USA

Bibliography

- Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V (eds) (2006) Compositional data analysis in the geosciences: from theory to practice. Geological Society of London, London, UK
- Chilès JP, Delfiner P (2012) Geostatistics: modeling spatial uncertainty. Wiley, Hoboken, New Jersey, USA
- Helsel DR (2012) Statistics for censored environmental data using Minitab® and R. Wiley, Hoboken, New Jersey, USA
- Little RJA, Rubin DB (2002) Statistical analysis with missing data. Wiley, Hoboken, New Jersey, USA
- Molenberghs G, Fitzmaurice G, Kenward M, Tsiatis B, Verbeke G (2014) Handbook of missing data methodology. CRC Press, Boca Raton, Florida, USA
- Palarea-Albaladejo J, Martín-Fernández JA (2015) zCompositions – R package for multivariate imputation of left-censored data under a compositional approach. *Chemometr Intell Lab Syst* 143:85–96
- Schafer JL (1997) Analysis of incomplete multivariate data. Chapman & Hall, Boca Raton, Florida, USA
- Schelin L, Sjöstedt-de Luna S (2014) Spatial prediction in the presence of left-censoring. *Comput Stat Data Anal* 74:125–141

Independent Component Analysis

Jaya Sreevalsan-Nair

Graphics-Visualization-Computing Lab, International Institute of Information Technology Bangalore, Electronic City, Bangalore, Karnataka, India

Definition

Independent component analysis (ICA) is a method of decomposing a mixture of signals, i.e., a measured signal, to its individual sources, and is widely understood using its “classical cocktail party” analog (Stone 2002). In a situation at a cocktail party where many people are talking and the speech is captured using a single microphone, ICA is a solution for extracting the different speech signals from the microphone output, given that the voices of different people are unrelated. This process of extraction of source signals is

called “blind source separation” (BSS). An example in geosciences is in using ICA for finding the different rock units in a stream sediment geochemical dataset for mineral exploration (Yang and Cheng 2015).

ICA is formulated as $\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{n}$, where $\mathbf{y}$ is the mixed signal, $\mathbf{x}$ is the matrix containing vectors of component signals, $\mathbf{M}$ is the mixing matrix, and $\mathbf{n}$ is the additive noise. If the noise is Gaussian, when using higher-order statistics in the equation, the noise term vanishes. This linear equation, upon removing the noise term, complies with the mixed signal being a linear combination of statistically independent source signals. However, in practice, the ordinary noise-free ICA methods are modified to remove the bias owing to the noise, for implementing the noisy ICA method (Hyvärinen 1999). The $\mathbf{x}$ maximizes a “contrast” function, as introduced by Gassiat in 1988 (Comon 1994), and for which “entropy” value is used. Both $\mathbf{M}$ and $\mathbf{x}$ are unknown, here, and the solution for $\mathbf{x}$ is arrived at when using the Moore–Penrose pseudoinverse of the mixing matrix, i.e., $\mathbf{x} = \mathbf{M}^\dagger \mathbf{y}$, and when $\mathbf{x}$ has maximum entropy.

Overview

The term independent component analysis was introduced by Herault and Jutten circa 1986 (Comon 1994), owing to its similarity with principal component analysis (PCA). A brief history of ICA and its evolution has been discussed by Comon (1994).

Alternative methods for BSS are PCA, factor analysis (FA), and linear dynamical systems (LDS). In ICA, the source signals, extracted as components, are assumed to be statistically independent and non-Gaussian. Unlike ICA, the popularly used alternative methods, PCA and FA, work on the basic assumption that the source signals are uncorrelated and Gaussian. Statistical independence between two signals implies that the values of one signal have no influence or information on those of the other signal. Being uncorrelated does not imply statistical independence as the value of one signal can provide information on the other signal. Thus, uncorrelated variables can be considered to be *partly independent* (Hyvärinen and Oja 2000). The source signals are extracted as eigenvectors and factors, respectively, in PCA and FA. In ICA and PCA, the source signals are referred to as “components.” In PCA, the uncorrelated signals are captured using eigenvectors of the correlation matrix, which are orthogonal components, by design. Thus, PCA removes correlations, which are second-order statistical dependencies, to extract components. That also implies that PCA does not infer higher-order statistical dependencies, e.g., skew and kurtosis. ICA, on the other hand, captures higher-order dependencies. There are known to be similarities between blind deconvolution and ICA (Hyvärinen 1999). When the blind

deconvolution is applied to a signal, which is independently and identically distributed (i.i.d.) over time, it relates closely with ICA.

A widely used data preprocessing step for ICA is the removal of correlations in the signal, referred to as “whitening” the signal. PCA has been used for whitening signals. The extraction of higher-order dependencies is reflected in the non-orthogonality of the components in ICA. The property of statistically independent signals having maximum entropy is exploited in finding independent components. ICA has been widely developed and used in cognitive science for both biomedical data analysis and computational modeling.

Solving PCA using eigenvalue decomposition (EVD) of covariance or correlation matrix is a linear process, whereas removing higher-order dependencies in ICA leads to a multilinear EVD (De Lathauwer et al. 2000). Instead of EVD of correlation matrices, singular value decomposition (SVD) of the raw data can be implemented. There are four widely used algebraic methods using multilinear EVD for solving ICA, namely,

- By means of higher-order eigenvalue decomposition (HOEVD).
- By means of maximal diagonality (MD).
- By means of joint approximate diagonalization of eigenmatrices (JADE).
- By means of simultaneous third-order tensor diagonalization (STOTD).

The HOEVD is the least accurate method, and among the remaining three, MD is the most computationally intensive. Thus, STOTD and JADE algorithms are the preferred methods, of which STOTD is more efficient owing to the readily available tensors that have to be diagonalized, whereas the dominant eigenspaces have to be additionally computed in the case of JADE.

Overall, the ICA algorithm has two aspects, namely, (1) the objective function that maximizes entropy, etc., and (2) the optimization algorithm that influences the algorithmic properties, such as convergence speed, memory complexity, and stability (Hyvärinen 1999). In practice, using stochastic gradient method leads to slower convergence. Hence, a fixed-point algorithm, called FastICA, was proposed by Hyvärinen in 1997 to solve the convergence issue, for which a brief summary is given in Hyvärinen 1999.

Extending the BSS/ICA model from one-dimensional to multidimensional signals, the standard algorithms for ICA can be used for multidimensional ICA (MICA) (Cardoso 1998). This can be done by using a geometric parameterization to avoid using the matrix algebraic technique.

The matrix formulation of noise-free ICA leads to two known ambiguities (Hyvärinen and Oja 2000). Firstly, the variances of the independent components cannot be

determined. Secondly, the order of independent components cannot be determined. This is because all independent components are equally important, unlike the reducing variances in the principal components in PCA that provides the ordering.

Applications

Common applications of ICA include separation of artifacts in the magnetoencephalography (MEG) data of cortical neurons, determination of hidden factors in financial data, and reducing noise in natural images (Hyvärinen and Oja 2000).

Similarly, ICA has also found widespread applicability in geosciences and related areas. When used as exploratory tools for geochemical data analysis, PCA has been found to be useful in finding significant geo-objects, whereas ICA is capable of separating geological populations in the dataset (Yang and Cheng 2015). Hyperspectral images have been classified into thematic land-cover classes using ICA and extended morphological attribute profiles (EAPs) of the images (Dalla Mura et al. 2010). The morphological attribute profiles are first computed for capturing the multiscale variation in geometry. Then, they are applied to the independent components of the image, computed using ICA. The support vector machine (SVM) classifier is then used with the EAPs for classification. ICA has been found to give more accurate classification outcomes than PCA for this application. Similarly, ICA has been useful for analyzing the variability of tropical sea surface temperature (SST) owing to its modeling of statistically independent components (Aires et al. 2000). The whitened SST has been found to be a mixture involving variabilities in several timescales in different geographical areas. Hence, ICA is best applicable in identifying the contributing components, which in turn determine the influence of SST variability in the El Niño Southern Oscillations (ENSO). This method has successfully shown the relationship between the SSTs of the northern and southern equatorial Atlantic Ocean through the “dipole,” and between the SSTs of the Pacific and Atlantic Oceans through the ENSO.

Future Scope

Some of the advances in ICA since 2000 include causal relationship analysis, testing independent components, and analysis of three-way datasets (Hyvärinen 2013). There have also been developments in solving for ICA using time-frequency decompositions, modeling distributions of the components, and non-negative models.

In summary, ICA is a powerful data exploratory tool for separating statistically independent and non-Gaussian components from a mixed signal.

Cross-References

- [Eigenvalues and Eigenvectors](#)
- [Principal Component Analysis](#)
- [Q-Mode Factor Analysis](#)
- [R-Mode Factor Analysis](#)

Bibliography

- Aires F, Chédin A, Nadal JP (2000) Independent component analysis of multivariate time series: application to the tropical SST variability. *J Geophys Res Atmos* 105(D13):17,437–17,455
- Cardoso JF (1998) Multidimensional independent component analysis. In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98* (Cat. No. 98CH36181), vol 4. IEEE, pp 1941–1944
- Comon P (1994) Independent component analysis, a new concept? *Signal Process* 36(3):287–314
- Dalla Mura M, Villa A, Benediktsson JA, Chanussot J, Bruzzone L (2010) Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis. *IEEE Geosci Remote Sens Lett* 8(3):542–546
- De Lathauwer L, De Moor B, Vandewalle J (2000) An introduction to independent component analysis. *J Chemom* 14(3):123–149
- Hyvärinen A (1999) Survey on independent component analysis. *Neural Computing Surveys* 2:94–128
- Hyvärinen A (2013) Independent component analysis: recent advances. *Philos Trans Royal Soc A Math Phys Eng Sci* 371(1984):20110,534
- Hyvärinen A, Oja E (2000) Independent component analysis: algorithms and applications. *Neural Netw* 13(4–5):411–430
- Stone JV (2002) Independent component analysis: an introduction. *Trends Cogn Sci* 6(2):59–64
- Yang J, Cheng Q (2015) A comparative study of independent component analysis with principal component analysis in geological objects identification, Part I: Simulations. *J Geochem Explor* 149:127–135

Induction, Deduction, and Abduction

Frits Agterberg

Central Canada Division, Geomathematics Section,
Geological Survey of Canada, Ottawa, ON, Canada

Definition

Induction is a method of reasoning from premises that supply some evidence without full assurance of the truth of the conclusion that is reached. It differs from deduction that starts from premises that are correct and lead to conclusions which are certainly true. Abduction is a rarely used form of logical inference advocated by the philosopher Peirce. It starts with observations from which it seeks to draw the simplest and most likely conclusion. All three words are based on the Latin verb “ducere,” meaning to lead. The prefix “de” means “from,” indicating that deduction derives truth from factual

statements; “in” means “toward,” and induction results in a generalization; while “ab” means “away,” i.e., taking away the best explanation.

Introduction

Induction and deduction are philosophical concepts that are more frequently used than abduction, which often is regarded as a kind of induction or as a form of exploratory data analysis. The philosopher Charles Sanders Pierce (see Staat 1993) made a distinction between inference and decision-making under uncertainty on the one hand, and “theorizing” on the other. He called the former induction and the latter abduction. Traditionally, both induction and theorizing are prevalent in the geological sciences where, commonly, limited observations are used to construct 3-D realities often along with their history in geologic time. This process of reconstructing geological realities is prone to uncertainties and has often resulted in fierce debates in the past.

Difficulties in the application of mathematics to solve many types geological problems stem from: (1) the nature of geological phenomena, especially paucity of exposure of bedrock and restriction of observations to a record of past events that have been only partially preserved in the rocks, and (2) the nature of traditional geological methods of research, which are largely nonmathematical (cf. Agterberg 1961, 1974). Induction is prevalent in geology. It could be argued that geology is a science of induction or abduction.

When mathematics is used to solve geologic problems, the parameters must be defined in a manner sufficiently rigorous to permit nontrivial derivations. During the design of models, it is important to keep in mind Chamberlin’s (1897) warning: “The fascinating impressiveness of rigorous mathematical analysis, with its atmosphere of precision and elegance, should not blind us to the defects of the premises that condition the whole process. There is, perhaps no beguilement more insidious and dangerous than an elaborate and elegant mathematical process built upon unfortified premises.” Problems related to mathematical applications in the geosciences will be discussed later in this entry. Special attention will be given to the application of Bayesian data analysis, which serves to change prior probabilities based on induction into posterior probabilities including true evidence.

Induction Versus Deduction in Mathematical Statistics

Mathematical statistics originated as a science during the seventeenth century with the definitions of “probability” in 1654, “expectation” in 1657, and other basic concepts (Hacking 2006) including the first application of a statistical

significance test Arbuthnot (1710). The idea of inductive logic was initially invented by philosophers under the banner “equipossibility” toward the end of the sixteenth century (Hacking 2006), but the idea remained dormant until the second half of the eighteenth century after the publication of the famous paper by Bayes (1763) that contains the general equation (known as Bayes’ rule or postulate):

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

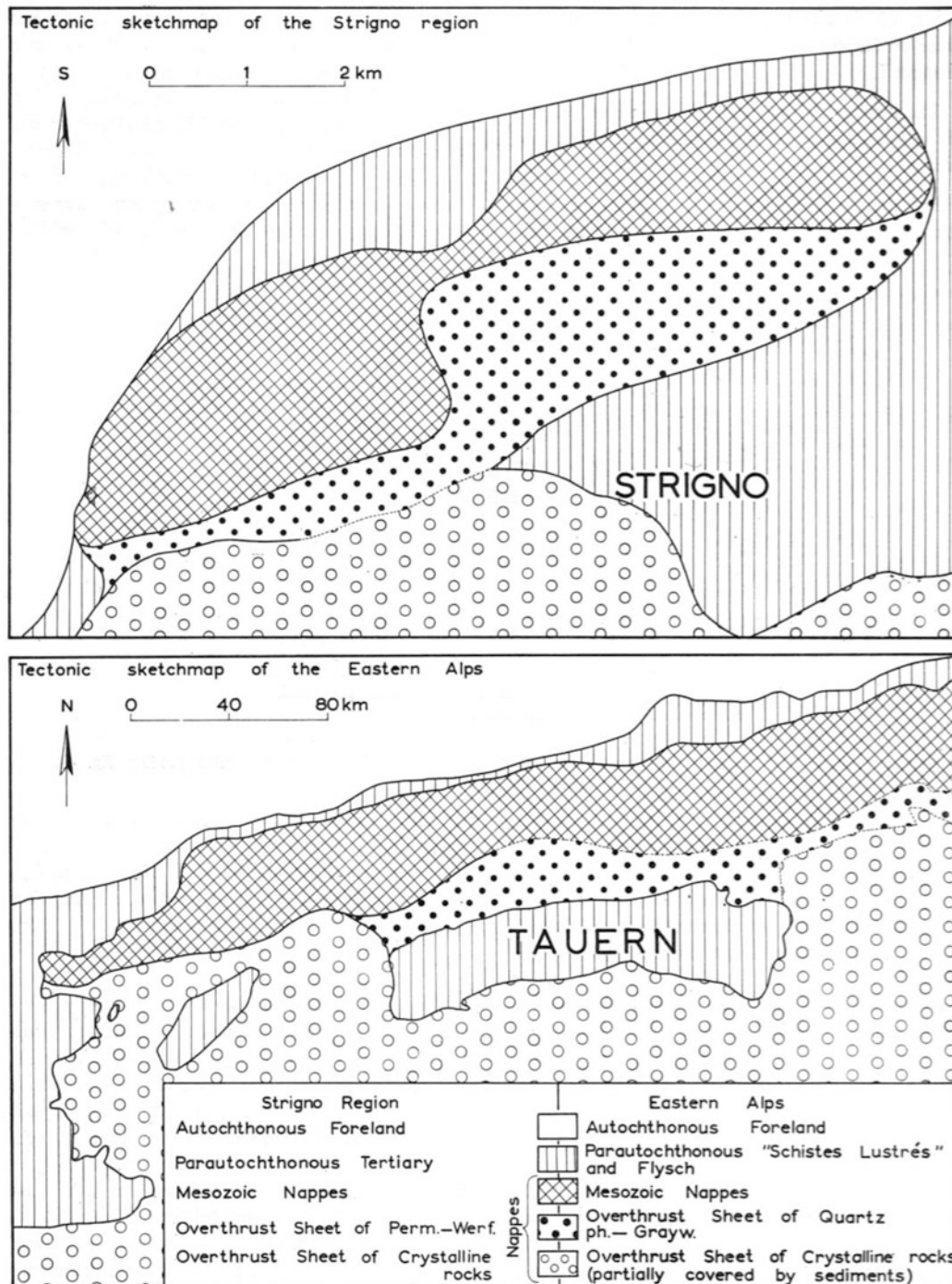
where A is a hypothesis (induction) and B is evidence. Proponents of Bayesian statistics included Jeffries (1939) who, in his book “*Theory of Probability*” insisted that inductive probabilities can only be defined in relation to evidence. Jeffries was immediately criticized by R.A. Fisher, who is generally regarded as the father of mathematical statistics, but who dismissed Jeffries’ book with the words: “He makes a logical mistake on the first page which invalidates all the 395 formulae in his book” (cf., Fisher-Box 1978). Jeffries’ “mistake” was to adopt Bayes’ postulate.

Mcgrayve (2011)’s book “*The theory that would not die*” describes in detail how Bayes’ rule emerged triumphant from two centuries of controversy. Bayesian data analysis is well described in Gelman et al. (2013). Examples of geologic examples using “equipossibility” and Bayes’ postulate will be given later in this entry.

Geological Data, Concepts, and Maps

The first geological map was published in 1815. This feat has been well documented by Winchester (2001) in his book: “*The Map that Changed the World – William Smith and the Birth of Modern Geology*.” According to the concept of stratigraphy, strata were deposited successively in the course of geologic time and they are often uniquely characterized by fossils such as ammonites that are strikingly different from age to age. It is remarkable that this fact remained virtually unknown until approximately 1800. There are several other examples of geologic concepts that became only gradually accepted more widely, although they were proposed much earlier in one form or another by individual scientists. The best-known example is plate tectonics: Alfred Wegener (1966) had demonstrated the concept of continental drift convincingly as early as 1912 but this idea only became acceptable in the early 1960s. One reason that this theory initially was rejected by most geoscientists was lack of a plausible mechanism for the movement of continents.

Along similar lines, Staub (1928) had argued that the principal force that controlled mountain building in the Alps was crustal shortening between Africa and Eurasia. Figure 1 (from Agterberg 1961, Fig. 107) shows tectonic sketch maps



Induction, Deduction, and Abduction, Fig. 1 Tectonic sketch maps of Strigno area on in northern Italy and Eastern Alps (after Agterberg 1961). Although area sizes differ by a factor of 1,600, the patterns showing overthrust sheets are similar. Originally, gravity sliding was assumed to be the major cause of these patterns. However, after development of the theory of plate tectonics in the 1960s, it has become

of two areas in the eastern Alps. The area in northern Italy shown at the top of Fig. 1 is 1600× smaller in area than the region for the eastern Alps at the bottom. The tectonic structure of these two regions is similar. Both contain overthrust

apparent that plate collision was the primary driving force of Alpine orogeny. The Strigno overthrust sheets are due to Late Miocene southward movement of the Eurasian plate over the Adria microplate, and the eastern Alpine nappes were formed during the earlier (primarily Oligocene) northward movement of the African plate across the Eurasian plate. (Source: Agterberg 1961, Fig. 107)

sheets with older rocks including crystalline basement overlying much younger rocks. It is now well known that the main structure of the map at the bottom was created during the Oligocene (33.9-23.0 Ma) when the African plate moved

northward over the Eurasian plate. On the other hand, the main structure of the relatively small Strigno area was created much later during the Late Miocene (Tortonian, 11.6–7.3 Ma) when the Eurasian plate moved southward overriding the Adria microplate that originally was part of the African plate.

Another geological idea that was conceived early on, but initially rejected as a figment of the imagination, is what later became known as the Milankovitch theory. Croll (1875) had suggested that the Pleistocene ice ages were caused by variations in the distance between the Earth and the Sun. Milankovitch commenced working on astronomical control of climate in 1913 (Schwarzacher 1993). His detailed calculations of orbital variations were published nearly 30 years later (Milankovitch 1941) showing quantitatively that amount of solar radiation drastically changes our climate. This theory was immediately rejected by climatologists because the changes in solar radiation due to orbital variations are miniscule, and by geologists as well because, stratigraphically, their correlation with ice ages apparently was not very good. However, in the mid-1950, new methods helped to establish the Milankovitch theory beyond any doubt. Subsequently, it resulted in the establishment of two new disciplines: cyclostratigraphy and astrochronology (Hinnov and Hilgen 2012). The latest geologic timescales of the Neogene (23.0–2.59 Ma) and Paleogene (66.0–23.0 Ma) periods are entirely based on astronomical calibrations, and “floating” astrochronologies are available for extended time intervals (multiple millions of years) extending through stages in the geologic timescale belonging to the older periods (Gradstein et al. 2020). The preceding three examples have in common that the underlying geologic processes were anticipated by individual geoscientists but remained in dispute until their existence could be proven mathematically beyond doubts.

Map-Making

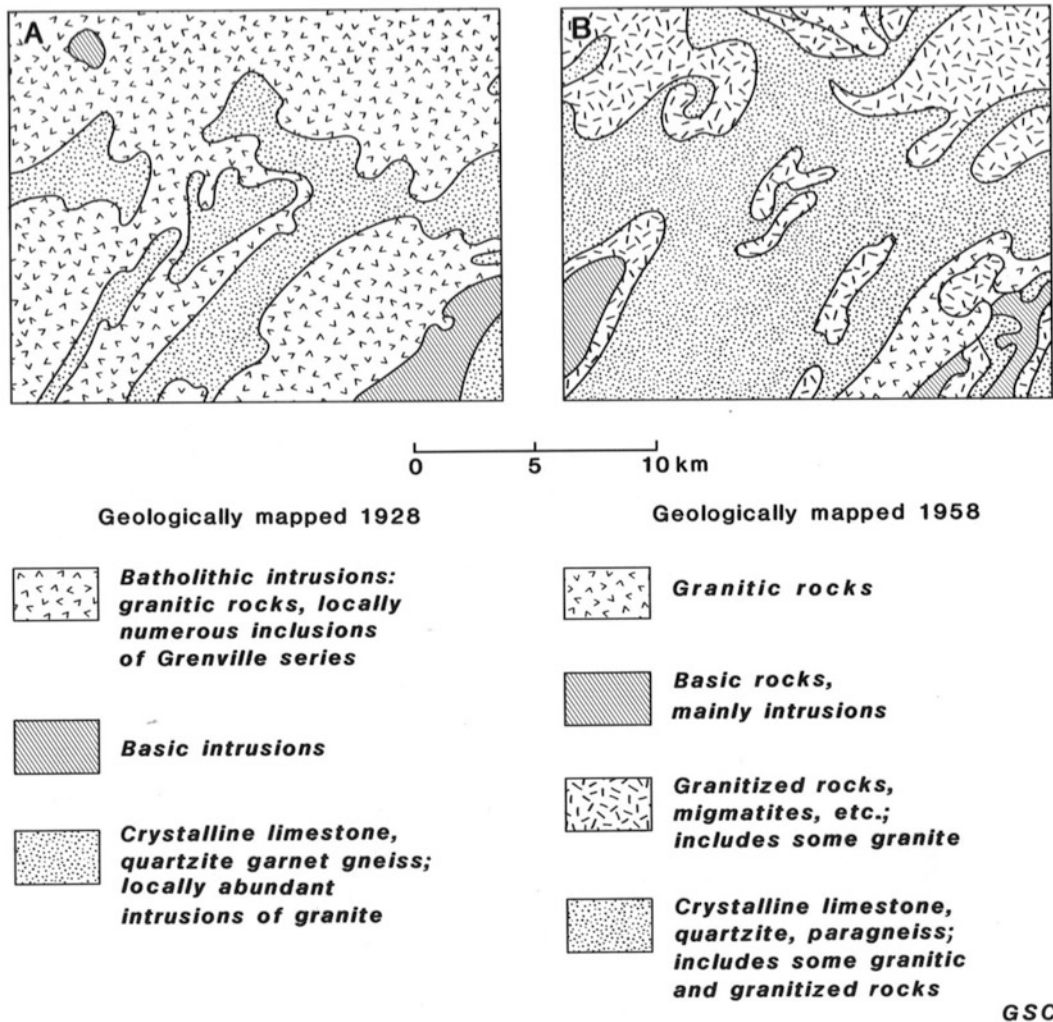
A field geologist is concerned with collecting numerous observations from those places where rocks are exposed at the surface. Observation is often hampered by poor exposure. In most areas, 90% or more of the bedrock surface is covered by unconsolidated overburden restricting observation to available exposures; for example, along rivers (cf. Agterberg and Robinson 1972). The existence of these exposures may be a function of the rock properties. In formerly glaciated areas, for example, the only rock that can be seen may be hard knobs of pegmatite or granite, whereas the softer rocks may never be exposed. Of course, drilling and geophysical exploration techniques provide additional information, but bore-holes are expensive and geophysics provides only partial, indirect information on rock composition, facies, age, and other properties of interest.

It can be argued that today most outcrops of bedrock in the world have been visited by competent geologists. Geologic maps at different scales are available for nearly all countries. One of the major accomplishments of stratigraphy is not only that the compositions but also the ages of rocks nearly everywhere at the Earth’s surface now are fairly well known. However, as Harrison (1963) has pointed out, although the outcrops in an area remain more or less the same during the immediate past, the geologic map constructed from them can change significantly over time when new geological concepts become available. A striking example is shown in Fig. 2. Over a 30-year period, the outcrops in this study area that were repeatedly visited by geologists had remained nearly unchanged. Discrepancies in the map patterns reflect changes in the state of geologic knowledge at different points in time.

Many geologists regard mapping as a creative art. From scattered bits of evidence, one must piece together a picture at a reduced scale, which covers at least most of the surface of bedrock in an area. Usually, this cannot be done without a good understanding of the underlying geologic processes that may have been operative at different geologic times. A large amount of interpretation is involved. Many situations can only be evaluated by experts. Although most geologists agree that it is desirable to make a rigorous distinction between facts and interpretation, this is hardly possible in practice, partly because during compilation results for larger regions must be represented at scales of 1: 25,000 or 1: 250,000, or less. Numerous observable features cannot be represented adequately in these scale models. Consequently, there is often significant discrepancy of opinions among geologists that can be bewildering to scientists in other disciplines and to others including decision-makers in government and industry.

Van Bemmelen (1961) has pointed out that the shortcomings of classical methods of geological observation constrain the quantification of geology. Much is left to the “feeling” and experience of the individual geologist. The results of this work, presented in the form of maps, sections, and narratives with hypothetical reconstructions of the geological evolution of a region, do not have the same exactitude as the records and accounts of geophysical and geochemical surveys, which are more readily computerized even though the results may be equally accurate in an interpretive sense. Geophysical and geochemical variables are determined by the characteristics of the bedrock geology which, in any given area, is likely to be nonuniform because of the presence of different rock types, often with abrupt changes at the contacts between them. The heterogeneous nature of the geologic framework will be reflected or masked in these other variables.

Geologists, geophysicists, and geochemists produce different types of maps for the same region. Geophysical measurements are mainly indirect. They include gravity and



Induction, Deduction, and Abduction, Fig. 2 Two geological maps for the same area on the Canadian Shield (after Harrison 1963). Between 1928 and 1958 there was development of conceptual ideas about the

nature of metamorphic processes. This, in turn, resulted in geologists deriving different maps based on the same observations

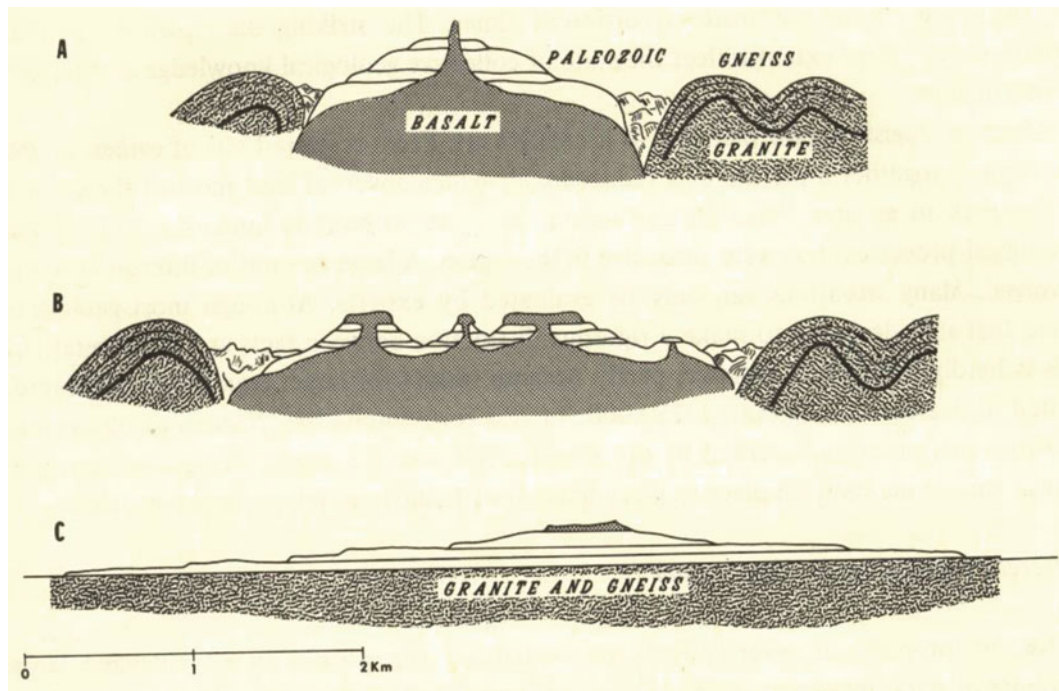
seismic data, variations in the Earth's magnetic field, and electrical conductivity. They generally are averages of specific properties of rocks over large volumes with intensities of penetration decreasing with distance and depth. Remotely sensed data are very precise and can be subjected to a variety of filtering methods. However, they are restricted to the Earth's surface. Geochemists mainly work with element concentration values determined from chips of rocks in situ, and also from samples of water, mud, soil, or organic material.

Geological Cross-Sections

The facts observed at the surface of the Earth must be correlated with one another; for example, according to a stratigraphic column. Continually, trends must be established and projected into the unknown. This is because rocks are 3-D

media that can only be observed in 2-D. The geologist can look at a rock formation but not inside it. Sound geological concepts are a basic requirement for making 3-D projections.

During the past two centuries, after William Smith, geologists have acquired a remarkably good capability of imagining 3-D configurations by conceptual methods. This skill was not obtained easily. In Fig. 3, according to Nieuwenkamp (1968), an example is shown of a typical geologic extrapolation problem with results strongly depending on initially erroneous theoretical considerations. In the Kinnehulle area in Sweden, the tops of the hills consist of basaltic rocks; sedimentary rocks are exposed on the slopes; granites and gneisses in the valleys. The first two projections into depth for this area were made by a prominent geologist (Von Buch 1842). It can be assumed that today most geologists would quickly arrive at the third 3-D reconstruction (Fig. 3c) by Westergård et al. (1943).



Induction, Deduction, and Abduction, Fig. 3 Schematic sections originally compiled by Nieuwenkamp (1968) showing that a good conceptual understanding of time-dependent geological processes is required for downward extrapolation of geological features observed at the surface of the Earth. Sections A and B are modified after von Buch

(1842) illustrating his genetic interpretation that was based on combining Abraham Werner's theory of Neptunism with von Buch's own firsthand knowledge of volcanoes including Mount Etna in Sicily with a basaltic magma chamber. Section C is after Westergård et al. (1943) (Source: Agterberg 1974, Fig. 1)

At the time of von Buch it was not yet common knowledge that basaltic flows can form extensive flows within sedimentary sequences. The projections in Fig. 3a, b reflect A.G. Werner's early pan-sedimentary theory of "Neptunism" according to which all rocks on Earth were deposited in a primeval ocean. Nieuwenkamp (1968) has demonstrated that this theory was related to philosophical concepts of F.W.J. Schelling and G.W.F. Hegel. When Werner's view was criticized by other early geologists who assumed processes of change during geologic time, Hegel publicly supported Werner by comparing the structure of the Earth with that of a house. One looks at the complete house only and it is trivial that the basement was constructed first and the roof last. Initially, Werner's conceptual model provided an appropriate and temporarily adequate classification system, although von Buch, who was a follower of Werner, rapidly ran into problems during his attempts to apply Neptunism to explain occurrences of rock formations in different parts of Europe. For the Kinnehulle area (Fig. 3) von Buch assumed that the primeval granite became active later changing sediments into gneisses, while the primeval basalt became a source for hypothetical volcanoes. Clearly, 3D time-dependent mapping continues to present important geoscientific challenges. It is likely that in future a major role will be played by new methodologies initially developed in mathematical morphology (Matheron 1981), a subject currently advanced by Sagar (2018) and colleagues.

Conditional Probability and Bayes' Theorem

Suppose that $p(D|B)$ represents the conditional probability that event D occurs given event B (e.g., a mineral deposit D occurring in a small unit cell underlain by rock type B on a geological map). This conditional probability obeys three basic rules (cf. Lindley 1987, p. 18):

1. Convexity: $0 \leq p(D|B) \leq 1$; D occurs with certainty if B logically implies D ; then, $p(D|B) = 1$, and $p(D^c|B) = 0$ where D^c represents the complement of D
2. Addition: $p(B \cup C|D) = p(B|D) + p(C|D) - p(B \cap C|D)$
3. Multiplication: $p(B \cap C|D) = p(B|D) p(C|B \cap D)$

These three basic rules lead to many other rules (cf. Agterberg 2014). For example, replacement of B by $B \cap D$ in the multiplication rule gives: $p(B \cap C|D) = p(B|D) p(C|B \cap D)$. Likewise, it is readily derived that: $p(B \cap C \cap D) = p(B|D) p(C|B \cap D) p(D)$. This leads to Bayes' theorem in odds form:

$$\frac{p(D|B \cap C)}{p(D^c|B \cap C)} = \frac{p(B|C \cap D)}{p(B|C \cap D^c)} \cdot \frac{p(D|C)}{p(D^c|C)}$$

or

$$O(D|B \cap C) = \exp(W_{B \cap C}) \cdot O(D|C)$$

where $O = p/(1-p)$ are the odds corresponding to $p = O/(1 + O)$, and $W_{B \cap C}$ is the “weight of evidence” for occurrence of the event D given B and C . Suppose that the probability p refers to occurrence of a mineral deposit D under a small area on the map (circular or square unit area). Suppose further that B represents a binary indicator pattern, and that C is the study area within which D and B have been determined. Under the assumption of equipossibility (or equiprobability), the prior probability is equal to total number of deposits in the study area divided by total number of unit cells in the study area. Theoretically, C is selected from an infinitely large universe (parent population) with constant probabilities for the relationship between D and B . In practical applications, only one study area is selected per problem and C can be deleted from the preceding equation. Then Bayes’ theorem can be written in the form:

$$\ln O(D|B) = W_B^+ + \ln O(D); \ln O(D|B^c) = W_B^- + \ln O(D)$$

for presence or absence of B , respectively. If the area of the unit cell underlain by B is small in comparison with the total area underlain by B , the odds O are approximately equal to the probability p . The weights satisfy:

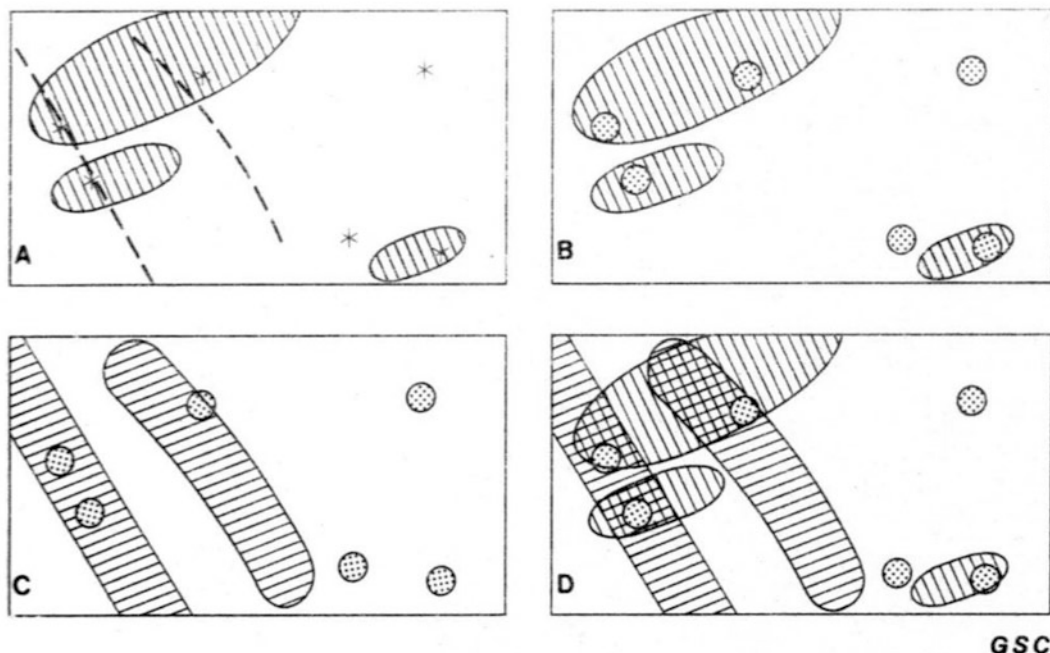
$$W_B^+ = \ln \frac{p(B \cap D)}{p(B \cap D^c)}; W_B^- = \ln \frac{p(B^c \cap D)}{p(B^c \cap D^c)}$$

As an example of this type of application of Bayes’ theorem, suppose that a study area C , which is a million times as large as the unit cell, contains 10 mineral deposits; 20% of C is underlain by rock type B , which contains 8 deposits. The prior probability $p(D)$ then is equal to 0.000 01; the posterior probability for a unit cell on B is $p(D|B) = 0.000 04$, and the posterior probability for a unit cell not on B is $p(D|B^c) = 2.5 \times 10^{-6}$. The weights of evidence are $W_B^+ = 0.982$ and $W_B^- = -1.056$, respectively. In this example, the two posterior probabilities can be calculated without use of Bayes’ theorem. However, the weights themselves provide useful information as will be seen in the next section.

The method of weights of evidence modeling was invented by Good (1950). In its original application to mineral resource potential mapping (Agterberg 1989; Bonham-Carter 1994), extensive use was made of medical applications (Spiegelhalter 1986).

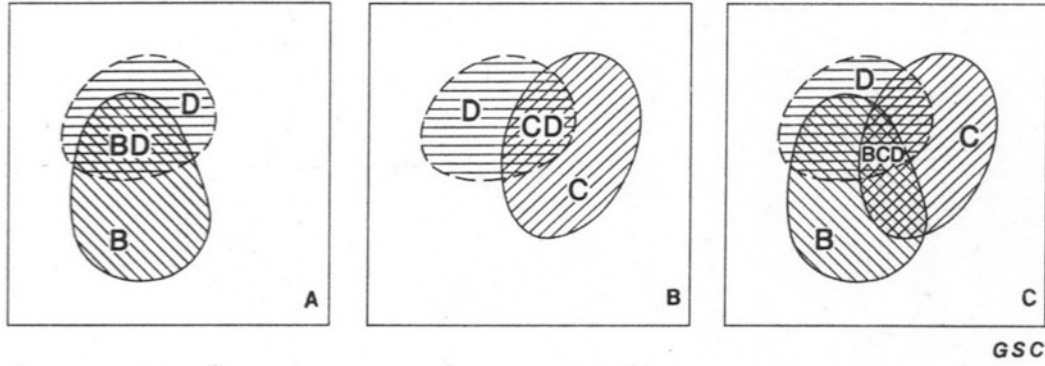
Basic Concepts and Artificial Example

Figure 4 illustrates the concept of combining two binary patterns for which it can be assumed that they are related to occurrences of mineral deposits of a given type. Figure 4a shows locations of six hypothetical deposits, the outcrop pattern of a rock type (B) with which several of the deposits may be associated (Fig. 4b), and two lineaments that have



Induction, Deduction, and Abduction, Fig. 4 Artificial example to illustrate the concept of combining two binary patterns related to occurrence of mineral deposits; (a) outcrop pattern of rock type, lineaments,

and mineral deposits; (b) rock type and deposits dilated to unit cells; (c) lineaments dilated to corridors; (d) superposition of three patterns. (Source: Agterberg et al. 1990, Fig. 1)



Induction, Deduction, and Abduction, Fig. 5 Venn diagrams corresponding to areas of binary patterns of Fig. 4; (a) is for Fig. 5.1B; (b) is for Fig. 5.1C; (c) is for Fig. 5.1D (Source: Agterberg et al. 1990, Fig. 2)

been dilatated in Fig. 4c. Within the corridors around the lineaments, the likelihood of locating deposits may be greater than elsewhere in the study area. In Fig. 4b–d, the deposits are surrounded by a small unit area. This permits estimation the unconditional “prior” probability $P(D)$ that a unit area with random location in the study area contains a deposit, as well as the conditional “posterior” probabilities $P(D|B)$, $P(D|C)$, and $P(D|BC)$ that unit areas located on the rock type, within a corridor and both on the rock type and within a corridor contain a deposit. These probabilities are estimated by counting how many deposits occur within the areas occupied by the polygons of their patterns. The relationships between the two patterns (B and C), and the deposits (D) can be represented by Venn diagrams as shown schematically in Fig. 5.

Operations such as creating corridors around line segments on maps and measuring areas can be performed by using Geographic Information Systems (GISs). The Spatial Data Modeller developed by Raines and Bonham-Carter (2007) is an example of a system that provides tools for weights of evidence, logistic regression, fuzzy logic, and neural networks. The availability of excellent software for WofE and WLR has been a factor in promoting widespread usage of these methods. Examples of applications to mapping mineral prospectivity can be found in Carranza (2004), and Porwal et al. (2010).

Bayes’ Rule for One or More Map Layers

When there is a single pattern B , the odds $O(D|B)$ for occurrence of mineralization if B is present is given by the ratio of the following two expressions of Bayes’ rule: $P(D|B) = \frac{P(B|D)P(D)}{P(B)}$; $P(\bar{D}|B) = \frac{P(B|\bar{D})P(\bar{D})}{P(B)}$ where the set $\bar{D}$ represents the complement of D . Consequently, $\ln O(D|B) = \ln O(D) + W^+$ where the positive weight for presence of B is: $W^+ = \ln \frac{P(B|D)}{P(B|\bar{D})}$.

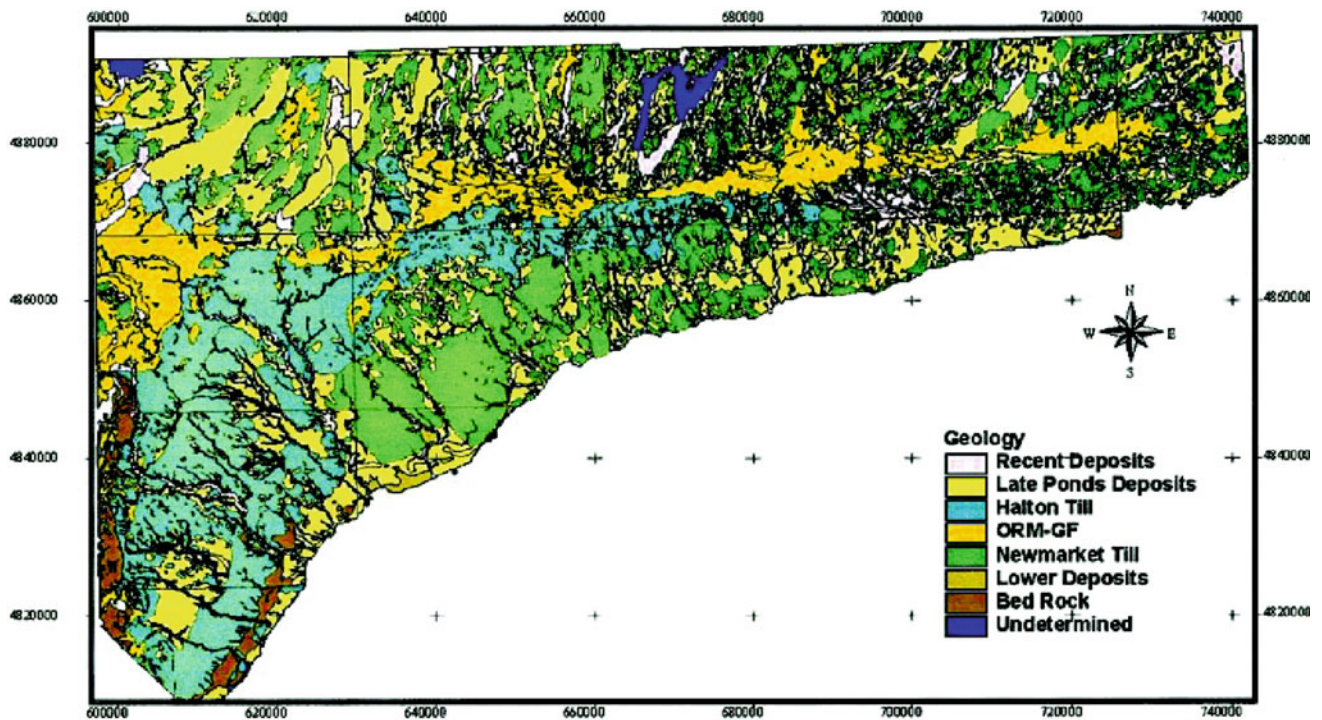
The negative weight for absence of B is: $W^- = \ln \frac{P(\bar{B}|D)}{P(\bar{B}|\bar{D})}$. The result of application of Bayes’ rule applied to a single map layer can be extended by using it as prior probability input for other map layers provided that there is approximate conditional independence (CI) of map layers. The order in which new patterns are added is immaterial. When there are two map patterns as in Fig. 5.1: $P(D|B \cap C) = \frac{P(B \cap C|D)P(D)}{P(B \cap C)}$ and $P(\bar{D}|B \cap C) = \frac{P(B \cap C|\bar{D})P(\bar{D})}{P(B \cap C)}$. Conditional independence of D with respect to B and C implies:

$$P(B \cap C|D) = P(B|D)P(C|D); P(B \cap C|\bar{D}) = P(B|\bar{D})P(C|\bar{D}).$$

Consequently,

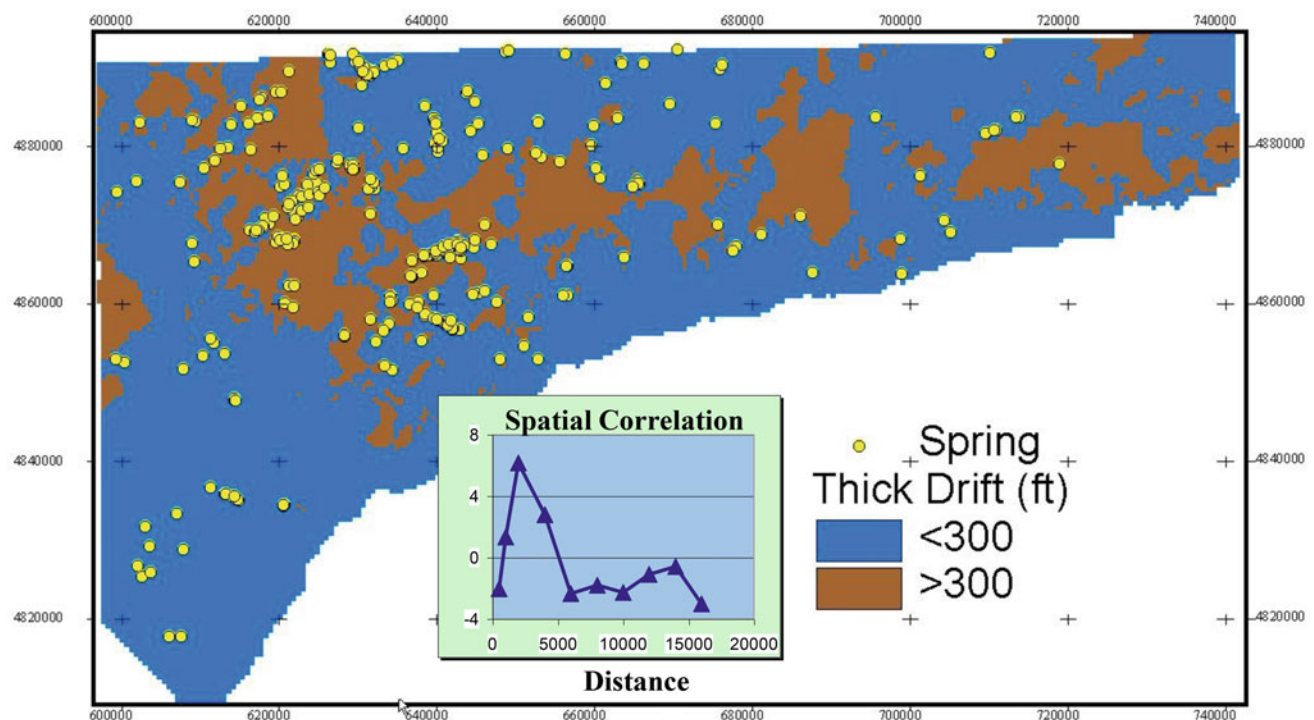
$$P(D|B \cap C) = P(D) \frac{P(B|D)P(C|D)}{P(B \cap C)}; P(\bar{D}|B \cap C) = P(\bar{D}) \frac{P(B|\bar{D})P(C|\bar{D})}{P(B \cap C)}.$$

From these two equations it follows that: $\frac{P(D|B \cap C)}{P(\bar{D}|B \cap C)} = \frac{P(D)P(B|D)P(C|D)}{P(\bar{D})P(B|\bar{D})P(C|\bar{D})}$. This expression is equivalent to $\ln O(D|B \cap C) = \ln O(D) + W_1^+ + W_2^+$. The posterior logit on the left side of this equation is the sum of the prior logit and the weights of the two map layers. The posterior probability follows from the posterior logit. Similar expressions apply when either one or both patterns are absent. Cheng (2008) has pointed out that, since it is based on a ratio, the underlying assumption is somewhat weaker than assuming conditional independence of D with respect to B_1 and B_2 . If there are p map layers, the final result is based on prior logit plus the p weights for these map layers. A good WofE strategy is first to achieve approximate conditional independence by pre-processing. A common problem is that final estimated probabilities usually are biased. If there are N deposits in a study area and the sum of all estimated probabilities is written as S ,



Induction, Deduction, and Abduction, Fig. 6 Surficial geology of Oak Ridge Moraine in study area according to Sharpe et al. (1997). (Source: Cheng 2004, Fig. 2)

Flowing vs. Distance from Thick Drift



Induction, Deduction, and Abduction, Fig. 7 Flowing wells versus distance from buffer zone around Oak Ridge Moraine in study area of Fig. 6. (Courtesy of Q. Cheng)

WofE often results in $S > N$. The difference $S - N$ can be tested for statistical significance (Agterberg and Cheng 2002). The main advantage of WofE in comparison with other methods such as weighted logistic regression is transparency in that it is easy to compare weights with one another. On the other hand, the coefficients resulting from logistic regression generally are subject to considerable uncertainty.

The contrast $C = W^+ - W^-$ is the difference between positive and negative weight for a binary map layer. It is a convenient measure for strength of spatial correlation between a point pattern and the map layer (Bonham-Carter et al. 1988; Agterberg 1989). It is somewhat similar to Yule's (1912) "measure of association" $Q = \frac{\alpha-1}{\alpha+1}$ with $\alpha = e^C$. Both C and Q express strength of correlation between two binary variables that only can assume the values 1 (for presence) or -1 (for absence). Like the ordinary correlation coefficient, Q is confined to the interval $[-1, 1]$. If the binary variables are uncorrelated, then $E(Q) = 0$.

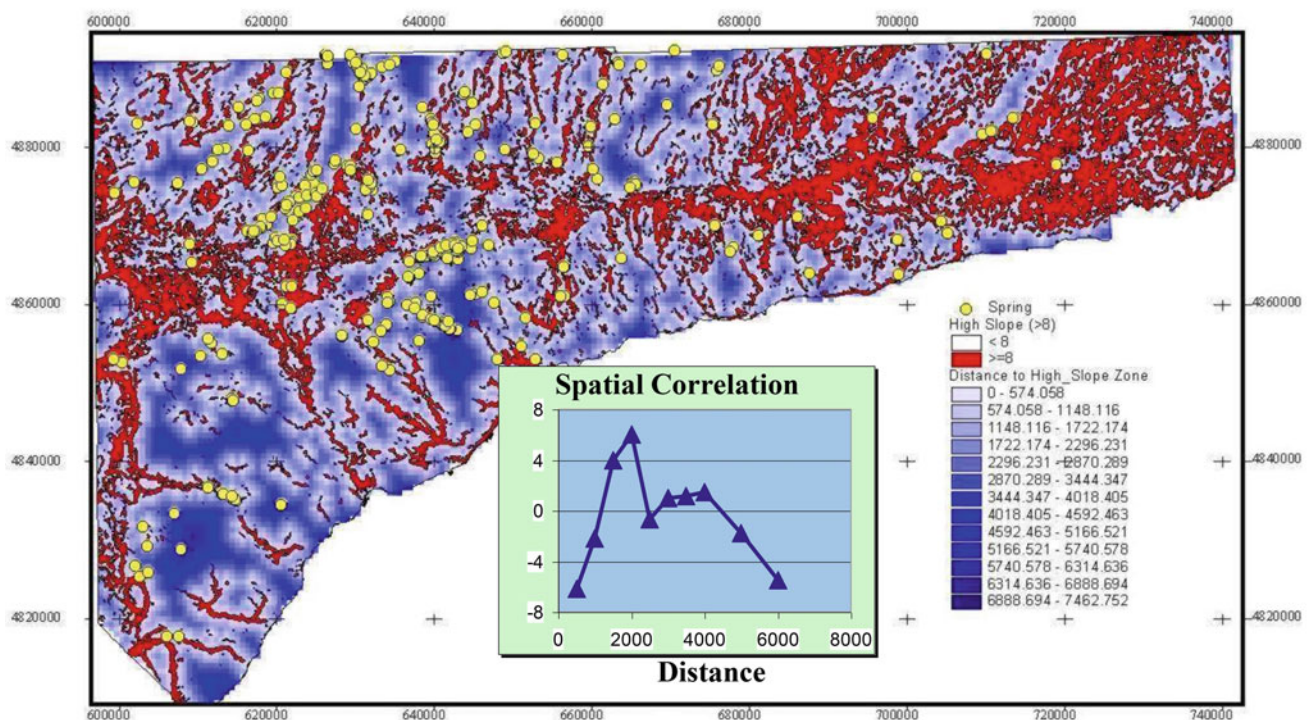
If a binary map layer is used as an indicator variable, the probability of occurrence of a deposit is greater when it is present than when it is absent, and W^+ is positive whereas W^- is negative. Consequently, C generally is positive.

However, in a practical application, it may turn out that C is negative. This would mean that the map layer considered is not an indicator variable. In a situation of this type, one could switch presence of the map layer with its absence, so that W^+ and C both become positive (and W^- negative). An excellent strategy often applied in practice is to create corridors of variable width x around linear map features that are either lineaments as in Fig. 4 or contacts between different rock types (e.g., boundaries of intrusive bodies) and to maximize $C(x)$. From $\frac{dQ}{dx} = \frac{2}{(\alpha+1)^2}$ being positive it follows that $Q(x)$ and $C(x)$ reach their maximum value at the same value of x (cf. Agterberg et al. 1990). Alternative approaches to WofE as discussed here have more recently been proposed by Baddeley et al. (2020).

Flowing Wells in the Greater Toronto Area

Cheng (2004) has given the following application of Weights-of-Evidence modeling. Figure 6 shows surficial geology Oak Ridge Moraine (ORM) for a study on assessment of flowing water wells in surficial geology in the Greater Toronto area,

Flowing Wells vs. Distance From High Slope Zone



Induction, Deduction, and Abduction, Fig. 8 Flowing wells versus distance from relatively steep slope zone in study area of Fig. 6. (Courtesy of Q. Cheng)

Ontario. The ORM is a 150 km long east-west trending belt of stratified glaciofluvial-glaciolacustrine deposits. It is 5–15 km wide and up to 150 m thick. For more detailed discussion of the geology of ORM, see Sharpe et al. (1997). It is generally recognized that the ORM is the main source of recharge in the area. Figures 7 and 8 show flowing wells together with two map patterns related to ORM thought to be relevant for flowing well occurrence.

Cheng (2004) used Weights-of-Evidence to test the influence of the ORM on locations of flowing wells. A number of binary patterns were constructed by maximizing the contrast C for distance from ORM. C reaches its maximum at 2 km meaning that a YES-NO binary pattern with YES on points belonging to ORM plus all points that occur less than 2 km from ORM, and NO for the remainder of the study area with points that are on ORM plus points more than 2 km away from ORM provides positive and negative weight that would be best to use for a binary pattern of enlarged ORM for which the contrast $C = W^+ - W^-$ reaches its maximum value of spatial correlation.

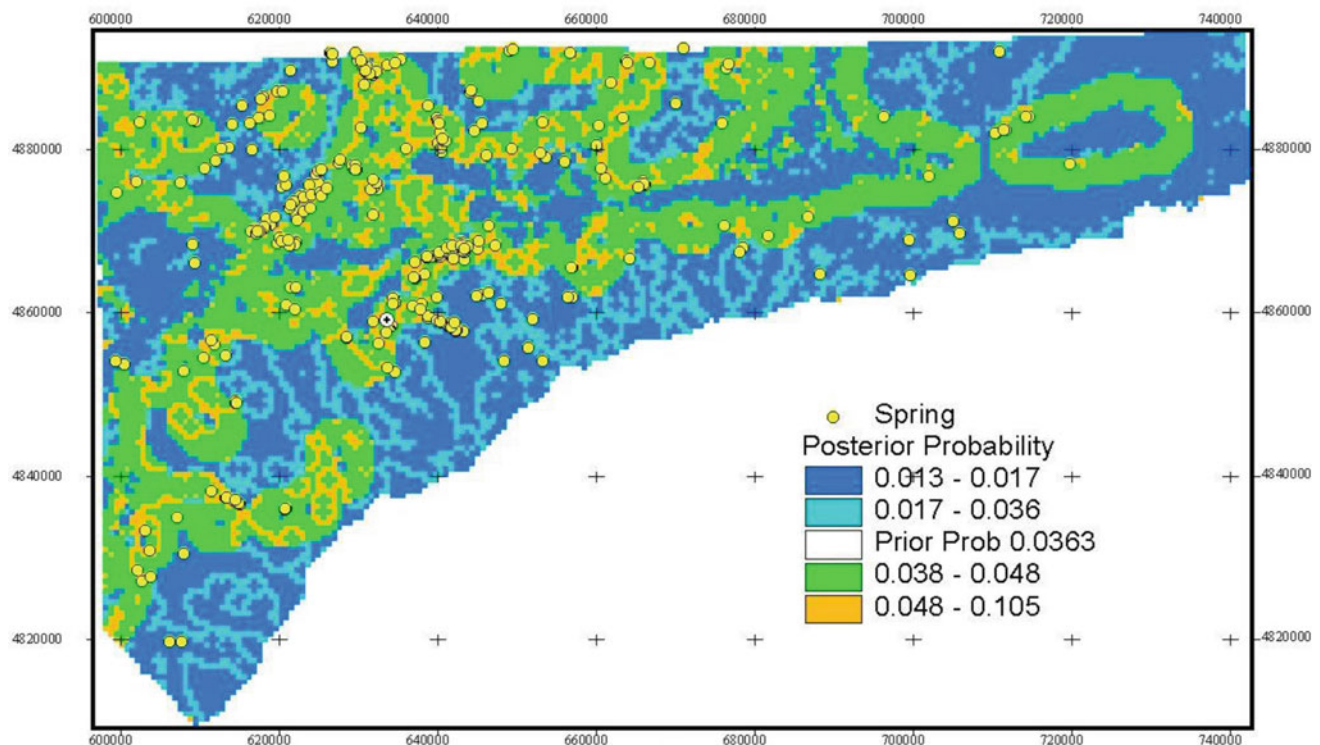
Two other binary patterns constructed in the same way by Cheng (2004) were distance from enlarged ORM (Fig. 7), and

distance from a relatively thick glacial drift layer (Fig. 8). The posterior probability map shown in Fig. 9 is based on binary maps at which the spatial correlation reaches their maximum values in Figs. 7 and 8, respectively. It not only quantifies the relationship between the known artesian aquifers and these two binary map layers, it also outlines areas with no or relatively few aquifers that have good potential for additional aquifers.

Summary and Conclusions

The philosophical terms “induction” and “deduction” are more frequently used than “abduction.” Induction is a method of reasoning from premises that supply some evidence without full assurance of the truth of the conclusion that is reached. It differs from deduction that starts from premises that are correct and leads to conclusions that are certainly true. Abduction starts with observations from which it seeks to find the simplest and most likely conclusion. Originally, geology was a science of induction and abduction.

Posterior Probability Map calculated by Arc-WofE from buffer zones around ORM and steep slope zones



Induction, Deduction, and Abduction, Fig. 9 Posterior probability map for flowing wells based on buffer zone around Oak Ridge Moraine (Fig. 7) and steep slope zone area (Fig. 8) in study area. (Courtesy of Q. Cheng)

Mathematical statistics commenced as a science of deduction. Induction was introduced slowly with the concept of “equiprobability” and after increasing popularity of Bayes’ rule. Until recently, there remained significant disagreement between Bayesian statisticians and those, sometimes called “frequentists,” who avoided subjective notions in their statistical modeling. Bayesian data analysis aims to improve estimates of statistical parameters initially obtained by subjective reasoning. In the example of flowing wells given in this entry the prior probabilities are based on the assumption that the wells are randomly distributed across the study area but the posterior probabilities incorporate various types of evidence pertaining to the geographic locations of the wells.

Cross-References

- [Bayes’s Theorem](#)
- [Digital Geological Mapping](#)
- [Logistic Regression, Weights of Evidence, and the Modeling Assumption of Conditional Independence](#)

Bibliography

- Agterberg FP (1961) Tectonics of the crystalline basement of the Dolomites in North Italy. Kemink, Utrecht, 232 pp
- Agterberg FP (1974) Geomathematics. Elsevier, Amsterdam, 596 pp
- Agterberg FP (1989) Computer programs for mineral exploration. *Science* 245:76–81
- Agterberg FP (2014) Geomathematics: theoretical foundations, applications and future developments. Springer, Heidelberg, 553 pp
- Agterberg FP, Cheng Q (2002) Conditional independence test for weights-of-evidence modeling. *Nat Resour Res* 11:249–255
- Agterberg FP, Robinson SC (1972) Mathematical problems in geology. *Proc 38th Sess Intern Stat Inst, Bull* 38:567–596
- Agterberg FP, Bonham-Carter GF, Wright DF (1990) Statistical pattern integration for mineral exploration. In: Gaal G, Merriam DF (eds) *Computer applications in resource estimation, prediction and assessment for metals and petroleum*. Pergamon, Oxford, pp 1–21
- Arbuthnot J (1710) An argument for divine providence from the constant regularity observed in the births of bot sexes. *Phil Trans R Soc London* 27:186–190
- Baddeley A, Brown W, Milne RK, Nair G, Rakshit S, Lawrence T, Phatak A, Fu SC (2020) Optimal thresholding of predictors in mineral prospectivity analysis. *Nat Resour Res*. <https://doi.org/10.1007/s1053-020-09769-2>
- Bayes T (1763) An essay towards solving a problem in the doctrine of chances. *Phil Trans R Soc London* 53:370–418
- Bonham-Carter GF (1994) Geographic information systems for geoscientists: modelling with GIS. Pergamon, Oxford, 398 pp
- Carranza EJ (2004) Weights of evidence modeling of mineral potential: a case study using small number of prospects, Abra, Philippines. *Nat Resour Res* 13(3):173–185
- Chamberlin TC (1897) The method of multiple working hypotheses. *J Geol* 5:837–848
- Cheng Q (2004) Application of weights of evidence method for assessment of flowing wells in the Greater Toronto Area, Canada. *Nat Resour Res* 13(2):77–86
- Cheng Q (2008) Non-linear theory and power-law models for information integration and mineral resources quantitative assessments. In: Bonham-Carter GF, Cheng Q (eds) *Progress in geomathematics*. Springer, Heidelberg, pp 195–225
- Croll J (1875) Climate and time in their geological relations. Appleton, New York
- Fisher-Box J (1978) R.A. Fisher – The life of a scientist. Wiley, New York, 512 pp
- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari RDB (2013) Bayesian data analysis, 3rd edn. CRC Press, Boca Raton, 660 pp
- Good H (1950) Probability and the weighing of evidence. Griffin, London
- Gradstein FM, Ogg JG, Schmitz MB, Ogg GM (eds) (2020) Geologic time scale, two volumes. Elsevier, Amsterdam, 809 pp
- Hacking I (2006) The emergence of probability, 2nd edn. Cambridge University Press, 209 pp
- Harrison JM (1963) Nature and significance of geological maps. In: Albritton CC Jr (ed) *The fabric of geology*. Addison-Wesley, Cambridge, MA, pp 225–232
- Hinnov IA, Hilgen FJ (2012) Cyclostratigraphy and astrochronology. In: Gradstein FM, Ogg JG, Schmitz M, Ogg G (eds) *The Geologic Time Scale 2012*. Elsevier, Amsterdam, pp 63–113
- Jeffries H (1939) The theory of probability, 3rd edn. Oxford University Press, 492 pp
- Lindley DV (1987) The probability approach to the treatment of uncertainty in artificial intelligence and expert systems. *Stat Sci* 2(1):17–24
- Matheron G (1981) Random sets and integral geometry. Wiley, New York
- Mcgrayve SB (2011) The theory that would not die. Yale University Press, New haven, 320 pp
- Milankovitch M (1941) Kanon der Erdbestrahlung und eine Auswendung auf das Eiszeitproblem. *R Serb Acad Spec Publ* 32, Belgrade
- Nieuwenkamp W (1968) Natuurfilosofie en de geologie van Leopold von Buch. *K Ned Akad Wet Proc Ser B* 71(4):262–278
- Porwal A, González-Álvarez I, Markwitz V, McCuaig TC, Mamuse A (2010) Weights-of-evidence and logistic regression modeling of magmatic nickel sulfide prospectivity in the Yilgarn Craton, Western Australia. *Ore Geol Rev* 38:184–196
- Sagar BSD (2018) Mathematical morphology in geosciences and GISci: an illustrative review. In: Sagar BSD, Cheng Q, Agterberg FP (eds) *Handbook of mathematical geosciences. Fifty years of IAMG*. Springer, Heidelberg, pp 703–740
- Schwarzacher W (1993) Cyclostratigraphy and the Milankovitch theory. Elsevier, Amsterdam
- Sharpe DR, Barnett PJ, Brennand TA, Finley D, Gorrel G, Russell HAJ (1997) Surficial geology of the greater Toronto and Ridges Moraine areas, compilation map sheet. *Geol Surv Can Open File* 3062, map 1: 200,000
- Spiegelhalter DJ (1986) Uncertainty in expert systems. In: Gale WA (ed) *Artificial intelligence and statistics*. Addison-Wesley, Reading, pp 17–55
- Staat W (1993) On abduction, deduction, induction and the categories. *Transact Charles S Peirce Soc* 29:97–219
- Staub R (1928) *Bewegungsmechanismen der Erde*. Borntraeger, Stuttgart
- Van Bemmelen RW (1961) The scientific nature of geology. *J Geol*:453–465
- Von Buch L (1842) Ueber Granit und Gneiss, vorzüglich in Hinsicht der äusseren Form, mit welcher diese Gebirgsarten auf Erdoberfläche erscheinen. *Abh K Akad Berlin IV/2*(18840):717–738
- Wegener A (1966) The origin of continents and oceans, 4th edn. Dover, New York

- Westergård AH, Johansson S, Sundius N (1943) *Beskrivning till Kartbladet Lidköping*. Sver Geol Unders Ser Aa 182
- Winchester S (2001) *The map that changed the world*. Viking, Penguin Books, London
- Yule GU (1912) On the methods of measuring association between two attributes. *J Roy Stat Soc* 75:579–642

International Generic Sample Number

Sarah Ramdeen¹, Kerstin Lehnert¹, Jens Klump² and Lesley Wyborn³

¹Columbia University, Lamont Doherty Earth Observatory, Palisades, NY, USA

²Mineral Resources, CSIRO, Perth, WA, Australia

³Research School of Earth Sciences, Australian National University, Canberra, ACT, Australia

Synonyms

IGSN

Definition

IGSN (International Generic Sample Number) is a persistent identifier for material samples and specimens. Material samples (also referred to as physical samples) include objects such as cores, minerals, fossil specimens, synthetic specimens, water samples, and more. The concept of the IGSN was originally developed for the earth sciences but has seen growing adoption in other domains including, but not limited to, natural and environmental sciences, materials sciences, physical anthropology, and archaeology.

Persistent identifiers are unique labels given to an object (often a digital resource) in order to facilitate unique identification. In the case of IGSN, the identifier is assigned as part of a metadata record representing a material sample. The metadata record includes provenance and other information from along the samples lifecycle such as where the sample was collected, who collected it, and descriptive metadata ranging from spatial coordinates, collection method, mineral classification to identifiers for related publications. Once an IGSN is assigned, it does not change, it is persistent and globally unique. The metadata record may be updated but the identifier will remain unchanged. IGSN provides sample citations in the literature which can be used to unambiguously track a sample and build connections between data and analysis derived from the same samples in different applications (Conze et al. 2017).

A key feature of an IGSN is that it is “resolvable.” IGSNs can be used to retrieve sample metadata by using Internet protocols to obtain the digital resource the IGSN identifies.

IGSN uses the Handle system (<http://www.handle.net/>), the same technical infrastructure used by Digital Object Identifiers (DOIs), to resolve an IGSN to the appropriate resource location on the web. To resolve an IGSN, one would append the URL “<https://hdl.handle.net/10273/>” with the identifier. For example, to resolve the IGSN “BFBG-154433,” visit <https://hdl.handle.net/10273/BFBG-154433>. The resulting URL will “resolve” to the webpage hosting the metadata profile for the sample represented by the IGSN.

While IGSN functions similarly to DOIs, they do have important differences. DOIs are used to identify objects like journal articles, publications, and datasets. DOIs have standard metadata requirements, which are optimized for the types of objects they support. IGSN have metadata requirements that are tailored to material samples. They include metadata fields that document the scientific nature of the samples and their provenance. The metadata fields can be community specific, capturing the critical metadata for a particular community or domain.

IGSNs, like other persistent identifiers, are important because they help support data management and comply with the principles system for making scientific data and samples FAIR: findable, accessible, interoperable, and reusable (Wilkinson et al. 2016). FAIR samples support reproducible research, enable access to publicly funded assets, and allow researchers to leverage existing research investments by way of new analytical techniques or by augmenting existing data through reanalysis. For example, the use of IGSN allows scientists to unambiguously link analytical data for individual samples which were generated in different labs and published in different articles or data systems. Ultimately these identifiers will allow researchers to link data, literature, investigators, institutions, etc., creating an “Internet of samples” (Wyborn et al. 2017). IGSNs are a recommended persistent identifier by the Coalition for Publishing Data in the Earth and Space Sciences (COPDESS) and publishers such as AGU and Elsevier.

History and Evolution of IGSN

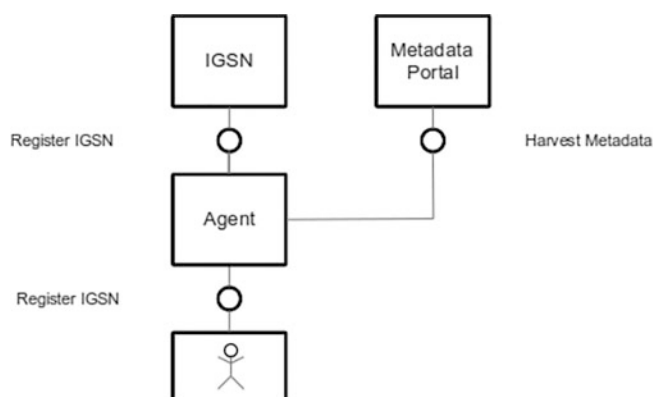
The concept of the IGSN was developed in 2004 at the Lamont-Doherty Earth Observatory at Columbia University with support from the National Science Foundation (NSF) (Lehnert et al. 2004). The first registration agent (also known as an allocating agent) to issue IGSNs was SESAR, the System for Earth Sample Registration (www.geosamples.org). The IGSN e.V. (www.igsn.org), the implementation organization of the IGSN, was formed in 2011. It is an international, member based, nonprofit organization which oversees the federated IGSN registration service and operates the IGSN Central Registry. The IGSN e.V. has more than 20 members (full and affiliate). The members are defined as

organizations who operate or are developing an allocating agent. Affiliate members are organizations which support the IGSN community and use existing allocating agents for sample registration. As part of the federated IGSN system, allocating agents “allocate” IGSN identifiers and ensure the persistence of metadata records and IGSN metadata profile pages.

In 2018, the Alfred P. Sloan Foundation funded the IGSN 2040 project “to re-design and mature the existing organization and technical architecture of the IGSN to create a global, scalable, and sustainable technical and organizational infrastructure for persistent unique identifiers (PID) of material samples” (IGSN e.V. 2019). The primary outcomes of the project have led to negotiations with DataCite to establish a partnership which will provide scalable and sustained IGSN registration, development of a samples “Community of Communities” to scale activities such as standards for sample identifiers, community engagement, broader adoption of IGSN, and a redesigned technical architecture which embraces modern web technologies (Lehnert et al. 2021a).

Current System Architecture of the IGSN e.V.

In simplified terms, the IGSN system architecture (Fig. 1) is made up of four roles: the IGSN Central Registry, allocating agents, end users, and metadata portals. As previously mentioned, the **IGSN Central Registry** is based on the Handle System. The Handle System is a distributed system for assigning identifiers (handles) for digital objects while ensuring that the identifier is unique and persistent (Arms and Ely 1995). The IGSN Central Registry is a global server which stores the identifiers, registration metadata about the objects being identified, and provides a service to resolve identifiers. The IGSN Central Registry uses a REST API (<https://doibb.wdc-terra.org/igsn/static/apidoc>) for its federated registration service.



International Generic Sample Number, Fig. 1 Simplified system architecture of the IGSN registration. Source: <https://igsn.github.io/system/>

Allocating agents are the local service providers. They provide a range of services including registration (or minting) services to one or more sample repositories or data centers. See Table 1 below for a list of IGSN allocating agents and the community they serve. New agents will emerge as new communities adopt IGSN.

Allocating agents capture registration metadata and descriptive metadata, which make up the IGSN metadata kernel. Registration metadata are the metadata elements required for the registration and administration of IGSN (<http://schema.igsn.org/registration/>). Descriptive metadata are dependent on the “community of practice” of the allocating agent and capture information used to catalog samples. This metadata is used to support discovery and access to samples. Descriptive metadata are not stored by the IGSN Central Registry. Allocating agents are responsible for storing the descriptive metadata and providing access to it for harvesting using the IGSN metadata kernel or Dublin Core.

End users of IGSN range from individual researchers or research teams to curators at sample repositories or museums. They may be using IGSN in their management of material sample collections, small and large. These users make decisions about what samples to register with IGSN and create metadata records. They register their sample metadata at an allocating agent. Allocating agents ensure the persistence of the submitted metadata and, depending on the agent, may provide curatorial review of the records to ensure they adhere to community standards. Ultimately, the users are responsible for maintaining the accuracy and quality of their metadata content. End users of IGSN also include those looking for existing samples to reuse. Samples may be reused in new and different ways or to validate previous work. Rich descriptive metadata allows new users to evaluate existing samples and may help to reduce the need to conduct field research to gather new samples at the same site.

Metadata portals are sample catalogs which harvest metadata from IGSN allocating agents. These metadata portals are used to support discovery and access to material samples. Metadata portals harvest sample metadata from allocating agents using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) protocols. IGSN metadata portals serve specific communities of practice. The various sample communities require different metadata to identify and describe the scientific qualities of their research objects as defined by their community. There isn’t a “one size fits all” approach to sample metadata. Example community portals can be viewed at <http://igsn2.csiro.au/> and <http://www.geosamples.org/>.

Syntax and Metadata

IGSN are represented by string values which can contain the following characters `^[A-Za-z]{2-4}[A-Za-z0-9.-]{1-71}$`. They are not intended to be read by humans but can contain some information embedded in the identifier. IGSN begins

International Generic Sample Number, Table 1 List of IGSN allocating agents

Allocating agent name	URL	Community of practice
Australian Research Data Commons (ARDC)	https://ardc.edu.au/services/identifier/igsn/	Sample holders and curators affiliated with Australian research organizations that do not offer their own IGSN minting service
Christian Albrechts Universität Kiel	https://www.ifg.uni-kiel.de/en/	Kiel facilities and staff
CSIRO Mineral Resources	http://www.csiro.au/en/Research/MRF	CSIRO facilities and staff
Cyber Carothèque Nationale (CNRS)	https://www.igsn.cnrs.fr/	Under development
Geoscience Australia	http://ldweb.ga.gov.au/igsn/	Australian Government Geological Surveys and other Australian state and federal government users
German Research Centre for Geosciences GFZ, Potsdam	http://dataservices.gfz-potsdam.de/mesi/overview.php?id=38	GFZ facilities and staff
Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER)	https://campagnes.flotteoceanographique.fr/search	IFREMER facilities and staff
Korea Institute of Geoscience & Mineral Resources (KIGAM)	https://data.kigam.re.kr/	Korea's geoscience research and academic sectors
MARUM Centre for Marine Environmental Sciences, Univ. Bremen	http://www.marum.de/	MARUM facilities and staff
System for Earth Sample Registration (SESAR)	http://www.geosamples.org/	Supports registration by the general public

with a prefix which includes “super-namespace” codes among the characters which represent the allocating agent where the IGSNs were minted (e.g., AU for Geoscience Australia). At SESAR, IGSNs have a five-digit prefix, which includes the super-namespace code IE, followed by a three-character personal namespace selected by the user. IGSNs are typically less than 20 characters in length (e.g., AU999971; IEUJA0011; ICDP5054EHW1001). The remaining characters after the super-namespace and personal namespace are randomly assigned during the registration process. Depending on the allocating agent, the characters after the super-namespace and personal namespace may also be user assigned.

There are some exceptions to the five-digit prefix. IGSNs registered at SESAR before the founding of the IGSN e.V. are grandfathered in with three-digit prefixes. These IGSN do not include the super-namespace code signifying the allocating agent where the samples were registered. This concept of the super-namespace was added once the federated concept of the IGSN e.V. was developed.

IGSN operates using the handle-namespace “10273.” The address of the IGSN resolver is <http://hdl.handle.net/10273>. As previously outlined, IGSN can be resolved by appending the IGSN to this address, for example, <http://hdl.handle.net/10273/IEUJA0011>. This address then redirects to the metadata profile page (Fig. 2) maintained by the allocating agent which registered the sample.

As previously mentioned, the IGSN metadata kernel is divided into two major categories, registration metadata and descriptive metadata. Registration metadata are the metadata elements required by the IGSN e.V. as part of the minting process. Descriptive metadata are dependent on the community of practice of the allocating agent. The IGSN e.V. has a

suggested list of descriptive metadata which “provide a minimum set of elements to describe a geological sample. The schema only contains elements that do not change over time, in analogy to being a ‘birth certificate’ of a sample” (IGSN e.V. 2016). Descriptive metadata is not prescribed by the IGSN e.V. as the allocating agents are best situated to determine what metadata is required to support their community of practice. The IGSN metadata kernel aligns with the DataCite Metadata Schema 4.0.

Future Technical Architecture of the IGSN

In order to increase the sustainability and scale IGSN registration capabilities to the billions, the IGSN 2040 project has developed a new architecture for the IGSN system and services (Klump et al. 2020). The redesign involves moving from XML based metadata schemas to using web architecture principles where metadata would be stored using JSON-LD. The use of JSON-LD will allow allocating agents greater flexibility in supporting community specific variations on their metadata profiles. Ultimately the use of web architecture principles will support the scaling of publishing and harvesting of IGSNs to billions of samples, which will lead to broader services for metadata portals (referred to as metadata aggregators in the proposed architecture) and end users. These services include the support of knowledge graphs which might connect samples to other related identifiers or resources.

Adoption of IGSN

As of 2021, nearly ten million samples have been registered with IGSN. IGSN has been adopted by geological surveys in

International Generic Sample Number, Fig. 2 Sample metadata profile. (Source: IEUJA0011)

IGSN: IEUJA0011



146_AL06-38.JPG (primary image)



IGSN: IEUJA0011
Sample Name: Al06-38
Other Name(s):
Sample Type: Individual Sample>Specimen
Parent IGSN: Not Provided

Description

Material:	Rock
Classification:	Metamorphic>Meta-Ultramafic
Field Name:	Orthopyroxene-bearing serpentinite

the USA, the UK, Australia, Korea, and other organizations and universities globally. In the USA, current users include the Smithsonian Institution, the US Department of Energy, NOAA-NCEI, US core repositories, and the International Continental Drilling Program.

IGSN is a recommended best practice in support of Open and FAIR samples by publishers and sample focused communities. IGSN can be used to track samples and link sample data along its lifecycle from field collection to analysis and publication and to archive and reuse. IGSN are being incorporated into data systems in order to create an efficient, collaborative pipeline (Lehnert et al. 2021b). This includes iSamples (www.isamples.org), StraboSpot (www.strabospot.org), EarthChem (www.earthchem.org), SPARROW (sparrow-data.org), and Macrostrat (www.macrostrat.org).

Moving Forward

IGSN e.V. and DataCite will finalize their partnership in late 2021. As a result, the system and technical architecture of the IGSN may undergo some changes. This includes the implementation of the newly developed technical architecture built on web architecture principles.

As interest and adoption of IGSN are expanding beyond the earth sciences, the IGSN e.V. will be changing the definition of the IGSN acronym. The change from geo to something more inclusive will better represent the diverse sample community around the IGSN. The change is expected to be announced in early 2022.

Summary

IGSN is a persistent identifier for material samples and specimens. They are used as citations to samples which can be tracked throughout their lifecycle. IGSN are globally unique and are resolvable, meaning they use Internet protocols to direct users to a persistent webpage which hosts metadata about the sample they represent. IGSN can be obtained from one of the allocating agents listed in Table 1. Allocating agents represent different communities of practice and may have different requirements for the descriptive metadata about a sample. The IGSN e.V., the implementation office of the IGSN, supports the IGSN Central Registry. Allocating agents mint IGSN through the Central Registry and must make their sample metadata available for harvesting by metadata portals. Allocating agents may offer these services themselves. The IGSN 2040 project made recommendations to scale the IGSN towards the future. This includes a partnership with DataCite and the design of a web architecture-based system. These changes will enable broader services related to the IGSN and allow IGSN to scale towards the future.

Cross-References

- [Data Acquisition](#)
- [Data Life Cycle](#)
- [FAIR Data Principles](#)
- [Geoinformatics](#)

- Machine Learning
- Metadata
- Open Data

Bibliography

- Arms W, Ely D (1995) The handle system: a technical overview. Available via <http://www.cnri.reston.va.us/home/ctr/handle-overview.html>. Accessed 30 Sept 2021
- Conze R, Lorenz H, Ulbricht D, Elger K, Gorgas T (2017) Utilizing the international generic sample number concept in continental scientific drilling during ICDP expedition COSC-1. *Data Sci J* 16:2. <https://doi.org/10.5334/dsj-2017-002>
- IGSN e.V. (2016) Metadata. In: Technical Documentation of the IGSN. Available via <https://igsn.github.io/metadata/>. Accessed 30 Sept 2021
- IGSN e.V. (2019) IGSN 2040. In: IGSN. Available via <https://www.igsn.org/igsn-2040/>. Accessed 30 Sept 2021
- Klump J, Lehnert K, Wyborn L, Ramdeen S (2020) IGSN 2040 technical steering committee meeting report. Zenodo <https://doi.org/10.5281/zenodo.3724683>
- Lehnert K, Goldstein S, Lenhardt W, Vinayagamoorthy S (2004) SESAR: addressing the need for unique sample identification in the Solid Earth Sciences. American Geophysical Union Fall Meeting 2004 SF32A-06
- Lehnert K, Klump J, Ramdeen S, Wyborn L, HaakL (2021a) IGSN 2040 summary report: defining the future of the IGSN as a global persistent identifier for material samples. Zenodo <https://doi.org/10.5281/zenodo.5118289>
- Lehnert K, Quinn D, Tikoff B, Walker D, Ramdeen S, Profeta L, Peters S, Pauli J (2021b) Linking data systems into a collaborative pipeline for geochemical data from field to archive. EGU General Assembly Conference Abstracts 2021EGUGA..2313940L
- Wilkinson M, Dumontier M, Aalbersberg I et al (2016) The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 3:160018. <https://doi.org/10.1038/sdata.2016.18>
- Wyborn L, Klump J, Bastrakova I, Devaraju A, McInnes B, Cox S, Karssies L, Martin J, Ross S, Morrissey J, Fraser R (2017) Building an Internet of Samples: The Australian Contribution. EGU General Assembly Conference Abstracts 2017EGUGA..1911497W

Interpolation

Jaya Sreevalsan-Nair
Graphics-Visualization-Computing Lab, International
Institute of Information Technology Bangalore, Electronic
City, Bangalore, Karnataka, India

Interpolation is the methodology by which unknown values of a variable are computed within a range of discrete data points using their known values. These values, both known and unknown, are governed by the assumption of an underlying continuous function. Interpolation involves identifying or approximating this continuous function, which is then used to compute the values at desired data points.

Definition and History

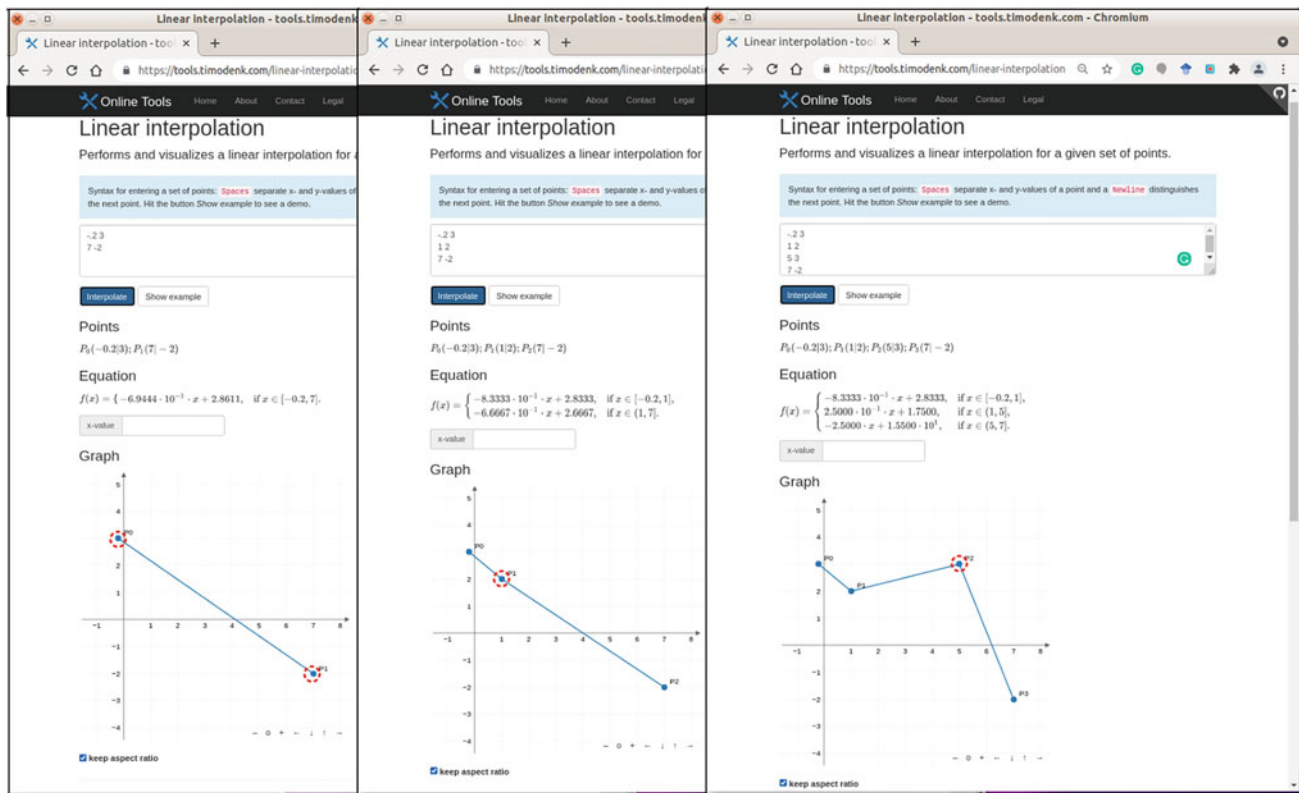
For interpolation of a single variable x , the ordered set of tuples $\mathcal{S}$ is known, such that $\mathcal{S} = \{(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)\}$, where x_i for $i = 0, 1, \dots, n$ is sorted in increasing order, then the interpolating function, or *interpolant*, $f(x)$ is determined, such that $f(x_i) = y_i \forall i \in [0, n]$, and $i \in \mathbb{Z}^+$. This is referred to as *univariate interpolation*. Hereafter, the “data points” (x, y) will be simply referred to as “points” for the sake of readability. Similarly, *multivariate interpolation* deals with interpolation of multiple variables, which can be seen analogous to reconstruction of a continuous function in a multi-dimensional space. These functions are represented in the form of polynomials.

Thorvald Thiele has defined in his book *Interpolationsrechnung* in 1909 that interpolation is the “art of reading between the lines in a numerical table” (Meijering 2002). In science and engineering, mathematically modeling discrete values as a continuous function gives a *tensor field*, which is a generalization of scalar field and vector field in its zeroth and first order, respectively. The discrete points, which are used for determining the interpolant, are obtained through sampling, observation, or experimentation. Interpolation is the process by which the continuous function of a variable or the field is reconstructed from its discrete values. In this regard, sampling based on Nyquist theorem is appropriately needed to reduce reconstruction error. In the scope of reconstruction, there is a salient difference between approximation and interpolation. The approximation theorem states that every continuous function on a closed interval can be approximated uniformly by a polynomial to a predetermined accuracy. However, the approximating function, unlike the interpolating one, need not guarantee that the function value at the discrete points is equal to the known values.

The history of interpolation dates back to seventeenth century in the age of scientific revolution, even though there is evidence of usage of history in texts from ancient Babylon and Greece, the early medieval China and India, and the late medieval Arab, Persia, and India (Meijering 2002). Interpolation theory was further developed in the Western countries by scientists, notably including Copernicus, Kepler, Galileo, Newton, and Lagrange, to promote their work in astronomy and physics.

Univariate Interpolation

The simplest forms of interpolation are implemented using univariate interpolants. There are several widely used interpolants – linear, piecewise linear, polynomial, trigonometric interpolants, etc. Linear interpolation is by far the most popular one used across different domains as it has the best tradeoff between accuracy and efficiency. The criteria for picking an interpolant for a specific computational application



Interpolation, Fig. 1 Piecewise linear interpolation generated by adding points progressively. The newly added points are annotated by the red circles on the visualization of linear interpolation. (Image source:

The screenshots of visualization of the interpolation were generated from an online browser-based application, Online Tools, developed by Timo Denk, <https://tools.timodenk.com>)

depend on both the required accuracy and its efficiency, measured using the computation time. For example, computer graphics applications, including geovisualization, use linear interpolation owing to two reasons. Firstly, the prescribed accuracy for the color mapping, i.e., mapping colors to real values, is limited by the human perception of color. Secondly, the expected computational efficiency is in creating interactive graphical user interface (GUI) applications. Linear interpolation entails finding the line equation for given two points, described by the two tuples of (x, y) values. For two discrete points $(x_1, f(x_1))$ and $(x_2, f(x_2))$, the linear interpolant is given by $f(x) = \frac{f(x_2) - f(x_1)}{x_2 - x_1} \cdot (x - x_1) + f(x_1)$.

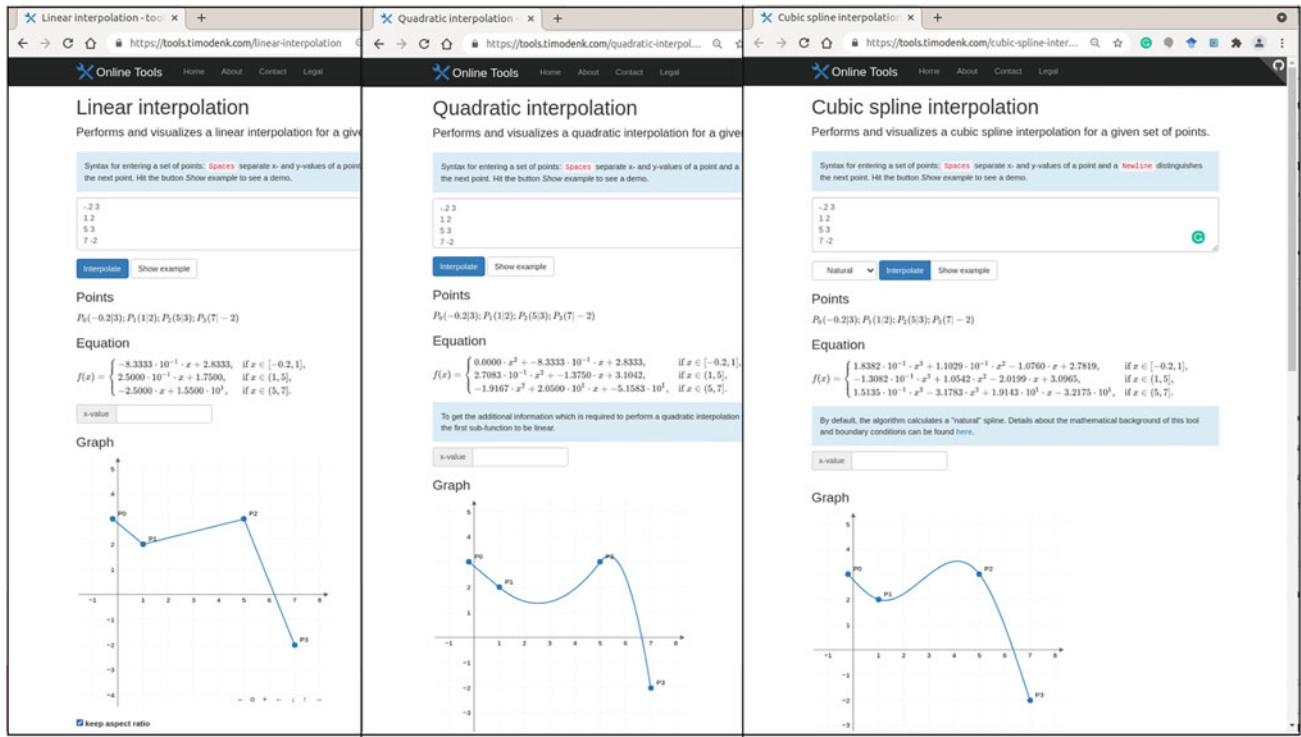
In order to reduce the reconstruction error, linear interpolation can be extended as *piecewise linear interpolation*, which introduces linear interpolants between two consecutive tuples in the ordered list of the tuples, sorted by their x -value. Figure 1 shows how the piecewise linear interpolation can be progressively generated. The degree of the interpolant can be increased by including more points to generate each interpolant, as shown in Fig. 2. In such a scenario, each interpolant between a curve is substituted by an interpolant of a higher degree using derivatives in the case of quadratic function and knots in a spline function in the case of a cubic spline function.

Extending the examples so far, each segment in the piecewise interpolants can include more than two points, thus using n points in each segment to have a polynomial interpolant of degree $(n - 1)$, at the most. The key aspect of piecewise interpolants is that the points at the end of the segment guarantee up to C^0 continuity, i.e., continuity by the function value or the zeroth derivative.

The interpolant type and application revolve around the mathematical function that gives the function values at the sample points. Apart from polynomials, the interpolants can be exponential, logarithmic, trigonometric, etc. There are also specific interpolation methods through the use of Gaussian processes, such as kriging, which has several variants of its own, e.g., simple, ordinary, and regression kriging (see the chapter on “Kriging” for more details).

Multivariate Interpolation

Now, interpolation can be extended from a single variable to multiple ones. Multivariate interpolation leads to higher dimensional interpolation, e.g., bilinear, and trilinear interpolants. These extensions of linear interpolants involve independent interpolation across different dimensions. This is



Interpolation, Fig. 2 Increasing the degree of interpolation in piecewise functions for the same set of points used in piecewise linear interpolation given in Fig. 1, going from linear to quadratic to cubic,

done by adding more points that are interpolated in one dimension, but used as sample points in another dimension, with the constraint that all dimensions are used for interpolation only once.

For two-dimensional (2D) and three-dimensional (3D) spatial applications, it is important to discuss bivariate and trivariate interpolation, respectively. As already discussed, interpolation involves the identification of the known values that would generate the interpolant. These known values in all applications are essentially immediate or local neighbors. Finding neighborhoods is straightforward in gridded and structured datasets as the indices, which are used to search and retrieve a point, are also helpful in finding the neighbors. For instance, in a 3D lattice structure, a grid or lattice point $P(i, j, k)$ with indices along x , y , and z axes, respectively, has 26 local neighbors whose indices are constrained within $[i - 1, i + 1]$, $[j - 1, j + 1]$, and $[k - 1, k + 1]$ simultaneously.

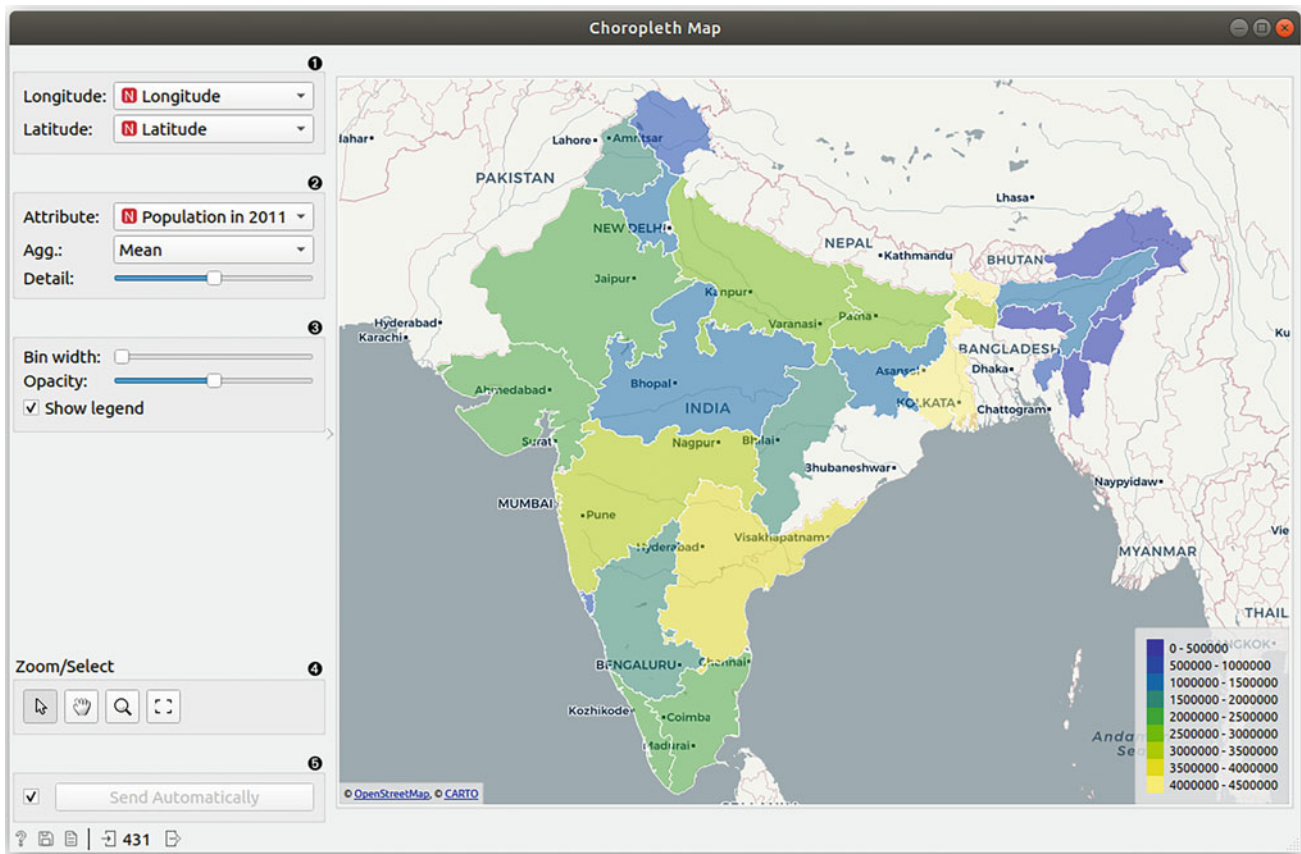
However, in the case of scattered (2D or 3D) points, there are several methods that could be used once neighbors are identified. Identifying neighbors is conventionally based on sorting the distance between the concerned point and the plausible neighboring points. Several constraints can be applied as neighborhood search criteria, e.g., spherical radius and k -nearest neighbors. Since the neighborhood is determined using the distance measure, which itself could be the

from left to right. (Image source: The screenshots of visualization of the interpolation were generated from an online browser-based application, Online Tools, developed by Timo Denk, <https://tools.timodenk.com>)

conventionally used Euclidean, or Hamming, Chebyshev, etc., special interpolants are used for scattered points. These include Shepard's method [inverse distance weighting (IDW)] (Shepard 1968), Sibson's interpolation (natural neighbor interpolation) (Sibson 1981), and radial basis function interpolation (Hardy 1971). These interpolants have a global as well as a local variant. In the global variant, all points are considered, whereas only local neighbors are considered in the local variant of the interpolant. The global or the local extent of the points considered is selected for computing weights for the known values. Consider the case of IDW to understand the weighting process better, where the interpolant $f(\mathbf{x})$ is computed using n selected points, such that, $\mathbf{x}, \mathbf{x}_i \in \mathbb{R}^d$, i.e., d -dimensional space, and $f(\mathbf{x}_i)$ are the known values. Then, with distance between $\mathbf{x}$ and $\mathbf{y}$ given as $d(\mathbf{x}, \mathbf{y})$ and the distance-decay parameter $p \in \mathbb{Z}^+$,

$$f(\mathbf{x}) = \begin{cases} f(\mathbf{x}_k), & \text{if } d(\mathbf{x}, \mathbf{x}_k) = 0, \text{ where } i = k, \\ \frac{\sum_{i=1}^n w_i(\mathbf{x}) f(\mathbf{x}_i)}{\sum_{i=1}^n w_i(\mathbf{x})}, & \\ \text{else } d(\mathbf{x}, \mathbf{x}_i) \neq 0, \forall i \text{ and } w_i(\mathbf{x}) = (d(\mathbf{x}, \mathbf{x}_i))^{(-p)} \end{cases}$$

In a variant of IDW, the distance-decay parameter has been adaptively computed using the point pattern of the local neighborhood (Lu and Wong 2008). IDW and its adaptive



Interpolation, Fig. 3 An example of a choropleth map, as generated of India, using the Orange (anaconda) Python library, an open-source data visualization and mining software. (Image courtesy: <https://orangedatamining.com/widget-catalog/geo/choroplethmap/>)

variants have been routinely used for 3D geological modeling (Liu et al. 2021).

Applications

There are several uses of interpolation, of which data visualization has several applications. Contour extraction is a visualization technique that entails the use of linear interpolation and its higher-dimensional variants in gridded datasets for finding contour lines in 2D and isosurfaces in 3D space. Contour lines are computed using Marching Squares and isosurfaces using the Marching Cubes method (Lorensen and Cline 1987), in 2D and 3D datasets, respectively. Contour lines or isosurfaces are defined as a locus of points all having the same function value. For example, contour tracking has been effectively used for superresolution border segmentation of remotely sensed images for geomorphologic studies (Cipolletti et al. 2012).

Linear interpolation is also widely used in several visualization techniques such as streamline computation, direct volume rendering, graphical rendering techniques with shading models, etc. In visualization, linear interpolation is used in

color mapping that has varied applications, such as choropleth maps, i.e., cartographic maps with colors in regions within geopolitical or other boundaries. As shown in Fig. 3, the color legend in the choropleth map indicates how color is linearly interpolated based on the known values and discretized for the numerical bins of the values.

The choice of interpolation techniques is based on performance – efficiency and accuracy. Reconstruction errors are to be studied for the latter. Apart from these metrics, the level of continuity that is needed from the interpolant, e.g., C^0 , C^1 continuity, is determined by the application, and it governs the choice of the interpolant.

Future Scope

Interpolation has been ubiquitous for mathematical modeling required for a variety of applications in geoscience. It has found its place as a key step in modern data science workflows, such as deep learning. For instance, residual learning networks (ResNets) have been used for seismic data interpolation (Wang et al. 2019).

Cross-References

- [Data Visualization](#)
- [K-nearest Neighbors](#)
- [Kriging](#)
- [Neural Networks](#)

Bibliography

- Cipolletti MP, Delrieux CA, Perillo GM, Piccolo MC (2012) Super-resolution border segmentation and measurement in remote sensing images. *Comput Geosci* 40:87–96
- Hardy RL (1971) Multiquadric equations of topography and other irregular surfaces. *J Geophys Res* 76(8):1905–1915
- Liu Z, Zhang Z, Zhou C, Ming W, Du Z (2021) An adaptive inverse-distance weighting interpolation method considering spatial differentiation in 3D geological modeling. *Geosciences* 11(2):51
- Lorensen WE, Cline HE (1987) Marching cubes: a high resolution 3D surface construction algorithm. *ACM SIGGRAPH Comp Graph* 21(4):163–169
- Lu GY, Wong DW (2008) An adaptive inverse-distance weighting spatial interpolation technique. *Comput Geosci* 34(9):1044–1055
- Meijering E (2002) A chronology of interpolation: from ancient astronomy to modern signal and image processing. *Proc IEEE* 90(3):319–342
- Shepard D (1968) A two-dimensional interpolation function for irregularly-spaced data. In: *Proceedings of the 1968 23rd ACM national conference*, pp 517–524
- Sibson R (1981) A brief description of natural neighbour interpolation (Chapter 2). In: *Interpreting multivariate data*, pp 21–36
- Wang B, Zhang N, Lu W, Wang J (2019) Deep-learning-based seismic data interpolation: a preliminary result. *Geophysics* 84(1):V11–V20

Interquartile Range

Alejandro C. Frery

School of Mathematics and Statistics, Victoria University of Wellington, Wellington, New Zealand

Synonyms

H-spread; Middle 50%; Midsread

Definition

The interquartile range is a measure of dispersion or scale based on order statistics.

Consider the univariate sample $\mathbf{x} = (x_1, x_2, \dots, x_n)$. For simplicity, assume that the sample size n is such that $n/4$ is an integer. Denote $\tilde{\mathbf{x}} = (x_{1:n}, x_{2:n}, \dots, x_{n:n})$ the sample $\mathbf{x}$ sorted

in nondecreasing order, i.e., $x_{1:n} \leq x_{2:n} \leq \dots \leq x_{n:n}$ are the order statistics of the sample. The lower quartile is $q_{1/4}(\mathbf{x}) = x_{\frac{n}{4}:n}$, the upper quartile is $q_{3/4}(\mathbf{x}) = x_{\frac{3n}{4}:n}$, and the interquartile range is $\text{IQR}(\mathbf{x}) = q_{3/4}(\mathbf{x}) - q_{1/4}(\mathbf{x})$. Sample sizes for which $n/4$ is not an integer require a special definition.

Usage and Interpretation

The interquartile range is analogous to the standard deviation in the sense that it measures the dispersion or scale of a sample. While the standard deviation computes the spread around the mean, the interquartile range depends on the sample's order statistics. The interquartile range is, therefore, a robust measure in the sense that a few outlying observations have little or no effect on it. Such robustness justifies the use of the interquartile range in outlier identification procedures.

If $X \sim N(\mu, \sigma^2)$ denotes a random variable with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$, then its interquartile range is $2\sigma\phi^{-1}(3/4)$, where ϕ is its cumulative distribution function. This value can be approximated by 1.349σ or by $27\sigma/20$. With this in mind, the sample standard deviation and $20/27$ times the interquartile range of the sample are comparable, provided the data are outcomes of independent identically distributed random variables $N(\mu, \sigma^2)$ -distributed. Comparing these two quantities is a simple way of checking this last hypothesis.

The interquartile range appears in the boxplot as the distance between the hinges.

Definitions for All Cases of $n \geq 4$

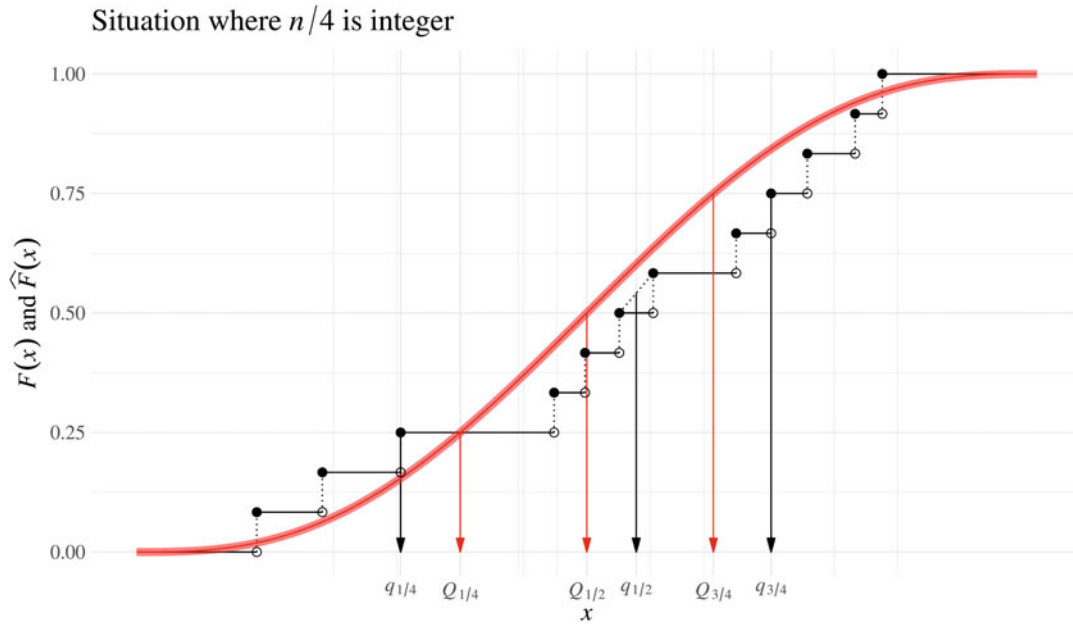
Hyndman and Fan (1996) provide a detailed discussion about the ways in which order statistics can be computed. These definitions are different estimators of the quantiles of a distribution.

The (population or distributional) quantile $\alpha \in (0, 1)$ of a distribution is

$$Q_\alpha = F^-(\alpha) = \inf_{x \in \mathbb{R}} \{F(x) \geq \alpha\},$$

where F is the cumulative distribution function that characterizes the distribution of the random variable X . This definition encompasses all types of distributions. Three values of α render well-known names: $Q_{1/4}$ is the lower or first quartile, $Q_{1/2}$ is the median or second quartile, and $Q_{3/4}$ is the upper or third quartile.

The thick red line in Fig. 1 shows the cumulative distribution function of a beta random variable with both shape parameters equal to 3, i.e.,



Interquartile Range, Fig. 1 Populational and sample quantiles

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{x^3}{3} - \frac{x^4}{2} + \frac{x^5}{5} & \text{if } 0 \leq x \leq 1, \\ 1 & \text{if } x > 1. \end{cases} \quad (1)$$

The population (or “distributional”) first quartile, the median, and the third quartile are, respectively, $Q_{1/4} = 0.359436$, $Q_{1/2} = 1/2$, and $Q_{3/4} = 0.640564$ (the first and last values were obtained numerically). The red arrows in Fig. 1 show how they stem as the inverse of F .

Figure 1 also shows a classical estimator of F based on the sample $\mathbf{x} = (x_1, x_2, \dots, x_n)$, namely, the empirical cumulative distribution function $\hat{F}(\mathbf{x})$:

$$\hat{F}(t) = \frac{1}{n} \# \{ \ell : x_\ell \leq t \}, \text{ for every } t \in \mathbb{R},$$

where we omitted the dependence of $\hat{F}$ on $\mathbf{x}$, and $\#A$ denotes the cardinality of the set A . This function appears as “stairs” of intervals closed to the left and open to the right, and height $1/n$. This function, or any other estimator of F , is the basis for computing sample quantiles of arbitrary order.

We now take a sample of size $n = 12$ from the Beta distribution characterized by the cumulative distribution function given in (1). This particular sample $\mathbf{x}$ is such that the first and third quartiles are $q_{1/4} = x_{3:12} = 0.29$, and $q_{3/4} = x_{9:12} = 0.70$ (after rounding). Black arrows show their position, and we notice that $q_{1/4} < Q_{1/4}$, and that $q_{3/4} > Q_{3/4}$.

In this case, we computed the sample median as the midpoint between $x_{6:n}^q$ and $x_{7:n}^q$. After rounding, its value is $q_{1/2} = 0.55$ and, therefore, $q_{1/2} > Q_{1/2}$. Similar approaches are taken to

compute the first and fourth quartiles of samples of sizes n such that $n/4 \notin \mathbb{N}$.

Hyndman and Fan (1996) define nine ways of computing $q_\alpha(\mathbf{x})$, for $\alpha \in (0, 1)$. All of these approaches are a weighted average of consecutive order statistics:

$$^{(i)}q_\alpha(\mathbf{x}) = (1 - ^{(i)}\gamma)x_{j:n} + ^{(i)}\gamma x_{j+1:n}, \quad (2)$$

where $i = 1, 2, \dots, 9$ is the method, $\frac{i-m}{n} \leq \alpha \leq \frac{i-m+1}{n}$, and $^{(i)}\gamma$ is a function of $j = \lfloor an + m \rfloor$ and $g = an + m - j$.

Any method should satisfy the following properties:

P1) $^{(i)}q_\alpha(\mathbf{x})$ is continuous on $\alpha \in (0, 1)$.

P2) The usual definition of median is respected:

$$^{(i)}q_{1/2}(\mathbf{x}) = \begin{cases} x_{\frac{n+1}{2}:n} & \text{if } n \text{ is odd,} \\ \frac{x_{\frac{n}{2}:n}^q + x_{\frac{n}{2}+1:n}^q}{2} & \text{if } n \text{ is even.} \end{cases}$$

P3) The tails of the sample are treated equally:

$$\# \{ \ell : x_\ell \leq ^{(i)}q_\alpha(\mathbf{x}) \} = \# \{ \ell : x_\ell \geq ^{(i)}q_{1-\alpha}(\mathbf{x}) \}.$$

P4) The sample and distributional quantiles are analogous, i.e.,

$$\# \{ \ell : x_\ell \leq ^{(i)}q_\alpha(\mathbf{x}) \} \geq \alpha n.$$

P5) Where the inverse $^{(i)}q^{-1}(x)$ is uniquely defined, holds that

- a) ${}^{(i)}q^{-1}(x_{1:n}) > 0$ and that ${}^{(i)}q^{-1}(x_{n:n}) < 1$,
 b) ${}^{(i)}q^{-1}(x_{k:n}) + {}^{(i)}q^{-1}(x_{n-k+1:n}) = 1$.

The `quantile` function provided by the R platform (R Core Team 2020) implements the nine alternatives discussed by Hyndman and Fan (1996), being $i = 7$ the one that operates unless another one is requested. It is noteworthy that only one of these options satisfies all the properties stated above, namely, $i = 5$. Nevertheless, the recommended one is $i = 8$ because of its approximate unbiasedness.

Summary

The Interquartile range is a measure of scale or spread based on order statistics. As there are many ways of computing such quantities, it is recommended to state which version was used in practice.

Cross-References

► [Standard Deviation](#)

Bibliography

- Hyndman RJ, Fan Y (1996) Sample quantiles in statistical packages. *Am Stat* 50(4):361. <https://doi.org/10.2307/2684934>
 R Core Team (2020) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. URL <https://www.R-project.org/>

Inverse Distance Weight

Narayan Panigrahi
 Scientist and Head of GIS Division, Center for Artificial Intelligence and Robotics (CAIR), Bangalore, India

Definition

This entry discusses one of the widely used spatial Interpolation method, the IDW (inverse distance weight) and its variants. The entry starts with the genesis and historic development of IDW, its definition, and its mathematical basis. The factors governing the IDW are discussed. The selections of the sample values depending on the neighborhood policy are discussed. Also discussed are some interesting mathematical variations of the method which makes it useful in various different phenomena with varying spatial distribution of sample locations. Finally the variants of the

method in 3D and its usage are discussed. The chapter ends with some important recommendations and scenario so as which variant to use in which application.

Genesis and Historic Development of IDW

Inverse Distance Weighting (IDW) is a type of deterministic method for multivariate spatial interpolation technique to interpolate the value at an location from known sampled locations scattered around the point of interest. The assigned values to unknown points are computed with a weighted average of the values available at the surrounding known locations.

IDW is a multivariate spatial interpolation technique which is easy to implement through computer programming. Computing the magnitude of missing spatial data of a location from among the given measured data of the phenomena such as height, pollution density, snow precipitation, and temperature is easier using IDW. It mimics the natural phenomena more closely in comparison to other spatial interpolation techniques.

The name IDW is because the methods use the weighted average of distance from the point of unknown data to data point where the value of the sample is known which is applied to compute the value of unknown value. Since it resorts to the inverse of the distance to each known point (“amount of proximity”) when assigning weights it is known as inverse distance weight.

The method IDW owes its genesis at the Harvard Laboratory for Computer Graphics and Spatial Analysis. At the beginning of 1965, a collection of scientists from varied domains converged to rethink, among other things, what we now call Geographic Information Systems (GIS) (Narayan 2014; Alcaraz et al. 2016). This workshop was led by Howard Fisher to discuss improved computer mapping program for GIS which lead to development of SYMAP. Fisher wanted an improved spatial interpolation technique to augment his work on SYMAP. During a brain storming session one freshman, Donald Shepard (Shepard 1968), decided to overhaul the interpolation techniques in SYMAP. This resulted in formulation and implementation of the basic inverse distance weighting (IDW). The first scientific article on IDW which is known as Sheppard’s algorithm is published in his famous article from 1968 (Shepard 1968; Fisher et al. 1987).

Definition and Mathematical Expression

Inverse distance weighting (IDW) is a type of deterministic method for multivariate interpolation with a known scattered

set of points (Łukaszuk 2004). The assigned values to unknown points are calculated with a **weighted** average of the values available at the known points.

The mathematical equation of IDW of a phenomena “Z” whose sample magnitude at different known locations Z_i are collected prior. One has to compute the magnitude of the phenomena at a location “ Z_0 ” using the IDW formula given in equation 1.

$$Z_0 = \frac{\sum_{i=1}^n Z_i \frac{1}{D_i^p}}{\sum_{i=1}^n \frac{1}{D_i^p}} \quad (1)$$

where Z_i is magnitude of samples measured at known “n” locations ($i = 1, \dots, n$)

Z_0 is the unknown magnitude of the phenomena to be computed.

D_i^p is the p th power of the distance computed between the point of unknown sample magnitude with that of the known sample points.

Factors Governing IDW

Inverse distance weighting is the simplest spatial interpolation method (George and David 2008; Ozelkan et al. 2015). A neighborhood about the interpolated point is identified and a weighted average is taken of the observation values within this neighborhood. The weights are a decreasing function of distance. The user has the control over the mathematical form of the weighting function, the size of the neighborhood (expressed as a radius or a number of points), in addition to other options.

Weighting Function

The simplest weighting function in the IDW is inverse power “ $w(d) = 1/d^p$ ”. The value of p is specified by the user. The most common choice is $p = 2$. For $p = 1$, the interpolated function is “cone-like” in the vicinity of the data points, where it is not differentiable.

The power factor “ p ” decides the degree of influence of the phenomena. If ($p = 1$), then it is a simple inverse distance weight interpolation. If ($p = 2$), then it is inverse distance square weight. The value of “ p ” can be any value ≥ 1 depending on the model under consideration.

Shepard’s method (Shepard, 1968) is a variation on inverse power, with two different weighting functions using two separate neighborhoods. The default weighting function for Shepard’s method is an exponent of 2 in the inner

neighborhood and an exponent of 4 in the outer neighborhood. A generic mathematical representation of the Shepard’s method is given in equation 2. The form of the outer function is modified to preserve continuity at the boundary of the two neighborhoods.

$$Z_0 = \frac{\sum_{i=1}^{n-k} Z_i \frac{1}{D_i^2}}{\sum_{i=1}^{n-k} \frac{1}{D_i^2}} + \frac{\sum_{i=k}^n Z_i \frac{1}{D_i^4}}{\sum_{i=k}^n \frac{1}{D_i^4}} \quad (2)$$

Neighborhood Size

The neighborhood size determines how many points are included in the inverse distance weighting. The neighborhood size can be specified in terms of its radius (in km), the number of points, or a combination of the two. If a radius is specified, the user also can specify an override in terms of a minimum and/or maximum number of points. Invoking the override option will expand or contract the circle as needed. If the user specifies the number of points, as an override of a minimum and/or maximum radius can be included. It also is possible to specify an average radius “ R ” based upon a specified number of points. Again, there is an override to expand or contract the neighborhood to include a minimum and/or maximum number of points. In Shepard’s method, there are two circular neighborhoods; the inner neighborhood is taken to be one-third the radius of the outer radius.

Anisotropy Correction

In many instances, the observation points are not uniformly spaced about the interpolation points. Several points are in a particular direction and fewer in direction orthogonal to the dominant direction. This condition is known as anisotropic condition. This situation produces a spatial bias of the estimate, as the clustered points carry an artificially large weight. The anisotropy corrector permits the weighted average to downweight clustered points that are providing redundant information. The user selects this option by setting the anisotropy factor to a positive value. A value of 1 produces its full effects, while a value of 0 produces no correction. This correction factor is defined by computing the angle between every pair of observation points in the neighborhood, relative to the observation point.

Gradient Correction

A disadvantage of the inverse weighted distance functions is that the function is forced to have a maximum or minimum at the data points (or on a boundary of the study region). The gradient corrector permits a nonzero gradient at the observation points. The implementation is such that the interpolated value is the sum of two values.

One of the key governing factor of equation 1 is the value of number of samples “n” and their spatial alignment around the location of the point of interest. Many geometric approaches are in practice for selection of data points. In most of the software implementations a circular filter with a radius “R” around the point of interest has been in use to select the location of sample values. Some other implementations use a rectangular or square filter around the point of interest. If the values of the phenomena have to be computed for gridded locations from the value of sampled scatter points such as generation of a DEM (Digital Elevation Model) from a scatter point of surveyed heights, then a moving circular disk or rectangle is considered for computing the values for uniformly spaced grid locations.

How Inverse Distance Weighted Interpolation Works?

Inverse distance weighted (IDW) interpolation explicitly makes the assumption that things that are close to one another are more alike than those that are farther apart. To predict a value for any unmeasured location, IDW uses the measured values surrounding the prediction location. The measured values closest to the prediction location have more influence on the predicted value than those farther away. IDW assumes that each measured point has a local influence that diminishes with distance. It gives greater weights to points closest to the prediction location, and the weights diminish as a function of distance, hence the name inverse distance weighted. Weights assigned to data points are illustrated in the example 1.

The geometric shape and dimension of the search neighbor is an important factor in effectiveness of IDW. It is decided depending upon the spatial distribution of the sampled data locations and the characteristic of the sampled data. Some of the frequently used geometric search patterns employed in IDW are: (a) radial search with mean radius “R”, (b) spherical geometry with mean radius “R” for scattered data in 3D, (c) rectangular geometry with (Length(L) x Width(W)), (d) elliptic geometry with (semi major axis(a) and semi minor axis (b)), and (e) disk-like structure with inner radius “r” and outer radius “R” as depicted in Fig 1 a, b, c, d.

The geometric structure and its dimension decide the search neighbor for inclusion and exclusion of the values of sampled points for participating in computation of unknown value using IDW.

Once search neighbor is decided the next step for software implementation of IDW is to answer the question whether location (x_i, y_i) from among the sampled locations is within the sample neighbor? This question can be answered using computational geometric queries such as the following:

- (a) Point_inside_circle(X_i, Y_i, R)
- (b) Point_inside_ellipse(X_i, Y_i, a, b)
- (c) Point_inside_rectangle(X_i, Y_i, L, W)
- (d) Point_inside_Sphere (X_i, Y_i, R)

If the Point Inside Neighbor geometry returns TRUE, then the corresponding value V_i of (X_i, Y_i) is included for computation using IDW else it is discarded.

Further if the spatial extend of the physical phenomena is spread over a large extend marked by a bounding rectangle ($L \times W$), then the process of computing the values at different locations from the scattered data is obtained by repeatedly computing using the moving neighbor operator for the entire spatial domain.

To consolidate the concept of IDW, let us analyze a real-life example. One need to compute the depth of snow cover at a high slope and hazard location from among the values of snow precipitation surveyed from a set of known locations surrounding the location under question. This scenario is depicted in the Fig. 2. The field observations of five locations surrounding the point and their distance data is tabulated (Table 1). Using IDW the value of precipitation at the location in question is computed using equation 3. Also how the variation in the computed value is manifesting when computed using inverse square distance weight (ISDW) is analyzed in equation 4. One can see the variation in the computed value by IDW and ISDW. The variation is very small implicating the effectiveness and the resilience of the method.

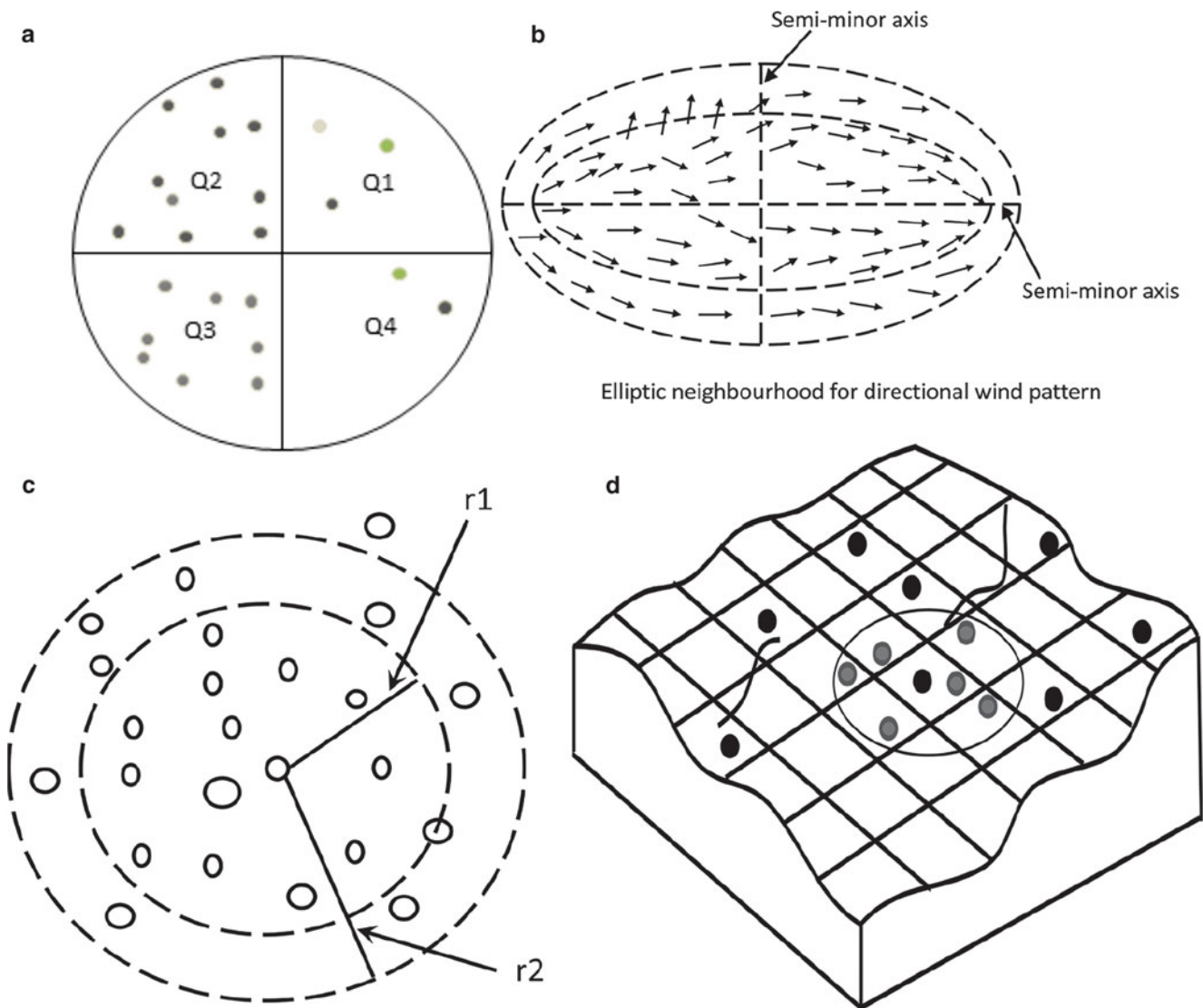
IDW and ISDW are applied to above data separately to compute the precipitation of the snow in the location as depicted in Fig. 1.

$$34\left(\frac{1}{5}\right) + 42\left(\frac{1}{8}\right) + 17\left(\frac{1}{14}\right) + 52\left(\frac{1}{12}\right) + 27\left(\frac{1}{9}\right) = 20.597589 \quad (3)$$

$$34\left(\frac{1}{5^2}\right) + 42\left(\frac{1}{8^2}\right) + 17\left(\frac{1}{14^2}\right) + 52\left(\frac{1}{12^2}\right) + 27\left(\frac{1}{9^2}\right) = 2.797387 \quad (4)$$

The values computed at the location using IDW and ISDW are 26.0439 and 22.35799, respectively.

Therefore it is highly encouraged to use IDW or ISDW method judiciously to compute an estimate of the phenomena. Often the choice of the type and dimension of the search neighbor for filtering the spatial locations which can participate in computation of the unknown value using IDW is governed by many factors:



Inverse Distance Weight, Fig. 1 (a) Sampled locations within a circular neighbor, (b) elliptic neighbor for flow like phenomena, (c) circular disk neighbor, (d) spherical neighbor with undulated terrain

- (a) Whether the phenomena is random and scattered spatially.
- (b) Whether the phenomenon is directional like flow of air, water, etc.
- (c) Whether the phenomena are static or influenced more by local factors such as seizure from brain.

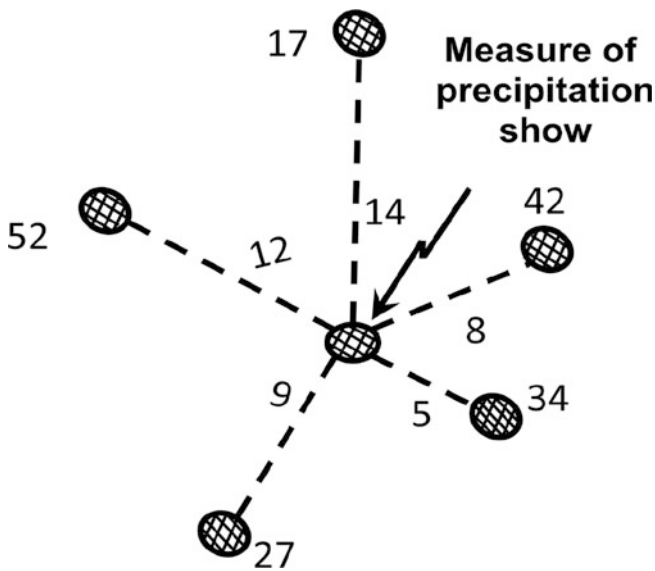
Depending on the type of physical phenomena and its spatial behavior the neighborhood geometry is selected.

- For random static phenomena the neighborhood geometry is circular with radius “R.”
- For air flow or fluid flow kind of phenomena it is recommended to use the elliptic structure.

- For chemical concentration or computation of value of precipitation of rain or snow circular neighborhood geometry should be more suitable.
- For computing the pressure and fault line in mines, slope and aspect of undulated terrain and geological exploration sphere like structure is more preferable.
- To find the exact location of brain from where seizure like EEG signals are emanating, spherical neighborhood geometry is more suitable.

The Behavior of Power Function

As mentioned above, weights are proportional to the inverse of the distance (between the data point and the location of prediction) raised to the power value p . As a result, as the



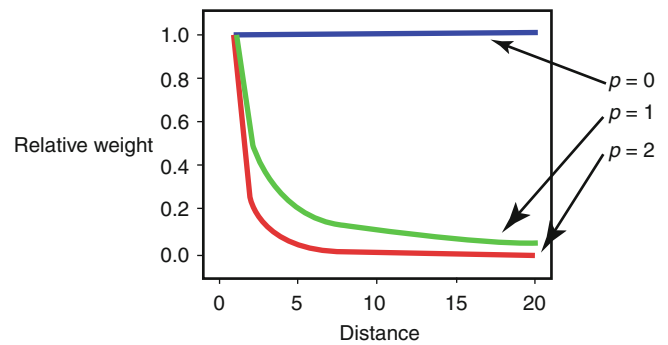
Inverse Distance Weight, Fig. 2 Snow precipitation measured at different Locations

Inverse Distance Weight, Table 1 Precipitation measured in known locations

Precipitation of snow in mm	Distance from the point of unknown value	Inverse of distance $1/D$	Square of inverse of distance $1/D^2$
34	5	$1/5 = 0.2$	$1/5^2 = 0.04$
42	8	$1/8 = 0.125$	$1/8^2 = 0.015625$
17	14	$1/14 = 0.071428$	$1/14^2 = 0.005102$
52	12	$1/12 = 0.083333$	$1/12^2 = 0.006944$
27	9	$1/9 = 0.111111$	$1/9^2 = 0.012345$

distance increases, the weights decrease rapidly. The rate at which the weight decrease is dependent on the value of “ p .” If $p = 0$, there is no decrease with distance, and because each weight D_i is the same, the prediction will be the mean of all the data values in the search neighborhood and IDW acts like an moving average interpolator. As p increases, the weights for distant points decrease rapidly. If the p value is very high, only the immediate surrounding points will influence the prediction. The nature of the weight becomes asymptotic with increase distance making the impact of the values of faraway locations less significant in the overall computation. This phenomena is depicted in the plot (Fig. 3).

In most of the geo-statistical analysis the value of p is chose to be 2, although there is no theoretical justification to prefer this value over others, and the effect of changing p should be investigated by previewing the output and examining the cross-validation statistics.



Inverse Distance Weight, Fig. 3 Relative weight of the known values decrease rapidly with increase of value of power factor “ p .” This is illustrated in the plot weight-distance with varying “ p ”

The Selection of Search Neighborhood

The philosophy of IDW modeling is that “things that are close to one another are more alike and related than those that are farther away.” As the locations get farther away, the measured values will have little relationship to the value of the prediction location. Also to speed us computations, one need to select a fixed set of known values from spatially close range and exclude the more distant points that will have little influence on the prediction. As a result, it is common practice to limit the number of measured values by specifying a search neighborhood. The shape of the neighborhood restricts how far and where to look for the measured values to be used in the prediction. Other neighborhood parameters restrict the locations that will be used within that shape. In the following image, five measured points (neighbors) will be used when predicting a value for the location without a measurement, the yellow point.

The shape of the neighborhood is influenced by the input data and the surface one is trying to create from the data. If there are no directional influences in the data, and observation points are equally distributed in all the directions then simple circular or spherical structure is preferred depending on the dimension of the data points. However, if there is a directional influence in the data, such as a prevailing wind, one may want to adjust for it by changing the shape of the search neighborhood to an ellipse with the major axis parallel to the direction of the flow of the wind. The adjustment for this directional influence is justified because one know that locations upwind from a prediction location are going to be more similar at remote distances than locations that are perpendicular to the wind but located closer to the prediction location.

Once a neighborhood shape has been specified, you can restrict which data locations within the shape should be used. One can define the maximum and minimum number of locations to use; also one can divide the neighborhood into sectors. If one divides the neighborhood into sectors, the maximum and minimum constraints will be applied to each sector.

The output surface generated using IDW is sensitive to clustering and the presence of outliers. IDW assumes that the phenomenon being modeled is driven by local variation, which can be captured (modeled) by defining an adequate search neighborhood. Since IDW does not provide prediction standard errors, justifying the use of this model may be problematic.

Applications of IDW

Given a set of known values such as elevation, rain in mm, pollution density or noise, there are different spatial interpolation techniques available to compute the value at unsampled location. IDW, Kriging, Voronoi tessellation, and moving average are some of the alternate tool. To estimate at points where you do not know its value, IDW uses spatial autocorrelation in its math. Closer values have more effect while farther away ones have less effect.

Application Considering Spatial Autocorrelation in 3D

Spatial interpolation is a main research method in 3D geological modeling, which has important impacts on 3D geological structure model accuracy. The inverse distance weighting (IDW) method is one of the most commonly used deterministic models, and its calculation accuracy is affected by two parameters: search radius and inverse-distance weight power value. Nevertheless, these two parameters are usually set by humans without scientific basis. To this end, we introduced the concept of “correlation distance” to analyze the correlations between geological borehole elevation values, and calculate the correlation distances for each stratum elevation. The correlation coefficient was defined as the ratio of correlation distance to the sampling interval, which determined the estimation point interpolation neighborhood. We analyzed the distance-decay relationship in the interpolation neighborhood to correct the weight power value. Sampling points were selected to validate the calculations. We concluded that each estimation point should be given a different weight power value according to the distribution of sampling points in the interpolated neighborhood. When the spatial variability was large, the improved IDW method performed better than the general IDW and ordinary kriging methods.

Spatial interpolation is a main research method in 3D geological modeling, which has important impacts on 3D geological structure model accuracy. The inverse distance weighting (IDW) method is one of the most commonly used deterministic models, and its calculation accuracy is affected by two parameters: search radius and inverse-distance weight power value. Nevertheless, these two parameters are usually set by humans without scientific basis. To this end, we introduced the concept of “correlation distance” to analyze the correlations between geological borehole elevation values,

and calculate the correlation distances for each stratum elevation. The correlation coefficient was defined as the ratio of correlation distance to the sampling interval, which determined the estimation point interpolation neighborhood. We analyzed the distance-decay relationship in the interpolation neighborhood to correct the weight power value. Sampling points were selected to validate the calculations. We concluded that each estimation point should be given a different weight power value according to the distribution of sampling points in the interpolated neighborhood. When the spatial variability was large, the improved IDW method performed better than the general IDW and ordinary kriging methods.

Syntax or Signature of API Implementing IDW

```
IDW(Surveyed_Data_Set_with_Location, Location_of_Unknown_Point, power_of_IDW, Dimension, z_field, search_neighborhood_geometry, cell_size, Output_Data_set)
```

To execute IDW

```
IDW(inPointFeatures, (Xi,Yi), zField, 2, 2, circular, 4, outDataSet)
```

Summary

IDW uses the measured values surrounding the prediction location to compute a value for any unsampled location, based on the assumption that things that are close to one another are more alike than those that are farther apart. It selects the neighborhood points from among the domain using a neighborhood policy which defines the geometry and search distance, or number of closest points. The power of the inverse distance decides the degree of influence of the phenomena. The search geometry and the search range are some of the decision one need to take before applying IDW. One need to choose higher power settings for more localized peaks and troughs, the directional behavior of the physical phenomena often plays a crucial role in choosing the shape of the filter, upper or lower bound of sample values to compute. Some of the key features of IDW are

- The predicted value is limited to the range of the values used to interpolate. Because IDW is a weighted distance average, the average cannot be greater than the highest or less than the lowest input. Therefore, it cannot create ridges or valleys if these extremes have not already been sampled.
- IDW can produce a bull's-eye effect around data locations.
- Unlike other interpolation methods such as Kriging, IDW does not make explicit assumptions about the statistical properties of the input data. IDW is often used when the

input data does not meet the statistical assumptions of more advanced interpolation methods.

- This method is well-suited to be used with very large input datasets.

The Inverse Distance Weighting interpolation method is as flexible and simple to program. But it is often the case that other interpolation techniques like kriging can help obtain a more robust model.

Bibliography

- Alcaraz M, Vazquez-Sune E, Velasco V, Diviu M (2016) 3D GIS-based visualisation of geological, hydrogeological, hydrogeochemical and geothermal models. *Zeitschrift Der Deutschen Gesellschaft Fur Geowissenschaften* 167(4):377–388
- Fisher NI, Lewis T, Embleton BJJ (1987) Statistical analysis of spherical data. Cambridge University Press, p 329
- George YL, David WW (2008) An adaptive inverse-distance weighting spatial interpolation technique. *Comput Geosci* 34:1044–1055
- Lukaszczuk S (2004) A new concept of probability metric and its applications in approximation of scattered data sets. *Comput Mech* 33(4): 299–304. <https://doi.org/10.1007/s00466-003-0532-2>
- Narayan P (2014) Computations in geographic information system. CRC Press. ISBN 978-1-4822-2314-9
- Ozelkan E, Bagis S, Ozelkan EC, Ustundag BB, Yucel M, Ormeci C (2015) Spatial interpolation of climatic variables using land surface temperature and modified inverse distance weighting. *Int J Remote Sens* 36(4):1000–1025
- Shepard D (1968) A two-dimensional interpolation function for irregularly-spaced data. In: *Proceedings of 23rd National Conference ACM*. ACM, pp 517–524

Inversion Theory

R. N. Singh¹ and Ajay Manglik²

¹Discipline of Earth Sciences, Indian Institute of Technology, Gandhinagar, Palaj, Gandhinagar, India

²CSIR-National Geophysical Research Institute, Hyderabad, India

Definition

Inverse theory – The mathematical theory for deducing from observations the cause(s) that produce them.

Introduction

Inverse theory in general deals with deducing the cause from observations through intricate mathematical formulations. In geophysics, physical fields are recorded by various sensors located on the earth's surface, within boreholes or above the ground by airborne and satellite observations. These fields are

responses of physical property variations within the earth, and, thus, inversion theory is used to image the earth's structure in terms of physical property variations from the observations of physical fields (Tarantola 1987; Aster et al. 2019). Inverse theory is also used in other branches of geosciences to understand the geological processes, e.g., estimation of magma source composition from inversion of trace elements (Albarede 1983; McKenzie and O'Nions 1991). The subject of inverse theory and its application in geosciences is vast, and a lot of development has taken place in this field. In this entry, we provide a brief description of inverse theory to give a flavor of the subject.

In inversion approach, first a forward problem (► [Forward and Inverse Models](#)) is defined that relates model parameters $\mathbf{m}$ to observed responses $\mathbf{d}$ through a functional f as

$$\mathbf{d} = f(\mathbf{m}). \quad (1)$$

Here, the functional f contains information about the physics of the given problem. Since in geophysical studies, we aim at obtaining the unknown model parameters $\mathbf{m}$ from the observed responses $\mathbf{d}$, in the simplest form, ignoring all complications of noise in data and inadequacy of data, assuming that the exact inverse exists the inverse problem can be posed as

$$\mathbf{m} = f^{-1}(\mathbf{d}). \quad (2)$$

Inverse of a Discrete Problem

In general form, the model parameters are a continuous function of the spatial coordinate system and the functional is a continuous operator. This requires solving Eqs. 1 and 2 in the continuous domain. There are methods such as the Backus-Gilbert method that are used to solve such inverse problems. However, a generally followed practice is to convert the problem into the discrete domain such that $\mathbf{m} = \{m_1, m_2, \dots, m_N\}^T$ and $\mathbf{d} = \{d_1, d_2, \dots, d_M\}^T$, where M, N are number of observations and model parameters, respectively, and then express Eq. 1 in matrix form as

$$\mathbf{d} = \mathbf{F}\mathbf{m}. \quad (3)$$

In the ideal situation of noise-free data and full-rank matrix ($M = N = K$, where K is the rank of the matrix), inverse of Eq. 3 is

$$\mathbf{m} = \mathbf{F}^{-1}\mathbf{d}. \quad (4)$$

This is the desired exact solution of an inverse problem. However, geophysical inverse problems are mostly nonlinear, rank-deficient, and ill-conditioned, and in practice it is difficult to get the exact solution (Menke 1984; Tarantola 1987). Therefore, a generalized inverse of the problem is obtained by

imposing some other constraints. There are two main scenarios; (i) $M > N$, and (ii) $M < N$ assuming that the rank $K = \min(M, N)$. In the first case, the number of observations is more than the number of unknown model parameters. It is called a strictly overdetermined system (Gupta 2011) for which the error of misfit is

$$\mathbf{e} = \mathbf{d} - \mathbf{F}\hat{\mathbf{m}}, \quad (5a)$$

where $\hat{\mathbf{m}}$ is the estimated model parameter vector and is minimized in the least square sense, i.e., $\mathbf{e}^T \mathbf{e}$ is minimized that yields the best fit solution as

$$\hat{\mathbf{m}} = (\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \mathbf{d}. \quad (5b)$$

In the second case, the number of observations is less than the number of unknown model parameters. It is called a strictly under-determined system. Here, the norm of the model parameter vector is minimized by minimizing $\mathbf{m}^T \mathbf{m}$ to yield the best fit solution as

$$\hat{\mathbf{m}} = \mathbf{F}^T (\mathbf{F} \mathbf{F}^T)^{-1} \mathbf{d}. \quad (6a)$$

There is another scenario when the rank of the matrix $K < \min(M, N)$. Inverse of an overdetermined system for such a case is obtained by minimizing an objective function that is a weighted sum of $\mathbf{e}^T \mathbf{e}$ and $\mathbf{m}^T \mathbf{m}$, i.e., minimizing $\mathbf{e}^T \mathbf{e} + \lambda^2 \mathbf{m}^T \mathbf{m}$, which yields

$$\hat{\mathbf{m}} = (\mathbf{F}^T \mathbf{F} + \lambda^2 \mathbf{I})^{-1} \mathbf{F}^T \mathbf{d}. \quad (6b)$$

Here, $\mathbf{I}$ is identity matrix and λ is the control parameter to assign relative weight to the data error and model norm, also known as the Marquardt parameter (Marquardt 1963). For an optimum value of λ , the L-curve is used which is constructed on a plane with axes as $\mathbf{e}^T \mathbf{e}$ and $\mathbf{m}^T \mathbf{m}$ for different values of λ . The value of λ is chosen to correspond to maximum curvature point of this curve. This is a simple case of regularization.

In case of noisy data, if the errors in data are given in terms of variances, then a weight matrix is introduced while minimizing errors as

$$\mathbf{e}^T \mathbf{C}_d^{-1} \mathbf{e}, \quad (7a)$$

where

$$\mathbf{C}_d = \begin{bmatrix} \sigma_{d1}^2 & 0 & \cdots & 0 & 0 \\ 0 & \sigma_{d2}^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & \sigma_{dN}^2 \end{bmatrix}. \quad (7b)$$

In this case the solution of the inverse problem is

$$\hat{\mathbf{m}} = (\mathbf{F}^T \mathbf{C}_d^{-1} \mathbf{F})^{-1} \mathbf{F}^T \mathbf{C}_d^{-1} \mathbf{d}. \quad (8)$$

Singular Value Decomposition

In the inversion approach described above, the generalized inverse is obtained by different matrix operations depending on the nature of the problem and the rank of the matrix. Singular value decomposition (SVD) (Singular Value Decomposition) does not require a priori information about the nature of the matrix and hence is very useful especially for inversion of singular or ill-conditioned matrices (Golub and Kahan 1965; Golub and Reinsch 1970). The technique was discovered independently by Beltrami in 1873 and Jordan in 1874 as a tool to solve square matrices and was extended to rectangular matrices by Eckart and Young in the 1930s (Stewart 1993; Manglik 2021). In SVD, a matrix $\mathbf{F}$ of dimension $M \times N$ and rank K is decomposed into data and parameter eigen matrices $\mathbf{U}(M \times K)$ and $\mathbf{V}(N \times K)$, respectively, and a diagonal matrix containing non-zero eigen values $\Sigma(K \times K)$ arranged in decreasing value of magnitude such that

$$\mathbf{F} = \mathbf{U} \Sigma \mathbf{V}^T. \quad (9)$$

An inverse of this matrix can be obtained as

$$\mathbf{F}^- = \mathbf{V} \Sigma^{-1} \mathbf{U}^T. \quad (10)$$

The data and parameter matrices satisfy the conditions $\mathbf{U}^T \mathbf{U} = \mathbf{V}^T \mathbf{V} = \mathbf{I}$ but $\mathbf{U} \mathbf{U}^T$ and $\mathbf{V} \mathbf{V}^T$ need not be the identity matrix for a geophysical problem. The first one is known as the information density matrix that provides information about the data sensing different model parameters, whereas the second one is called the parameter resolution matrix that is used to understand the resolution of model parameters. This matrix is very useful to find equivalence of model parameters. The SVD approach reduces the drudgery of computation and provides ease in interpretation.

Inverse of a Continuous Problem

Discrete inverse problem assumes that the model parameters are discrete in nature, i.e., the earth's structure is composed of discrete layers/cells each having constant values of physical properties. However, physical properties may vary continuously in spatial domain and observations made at discrete locations are a function of physical property variation in the entire domain. This leads to a situation where observed data are always less compared to the unknown model parameters (under-determined system). Backus and Gilbert (1967, 1968, 1970) developed an elegant approach to invert this continuous problem. This method tries to invert the parameters given by continuous functions. We have in this case

$$\int_0^1 F_i(x)m(x)dx = d_i, \text{ for the domain in the range, } 0 \leq x \leq 1, \quad (11)$$

where $m(x)$ represents continuous model parameter variation and $F_i(x)$ is a function describing the physics of the problem. The subscript i denotes the index of the data point ($i = 1, 2, \dots, N$). Here, we have finite set of data and wish to estimate the unknown function $m(x)$. To do this, we first express m at $x = x_0$ in terms of the Dirac delta function as

$$m(x_0) = \int_0^1 \delta(x - x_0)m(x)dx, \quad (12)$$

and assume that $m(x_0)$ is a linear combination of the data with weights a_i , ($i = 1, 2, \dots, N$), i.e.

$$\sum_{i=1}^N a_i d_i = m(x_0). \quad (13)$$

Substituting Eq. 11 into Eq. 13 gives

$$\sum_{i=1}^N a_i d_i = \sum_{i=1}^N a_i \int_0^1 F_i(x)m(x)dx = \int_0^1 \left\{ \sum_{i=1}^N a_i F_i \right\} m(x)dx. \quad (14)$$

Comparing Eq. 14 with Eq. 12, we get

$$\sum_{i=1}^N a_i F_i = \delta(x - x_0). \quad (15)$$

This is the ideal situation in which we get the exact solution of the model at $x = x_0$. However, this is mostly not possible and weights a_i are obtained such that $\sum_{i=1}^N a_i F_i$ is as close to the

Dirac delta function as possible and we get an estimate $\hat{m}(x_0)$ of the true model $m(x)$ around $x = x_0$. The spread of the function gives information about the resolution of the model. An example of application of the Backus-Gilbert method to magnetotelluric data can be found in Manglik and Moharir (1998).

Bayesian Inversion

In this inversion approach (► [Bayesian Inversion](#)), the model is considered as a random variable unlike the methods

described in previous sections that treat the model as deterministic. The model thus has a probability distribution and we invert for a posterior distribution of the model parameters from the known prior distribution for the model and the data. The posterior distribution of the model according to the Bayes theorem is expressed as

$$f_{M|d} = \frac{f_{D|m}f_M(m)}{f_D(d)}, \quad (16)$$

where $f_M(m)(f_D(d))$ is the probability density function (► [Probability Density Function](#)) for model (data) and $f_{M|d}(f_{D|m})$ is the conditional probability distribution of model given data (or data given model). This can be demonstrated for a normal distribution by considering that the model follows a Gaussian distribution so that the probability $f_M(m)$ is expressed as

$$f_M(m) \propto \exp \left[-\frac{1}{2} \left\{ (m - m_{\text{prior}})^T C_m^{-1} (m - m_{\text{prior}}) \right\} \right], \quad (17)$$

where m_{prior} and C_m are the mean of the prior distribution of the model and its covariance matrix. The conditional distribution of data given model is then

$$f_{D|m} \propto \exp \left[-\frac{1}{2} \left\{ (Fm - d)^T C_d^{-1} (Fm - d) \right\} \right], \quad (18)$$

where d represents observed data with C_d corresponding data covariance matrix. Using Bayes theorem, we can get model posterior conditional probability density function given data as

$$f_{M|d} \propto \exp \left[-\frac{1}{2} \left\{ (m - m_{\text{prior}})^T C_m^{-1} (m - m_{\text{prior}}) + (Fm - d)^T C_d^{-1} (Fm - d) \right\} \right]. \quad (19)$$

Equation 19 can be simplified to (Tarantola 1987)

$$f_{M|d} \propto \exp \left[-\frac{1}{2} (m - m_{\text{MAP}})^T C_{M'}^{-1} (m - m_{\text{MAP}}) \right], \quad (20)$$

where m_{MAP} is the maximum a posteriori model and modified covariance matrix $C_{M'}$ and is given by

$$C_{M'} = (F^T C_d^{-1} F + C_m^{-1})^{-1}. \quad (21)$$

The maximum a posteriori model m_{MAP} is obtained by minimizing the following expression

$$(m - m_{\text{prior}})^T C_m^{-1} (m - m_{\text{prior}}) + (Fm - d)^T C_d^{-1} (m - d). \quad (22)$$

Bayesian method is highly computationally intensive as a large number of random samples from probability distributions are generated.

Inversion of a Nonlinear Problem

Geophysical inverse problems are in general nonlinear in nature. The above discussed linear inverse theory for discrete systems is applied to nonlinear problem by quasi-linearizing the nonlinear problem through Taylor's series expansion. Here, the model is expanded around some initial model $\mathbf{m}_0$ assuming that the model perturbation $\delta\mathbf{m}$ is small

$$\mathbf{d} = f(\mathbf{m}) = f(\mathbf{m}_0 + \delta\mathbf{m}) \approx f(\mathbf{m}_0) + \mathbf{G}'\delta\mathbf{m} + \dots, \quad (23)$$

where

$$\mathbf{G}' = \left. \frac{\partial f}{\partial \mathbf{m}} \right|_{\mathbf{m}=\mathbf{m}_0}. \quad (24)$$

Ignoring second- and higher-order derivatives in Eq. 23, we get

$$\delta\mathbf{d} = \mathbf{d} - \mathbf{d}_0 = \mathbf{G}'\delta\mathbf{m}, \quad \mathbf{d}_0 = f(\mathbf{m}_0). \quad (25)$$

Thus, we get again a linear form of inverse problem

$$\delta\mathbf{d} = \mathbf{G}'\delta\mathbf{m}, \quad (26)$$

which is solved for perturbation in model parameters given perturbation in data. Once perturbation in model parameters is estimated, the model is updated and the process is repeated till the desired convergence is obtained. Thus, the solution is obtained iteratively.

The quasi-linearization approach tries to converge to a minimum closest to the chosen initial model $\mathbf{m}_0$. For a highly nonlinear problem having several minima, this approach might yield an incorrect solution corresponding to a local minimum. There are global optimization techniques such as simulated annealing and genetic algorithms (► [Genetic Algorithm](#)) that are used to circumvent the issue of local minima (► [Optimization Methods](#)). Further development has taken place in the form of application of artificial neural networks to construct model parameter space from observed geophysical data (► [Artificial Neural Network](#)). Recently, data-driven approaches of machine learning (► [Machine Learning and Geosciences](#)) and hybrid data-driven and physics-based approaches (► [Deep Learning](#)) have also made in-roads to

interpretation of geophysics data in terms meaningful models of the subsurface.

Geophysical Example

We demonstrate an application of the inverse theory in developing a joint inversion (JI) algorithm for interpretation of seismic and magnetotelluric geophysical data. These two methods sense different physical properties, namely, bulk and shear moduli, and Poisson's ratio through seismic velocities, and electrical resistivity, respectively, of the earth. JI algorithms allow inverting of such diverse and unrelated datasets together, under the assumption that the geometry of the causative structure is the same, to obtain an integrated model of the earth's structure. Here, we give the mathematical description in very brief. For the detailed formulation, one may refer to Manglik and Verma (1998) and Manglik et al. (2009) and for field application to Manglik et al. (2011).

In the forward problem of magnetotelluric method, apparent resistivity (ρ_a) and phase (ϕ_a) measured on the surface of the earth are linked to the resistivity distribution inside the earth in terms of surface impedance Z_1 as

$$\rho_a(\omega) = \frac{1}{|w|} Z_1^* Z_1, \quad \phi_a(\omega) = \tan^{-1} \frac{\text{Im}(Z)}{\text{Re}(Z)}, \quad (27)$$

where $w = \sqrt{-i\omega\mu}$, ω is the angular frequency, μ is the magnetic permeability of the medium, and “*” represents the complex conjugate. For a n -layered model (n^{th} layer begin the half-space), the impedance of the i^{th} layer is represented as

$$Z_i(\omega) = f(Z_{i+1}(\omega), \omega, \rho_i, h_i), \quad \text{with } Z_n(\omega) = w\sqrt{\rho_n}, \quad (28)$$

where ρ_i and h_i are the electrical resistivity and thickness, respectively, of the i^{th} layer. The partial derivatives of ρ_a and ϕ_a with respect to the unknown layer parameters ρ_i and h_i , needed to construct the matrix $\mathbf{G}'$ (Eq.24), can be obtained from Eqs. 27 and 28.

Seismic travel time modeling involves refraction and reflection travel times. For surface seismics, the time t_i^{rr} taken by head waves generated by a critically refracted wave at the i^{th} interface of an isotropic, horizontally layered earth, and recorded at an offset distance X can be expressed as

$$t_i^{\text{rr}} = f[X, v_k(k=1, \dots, i), h_k(k=1, \dots, i-1)], \quad (29)$$

where v_k is the seismic velocity of the k^{th} layer. Similarly, reflection travel time t_i^{rf} of a wave reflected from the i^{th}

interface and recorded at an offset distance x is related to the model parameters as

$$t_i^{rf} = f[R, v_k(k=1, \dots, i), h_k(k=1, \dots, i)]. \quad (30)$$

Here, R is called the ray parameter which is a nonlinear function of the offset distance and velocity and thickness of layers. The partial derivatives of these seismic observations with respect to the model parameters can be obtained from these equations.

Following the quasi-linearization approach (Eqs. 25 and 26), we can write a combined matrix for geometrical JI as (dropping δ for ease of writing expressions)

$$\begin{Bmatrix} d_\rho \\ d_\phi \\ d_{rr} \\ d_{rf} \end{Bmatrix} = \begin{bmatrix} G_\rho^\rho & G_\rho^h & 0 \\ G_\phi^\rho & G_\phi^h & 0 \\ 0 & G_{rr}^h & G_{rr}^v \\ 0 & G_{rf}^h & G_{rf}^v \end{bmatrix} \begin{Bmatrix} \rho \\ h \\ v \end{Bmatrix} \quad \text{or} \quad \mathbf{d} = \mathbf{Gm}. \quad (31)$$

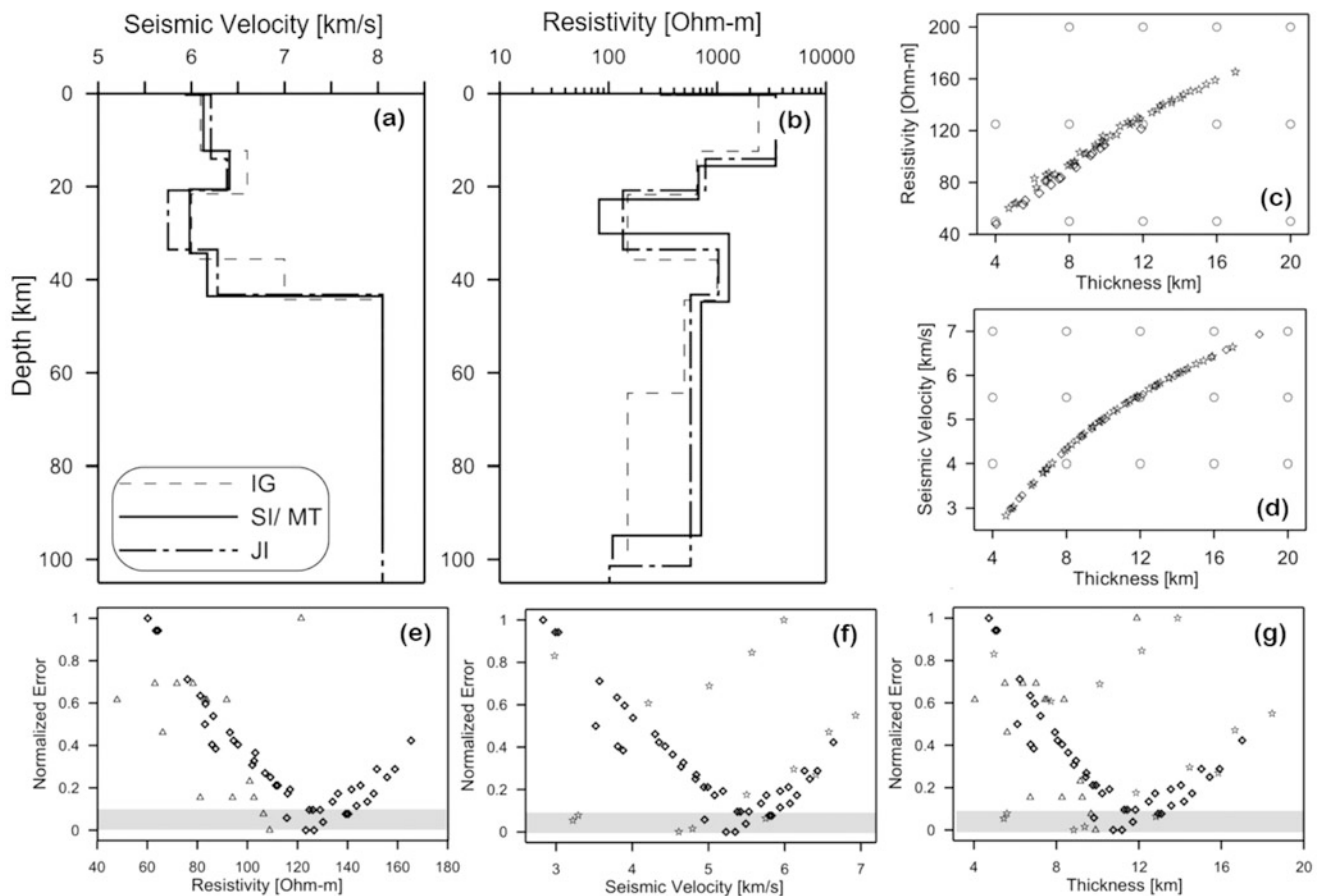
In this matrix, the data column vector $\mathbf{d}$ has the array dimension ($m^a \times 1$), where m^a is the sum of all observations corresponding to the four datasets and the model parameter vector $\mathbf{m}$ has array dimension $(3n - 1)$ corresponding to a layered earth model with $(n - 1)$ layers overlying a half-space. Equation 31 is solved by SVD (Manglik 2021).

A field example of application of this method is discussed by Manglik et al. (2011) who jointly inverted coincident deep crustal seismic (DCS) and magnetotelluric (MT) data at one shot-point along the Kuppam-Palani geotransect in the southern Indian shield. The geotransect was covered by different geophysical methods to delineate the crustal structure of various tectonic blocks of the Archaean to the Neoproterozoic age separated by shear zones. The study delineated a four-layer crustal model with a mid-crustal low velocity layer beneath the entire profile (Reddy et al. 2003) which is also found to be electrically conductive (Harinarayana et al. 2003). Quantification of physical properties and thickness of such low seismic velocity and high electrical conductivity layers (LVCL) sandwiched between high velocity and resistive layers is a challenging task because of certain limitation of these methods. In seismics, the absence of refraction from the top of a LVCL and trapping of seismic energy within the layer severely limits interpretation of seismic data for reliable estimates of the thickness and seismic velocity of the LVCL. Similarly, electrical and electromagnetic methods of geophysical exploration suffer from the well-known problem of equivalence due to which different combinations of conductivity and thickness of a subsurface layer for a 1-D earth can

produce similar response. This leads to a range of acceptable models, all satisfying the minimum root mean square (RMS) error criterion.

Manglik et al. (2011) performed JI of seismic reflection and refraction travel times (SI) and MT apparent resistivity and phase data to test the efficacy of the JI approach (Manglik and Verma 1998) in reducing the problem of equivalence and providing better constrained model of the LVCL compared to the models obtained by inversion of individual datasets. Another advantage of the JI approach is that the features of the model sensed by at least one method can be accounted for in the model and, thus, can yield an improved model of the subsurface structure. For example, a near-surface layer needed to explain the MT data was ignored by the SI method. Similarly, the Moho (a major interface between the crust and the mantle) is a prominent reflector in the SI data but may remain transparent in the MT data. Using these additional details, Manglik et al. (2011) modified the four-layer model of Reddy et al. (2003) to a six-layer model by including a thin near-surface conducting layer and a lithospheric mantle layer below the crust and jointly inverted the SI and MT data. The results are shown in Fig. 1a, b. For comparison, the results obtained by inversion of individual datasets are also shown in the figure. The JI results reveal that the LVCL is about 5 km thicker in comparison to the MT model but it is approximately the same as obtained by the SI method. However, the results depend on the choice of the initial guess due to the equivalence problem. Therefore, it is important to check for the sensitivity of the model to the choice of the starting model.

A model appraisal was performed by Manglik et al. (2011) to test if JI can better constrain the range of acceptable models compared to the range of SI and MT models. This was done by a systematic scan of the model parameter space for the LVCL covering the chosen resistivity, thickness, and velocity range of 50–200 $\Omega\cdot\text{m}$, 4–20 km, and 4–7 km/s, respectively, and constructing the starting models (shown by open circles in Fig. 1c, d) from this model space for inversion. Individual SI and MT inversions and JI inversion were performed for each of these starting models. The final models obtained after inversion corresponding to all starting models are also shown in Fig. 1c, d by diamonds for MT/SI and stars for JI. All these models fall in a narrow zone of minimum RMS error, suggesting that the parameters of LVCL suffer from the problem of equivalence. Since the parameters of LVCL converge in a long narrow zone of minimum RMS error even with JI, a cursory look at Fig. 1c, d would suggest almost no improvement. However, an analysis of the normalized RMS errors of all the final models has revealed a pattern. Here, normalization is done in such a way that the model having largest RMS error has normalized error of 1 and that having least RMS error now has a normalized error of 0. The results (Fig. 1e–g) corresponding to MT (triangles), SI (stars), and JI (diamonds)



Inversion Theory, Fig. 1 (a and b) Seismic velocity and electrical resistivity models obtained by inversion of individual seismic reflection and refraction (SI) data and magnetotelluric apparent resistivity and phase (MT) data as well as their joint inversion (JI). A six-layer earth model was considered for inversion. Starting model for inversion is shown as IG. (c and d) Appraisal of the seismic velocity and electrical resistivity of the LVCL. Circles, diamonds, and stars represent initial value, final value by individual methods, and final value by JI,

respectively. (e–g) Normalized RMS error vs. resistivity (e), seismic velocity (f), and thickness (g) for various models used in (c and d). Models with less than 10% error (shaded region) are considered as acceptable models. Triangle, stars, and diamonds represent MT, SI, and JI solutions, respectively. [Reprinted/adapted by permission from Springer Nature: Springer Nature, Joint Inversion of Seismic and MT Data – An Example from Southern Granulite Terrain, India by A. Manglik, S.K. Verma, K. Sain et al. (2011)].

reveal a narrowing of the zone of acceptable models if one considers the models falling within 10 percent of the least normalized error (shaded region) as acceptable. It can be seen that SI models have more scatter in the range of acceptable models. The LVCL with seismic velocity and thickness in the range of 3.25–5.75 km/s and 5.5–12.8 km, respectively, can fit the observed data equally well. MT gives a better constrained model of the LVCL with resistivity and thickness in the range of 106–110 $\Omega\cdot\text{m}$ and 9.8 km, respectively. In contrast, JI models fall in the seismic velocity range of 5.0–5.8 km/s with the minimum at 5.4 km/s, resistivity in the range of 115–140 $\Omega\cdot\text{m}$ with minimum at 122 $\Omega\cdot\text{m}$, and thickness in the range of 10–13 km with the minimum at 11.5 km. These results indicate that JI gives a well-constrained model of the LVCL and reduces the ambiguity in its parameters.

Summary

Inverse theory has been applied to interpret diverse geophysical and geochemical data to image the earth's interior and understand the processes since its use in seismology by Backus and Gilbert (1968) and Tikhonov and Arsenin (1977). Problem is essentially reduced to solving a linear algebra problem. Current endeavor is to use more data-intensive techniques such as machine learning to reduce uncertainty in the solutions.

Cross-References

- [Artificial Neural Network](#)
- [Bayesian Inversion in Geoscience](#)
- [Deep Learning in Geoscience](#)

- Forward and Inverse Stratigraphic Models
- Genetic Algorithms
- Machine Learning
- Optimization in Geosciences
- Probability Density Function

Acknowledgments AM carried out this work under the project MLP6404-28(AM) with CSIR-NGRI contribution number NGRI/Lib/2020/Pub-201

Bibliography

- Albarede F (1983) Inversion of batch melting equations and the trace element pattern of the mantle. *J Geophys Res* 88:10573–10583
- Aster RC, Borchers B, Thurber CH (2019) Parameter estimation and inverse problems. Elsevier
- Backus GE, Gilbert JF (1967) Numerical applications of a formalism for geophysical inverse problems. *Geophys J R Astron Soc* 13: 247–273
- Backus GE, Gilbert JF (1968) The resolving power of gross earth data. *Geophys J R Astron Soc* 16:169–205
- Backus G, Gilbert JF (1970) Uniqueness in the inversion of inaccurate gross earth data. *Philos Trans Roy Soc Lond A* 266:123–192
- Golub GH, Kahan W (1965) Calculating the singular values and pseudo-inverse of a matrix. *J Soc Indus Appl Math Ser B* 2:205–224
- Golub GH, Reinsch C (1970) Singular value decomposition and least squares solutions. *Numer Math* 14:403–420
- Gupta PK (2011) Inverse theory, linear. In: Gupta H (ed) *Encyclopedia of solid earth geophysics*. Encyclopedia of earth sciences series. Springer, pp 632–639
- Harinarayana T, Naganjaneyulu K, Manoj C, Patro BPK, Kareemunnisa Begum S, Murthy DN, Rao M, Kumaraswamy VTC, Virupakshi G (2003) Magnetotelluric investigations along Kuppam-Palani geotransect, South India – 2-D modeling results. In: Ramakrishnan M (ed) *Tectonics of southern Granulite terrain: Kuppam-Palani geotransect*. Memoir Nr 50. Geological Society of India, Bangalore, pp 107–124
- Manglik A (2021) Inverse theory, singular value decomposition. In: Gupta H (ed) *Encyclopedia of solid earth geophysics*. Encyclopedia of earth sciences series. Springer, Cham. <https://doi.org/10.1007/978-3-030-10475-7>
- Manglik A, Moharir PS (1998) Backus -Gilbert magnetotelluric inversion. In: Roy KK, Verma SK, Mallick K (eds) *Deep electromagnetic exploration*. Narosa Publishing House & Springer, pp 488–496
- Manglik A, Verma SK (1998) Delineation of sediments below flood basalts by joint inversion of seismic and magnetotelluric data. *Geophys Res Lett* 25:4042–4045
- Manglik A, Verma SK, Kumar H (2009) Detection of sub-basaltic sediments by a multi-parametric joint inversion approach. *J Earth Syst Sci* 118:551–562
- Manglik A, Verma SK, Sain K, Harinarayana T, Vijaya Rao V (2011) Joint inversion of seismic and MT data – an example from Southern Granulite Terrain, India. In: Petrovsky E, Ivers D, Harinarayana T, Herrero-Bervera E (eds) *The Earth's magnetic interior*. IAGA special Sopron book series. Springer, pp 83–90
- Marquardt DW (1963) An algorithm for least-squares estimation of nonlinear parameters. *SIAM J Appl Math* 11:431–441
- McKenzie D, O'Nions RK (1991) Partial melt distributions from inversion of rare Earth element concentrations. *J Petrol* 32:1,021–1,091
- Menke W (1984) *Geophysical data analysis: discrete inverse theory*. Academic, New York

Reddy PR, Rajendra Prasad B, Vijaya Rao V, Sain K, Prasada Rao P, Khare P, Reddy MS (2003) Deep seismic reflection and refraction/wide-angle reflection studies along Kuppam-Palani transect in the southern Granulite terrain of India. In: Ramakrishnan M (ed) *Tectonics of Southern Granulite Terrain: Kuppam-Palani geotransect*. Memoir Nr 50. Geological Society of India, Bangalore, pp 79–106

Stewart GW (1993) On the early history of the singular value decomposition. *SIAM Rev* 35:551–566

Tarantola A (1987) *Inverse problem theory: methods for data fitting and model parameter estimation*. Elsevier Science, Amsterdam

Tikhonov AN, Arsenin VY (1977) *Solutions of ill-posed problems*. V.H. Winston and Sons, New York

Inversion Theory in Geoscience

Shib Sankar Ganguli and V. P. Dimri

CSIR-National Geophysical Research Institute, Hyderabad, Telangana, India

Definition

Inverse problems are familiar to numerous disciplines of geosciences that relate physical properties characterizing a model, m , with its model parameters, m_i , to collected geoscientific observations, i.e., data, d . This demands a knowledge of the forward model competent in predicting geoscientific data for specific subsurface geological structures. The modeled data can be expressed as

$$d = G(m, m_i). \quad (1)$$

Here G is the forward modeling operator, which is a nonlinear operator in most geoscientific inverse problems. In practice, for an inverse method, a model of the feasible subsurface is assumed and the model response is computed, which is subsequently compared with the observed data. This procedure is repeated several times until a minimum difference between the model response and observations is obtained.

Inverse Theory with Essential Concepts

In any geoscientific inverse problem, the ultimate aim is to estimate the earth's physical properties from the surface or subsurface measurements, which have been experienced variation in properties of the earth element. Undeniably, inverse theory is also capable of deciphering the quality of the predicted model than just determining model parameters. It can assist in deciding which model parameters or which combinations of model parameters are the most appropriate

representatives of the earth's subsurface. Further, it can be also useful in analyzing the effect of noise on the stability of the solution. For all of these, it is essential to realize the theoretical response of an assumed earth model (i.e., simulation or forward problem), thereby establishing the mathematical relationship among discrete data and a continuously defined model. Generally speaking, a real geological scenario is approximated by considering a simple model, and relevant model parameters are inferred from the observed data. This is known as an inverse problem. Mathematically, it is also possible to formulate a forward problem as an inverse problem. The general forward and inverse problems can be expressed as:

Forward problem : model $\{ \text{model parameters, } m_i \} \rightarrow \text{data}.$

Inverse problem : data $\rightarrow \text{model } \{ \text{estimated model parameters, } m_i \}.$

Finding solutions to the geoscientific inverse problems has always been at the forefront of an active research area, and therefore methods such as direct inversion, linear and iterative inversions, and global optimization methods have been investigated methodically. Even though the forward problem can provide a unique solution, the inverse problem does not. Inverse problem has multiple solutions and hence a priori information becomes vital to obtain meaningful model parameters. In general, a priori information on the model parameters is characterized using a probabilistic point of view on model space. The model space is a hypothetical space with manifold space points, independent of any specific parametrization, each of which essentially signifies a plausible model of the system. Apart from the obvious goal of estimating a set of model parameters, it is equally important to estimate the formal uncertainties in the model parameters. The ultimate goal is, of course, to obtain a reasonable, valid, and acceptable solution to the inverse problem of interest. A comprehensive framework showing the real implementation of various inversion methods including the process workflow to infer earth physical properties is illustrated in Fig. 1. Each element of a typical geophysical inverse problem, central for obtaining a meaningful solution, is indicated by arrows (Fig. 1). It can be seen that the inversion process comprises inputs (e.g., geophysical data, the equations that consider governing physical laws, a priori information about the problem), implementation (e.g., forward simulation and the inversion approach), and evaluation (e.g., assessment of the model). During the implementation stage, it is ensured that the relevant components are properly defined so that a well-posed inverse problem is developed and is being solved mathematically. For an effective geophysical inversion, two abilities of the algorithm become important and these are:

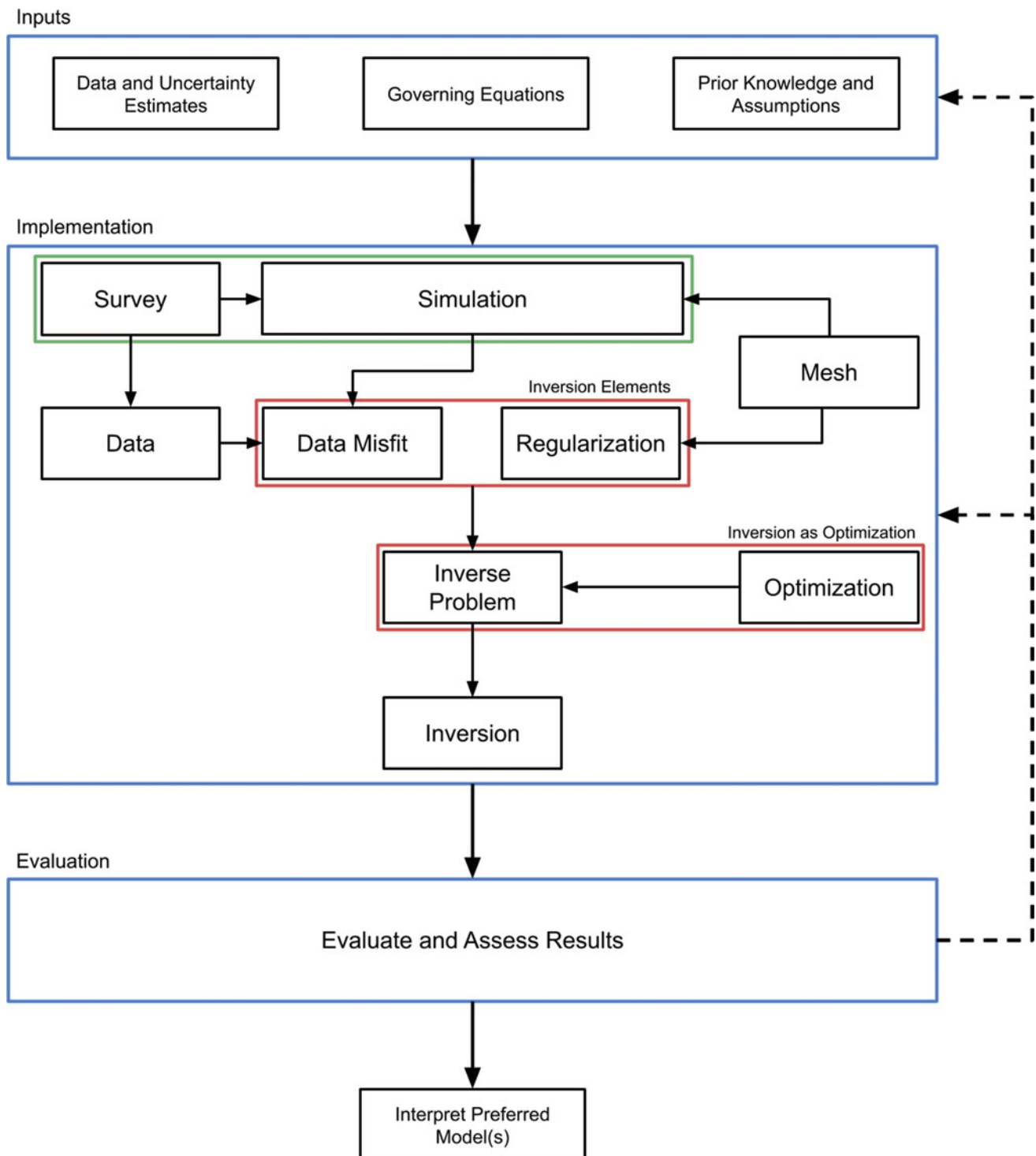
(a) the ability to run a forward simulation and generate synthetic data given a physical model and (b) the ability to appraise the model performance and update the model until the misfit function is optimized (Fig. 1). Regularization plays a critical role in the evaluation of the model's performance so that a suitable model is developed that is in good agreement with a priori information and assumed statistical distributions. Numerically, the inversion problem will be solved through optimization to obtain the unknown earth parameters. During optimization, often we need to alter the model regularization or change the aspects of numerical computations until the final model is accepted based on how well the model fits the observed data with less uncertainty. For more details, several excellent textbooks on the geophysical inverse theory are accessible, and interested readers are referred to Tarantola (1987), Menke (2012), Richter (2020), and others.

In any geoscientific inverse problem, the following significant issues related to the problem of finding solutions remain valid:

1. Does the solution exist for the problem?
2. If yes, then is it unique?
3. Is it stable and robust?

The existence of solution ($\{m, m_i\} = G^{-1}d$) is usually related to the mathematical design of the inverse problem. It is important to note that, from a physical perspective dealing with geological structures, a certain solution to the problem is anticipated. However, it may be also possible that the mathematical perspective could not provide an adequate numerical model that would fit the observations (data). There is an infinite possibility of models of a given subsurface geological condition that could match the data. In general, material properties within the subsurface are characterized by continuous functions, and the inversion method attempts to derive these functions from the set of observations of a finite number of data points. This leads to the problem of nonuniqueness in results. Let us assume two different models, m_1 and m_2 with two different sources, S_1 and S_2 that generate the same data d , then it is difficult to distinguish these two models from the given set of observations, and the solution will be nonunique. A solution will be considered as unique if these two diverse models, m_1 and m_2 , will produce two dissimilar data sets, d_1 and d_2 ($d_1 \neq d_2$).

By and large, geophysical data are invariably contaminated with some amount of noise. Hence, the last problem dealing with the stability and robustness of the solution is pivotal in geophysical inversion. Stability specifies how minor errors obtained during observations spread into the model and robustness deals with the magnitude of insensitivity concerning a small number of large errors within the data.



Inversion Theory in Geoscience, Fig. 1 A comprehensive framework of geophysical inversion including inputs, implementation of algorithms, evaluation, and interpretation (Cockett et al. 2015).

A stable and robust solution is insensitive to small errors in the data. An inverse problem is said to be *ill-posed* if the solution is unstable and nonunique; however, such inverse problems can result in physically and/or mathematically

meaningful solutions if regularization techniques are applied to them (Tikhonov and Arsenin 1977; Dimri 1992).

In practice, inversion methods are categorized into two broad classes: direct inversion methods and model-based

inversion methods. For the sake of completeness, these two are discussed in the subsequent parts of this chapter.

Direct Inversion Methods

In the case of direct inversion, the model is derived directly from the data without any assumptions on the model parameter values. These types of inversion schemes are implemented by a mathematical operator, which is derived based on the governing physics of the forward problem and subsequently applied to the data directly. The layer-stripping method is one of the most familiar direct inversion methods used in seismology to infer the medium properties of 1D acoustic or elastic medium. This is founded on the forward problem dealing with the reflectivity concept accounting for the plane wave response of the medium at individual frequency. A synthetic seismogram can be generated by combining the responses of all the plane waves for a layered earth model and applying inverse Fourier transform to the same. In practice, seismic data is transformed from the travel-time domain to the intercept time ray parameter domain and then the effects of each layer are pulled out by means of downward continuation during the estimation of layer properties. The foundation of this technique is on the assumption that seismic data is collected in noise-free conditions and is essentially in the broadband range. Such a scenario is difficult in a real field study, therefore, information about the frequency content needs to be supplied self-sufficiently in order to obtain stable model parameters. Apart from this, another direct inversion method is Born inversion, which is based on the inverse scattering approach (popular as Born theory). The important limitation of the layer-stripping algorithm is relaxed in this case by considering the earth to be varied in 2D or 3D. This is treated as a concomitant high-priority approach and is of interest since the Born theory implicitly contains migration with a proper band-limited layer property estimate. That said, the algorithm can estimate the precise time and amplitudes of all internal multiples independent of whether the medium is acoustic or elastic, deemed to be useful in deciphering subsurface structures. One of the major limitations of direct inversion methods is that these algorithms are sensitive to noise and often it becomes problematic for the operator to infer meaningful information from the data contaminated with noise. Such issues can be controlled, to some extent, in a model-based inversion approach where uncertainties associated with data are mostly addressed.

Model-Based Inversion Methods

A model-based inversion algorithm employs a forward model to compute synthetic data, which is then compared with the

observed data, unlike the direct inversion approach. The process of generating synthetic data is repeated until a proper match between the computed and observed is obtained. If the match between observed and computed is acceptable with minimum error, the model is accepted as the solution to the inverse problem. This repeated process, in essence, is viewed as an optimization problem in which the derived model elucidates the set of observations in a much more effective sense. Based on the approach of a search for an optimum model, optimization methods differ. The simplest and best-understood optimization method is one that can be represented with an explicit linear equation, which laid the foundation of linear discrete inverse theory. The discrete inverse theory can be expressed as

$$d_k = \sum_{l=1}^M G_{kl} m_l. \quad (2)$$

Whereas the continuous inverse theory considers discrete data, in which a continuous model function can be represented as

$$d_k = \int G_k(x) m(x) dx. \quad (3)$$

As stated earlier, the observed data in inverse problems are always discrete. Hence, a discrete inverse problem is taken into consideration by approximating the data and model as vectors, which are represented in the following forms

$$\begin{aligned} d &= [d_1, d_2, d_3, \dots, d_{ND}]^T, \text{ and } m \\ &= [m(x_1), m(x_2), m(x_3), \dots, m(x_{NM})]^T, \end{aligned}$$

where ND is the total number of data points, NM is the total number of model parameters, and T defines the transpose of a matrix.

Generally speaking, the solution to the linear inverse problem is based on measures of the size or length of the misfit (objective function) between the observed and calculated data. The length, namely Euclidean length, of the prediction error, is usually defined by “norm” and can be represented by the sum of the absolute values of the components of the error vector. Consequently, the best model is then the one that culminates in the smallest overall error (e_k):

$$e_k = d_k^{\text{obs}} - d_k^{\text{pre}} = d_k^{\text{obs}} - G(m). \quad (4)$$

The most commonly used norms are characterized by L_n norm (Menke 1984), where n is the power, takes the following form

$$L_1 \text{ norm : } \|e\|_1 = \left[\sum_{k=1}^{ND} |e_k|^1 \right], \quad (5)$$

$$L_2 \text{ norm : } \|e\|_2 = \left[\sum_{k=1}^{ND} |e_k|^2 \right]^{\frac{1}{2}}, \quad (6)$$

$$L_n \text{ norm : } \|e\|_n = \left[\sum_{k=1}^{ND} |e_k|^n \right]^{\frac{1}{n}}, \text{ where } n = 0, 1, 2, \dots, \infty \quad (7)$$

or, in vector notation, it can be expressed as:

$$L_n \text{ norm : } \|e\|_n = \left[(d_k^{\text{obs}} - G(m))^T (d_k^{\text{obs}} - G(m)) \right]^{\frac{1}{n}}. \quad (8)$$

Different norms offer different weights to outliers. Comparatively, the higher norms provide more weights to large elements of the misfit between the observed and predicted data. To understand how the concept of measure of length can be pertinent to the solution of an inverse problem, let us consider a simple problem of fitting data to (m, d) pairs with a single outlier, as depicted in Fig. 2. We wish to obtain a norm that leads to a straight line that passes across the set of data points (m, d) so that the sum of the squared residuals (errors) is minimal. Figure 2 demonstrates that three lines are fitting the data points corresponding to L_1 , L_2 , and L_∞ norms. As expected, the L_1 norm is least sensitive to the presence of outlier, whereas the L_∞ norm finds the curve which minimizes the largest single error between the observed and

predicted values. Also, the latter is strongly influenced by the outlier and is unable to eradicate the error (residual) since that may establish a large error elsewhere. Hence, the choice of norm determines how the total error can be quantified, wherein the norm is preferred to signify the nature of noise or scatter in the observed data.

Among the various available norms, the L_2 norm is the most widely applied norm in geophysical inverse problems. The preference of L_2 norm over L_1 norm is because the former weights the data outliers that lie nowhere near the average trend (Fig. 2) and obeys Gaussian statistics. Application of other kinds of norms, e.g., L_1 norm, has lately received attention in geophysical inversion (Wang et al. 2020). Generally speaking, a lower order norm is preferred for reliable estimates if the data are likely to be dispersed broadly about the average trend. A comparative analysis of these two norms is summarized in Table 1. It is evident that L_1 norm is more robust as compared to L_2 norm since the former is insensitive to a small number of big errors (outlier) in the data. This characteristic, in principle, gave new impetus to Huber (1964) to introduce the theory of robust estimation. Some interesting pieces of literature for more details about robust estimation are Hampel (1968) and Wilcox (2011).

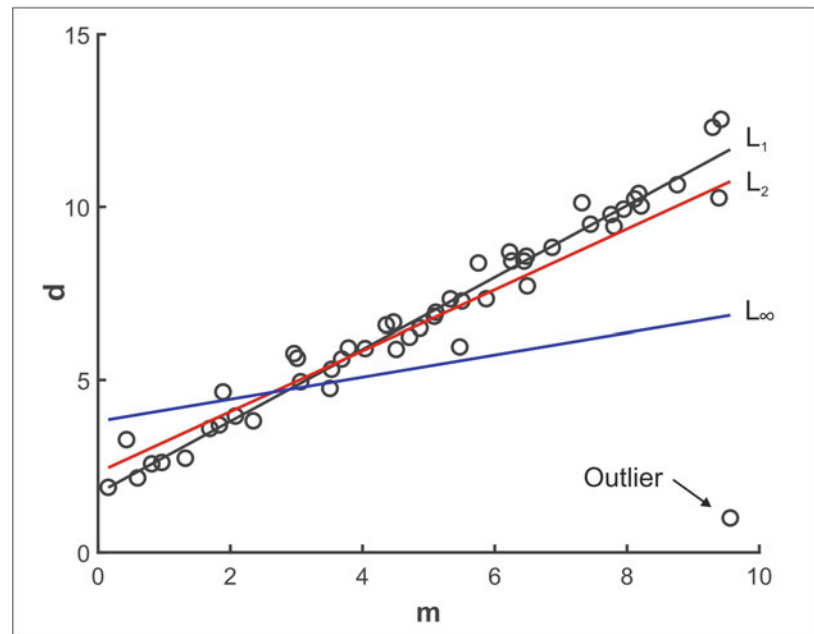
Based on the search method to obtain the optimal solution, the model-based inversion can be classified into four categories, as discussed below.

Linear Inversion Methods

As the name suggests, these types of inversion techniques presume that data and models have a linear relationship. These methods have been implemented using linear algebra

Inversion Theory in

Geoscience, Fig. 2 Fitting a straight line to (m, d) data sets where the measured error is represented by L_1 , L_2 , and L_∞ norms (after Menke 1984). Note the L_1 norm offers a minimum weight to the outlier



Inversion Theory in Geoscience, Table 1 Summary of comparative analysis of using L_1 and L_2 norms

Features	L_1 norm	L_2 norm
<i>Quality aspect</i>	Robust	Not very robust
<i>Output</i>	Sparse	Non-sparse
<i>Accuracy</i>	Better	Good
<i>Computational time</i>	More	Less
<i>Programming used</i>	Linear	Least squares

in many geophysical applications for the past several decades. Of course, these algorithms innately demand certain conditions to establish a linear relationship between data and model. For instance, in Zoeppritz equations, the reflection coefficients are nonlinearly related to the P-wave velocity (V_p), S-wave velocity (V_s), and density. In exploration seismology, the Zoeppritz equations are a set of equations that define the segregation of incident seismic energy at a planar interface, typically a boundary between two layers of rock. These equations become helpful in the detection of petroleum reservoirs, especially during the investigation of the governing factors affecting the amplitude of the plane waves when the angle of incidence is altered. However, these equations are essentially nonlinear and to solve this problem, researchers have derived a linearized form of these equations, which are valid for small angles of reflection. An alternative way to establish a linear relation is to consider a linear relationship among perturbations in data and that in the model. This linear relationship remains valid as long as the perturbation in the model from the initial model is small. The linear relationship between the data perturbation and model perturbation can be accomplished by assuming that due to a small perturbation δm to a reference model m_0 , the model is affected by the forward modeling operator (G) in a way that results in the observed data in the following form

$$d_{\text{obs}} = G(m_0 + \delta m), \text{ and } d_{\text{syn}} = G(m_0).$$

To obtain the linearize relation, we implemented Taylor's series expansion about some initial guess of m_0 as

$$G(m_0 + \delta m) = G(m_0) + \left. \frac{\partial G(m_0)}{\partial m} \right|_{m=m_0} \Delta m + \dots \quad (9)$$

Now neglecting the second- and higher-order terms, we get

$$d_{\text{obs}} = d_{\text{syn}} + \left. \frac{\partial G(m_0)}{\partial m} \right|_{m=m_0} \Delta m \text{ or } \Delta d = G_0 \Delta m, \quad (10)$$

where Δd is the data misfit, i.e., vector difference between the observed data and modeled data, and G_0 is the sensitivity matrix comprises partial derivatives of synthetic data concerning model parameters. After knowing G , and that its

inverse operator exists, we can easily solve for the model vector by implementing the inverse operator for G , i.e., G^{-1} to the data vector. Also, note that it is possible to get an update to the model perturbation by merely applying G_0^{-1} to the data residual vector.

The solution to the linear inverse problem ($d = Gm$) is straightforward and can be derived from the widely used least-square inverse method, as given below:

$$G^T d = (G^T G) m. \quad (11a)$$

Now, assuming that $(G^T G)^{-1}$ exists, we can obtain the model parameter estimates in the form of

$$m^{\text{est}} = (G^T G)^{-1} G^T d. \quad (11b)$$

Thus, from knowledge of the matrix $(G^T G)^{-1} G^T$, which is known as Moore-Penrose generalized inverse (Dimri 1992), and its operation on the data, we can infer about the model parameters of interest. In order to get one unique solution, one of the important criteria is that there are as many systems of equations as the number of unknown model parameters. Such a system involves a situation that the data has enough information on all the model parameters, which means the number of data points is equal to the number of model parameters to be determined, i.e., $ND = NM$. This is a case of an *even-determined* problem. In a system, where $NM > ND$, i.e., the model parameters are more as compared to the observed data, then it is known as an *under-determined* problem. The system of equations, in this case, can determine uniquely some of the model parameters and fails to offer adequate relevant information to estimate uniquely all the model parameters. Generally speaking, this would typically be the situation with almost every geoscientific inverse problem, in which we strive to infer continuous (mainly infinite) earth model parameters from a finite set of observed data. Often, the problem is reduced to an *even-determined* or *over-determined* ($ND > NM$) problem by simply splitting the earth model into a set of discrete layers in lieu of solving for the earth parameters as a continuous function of spatial coordinates. Another possibility of a case may occur in practice, in which the observed data may encompass complete information on several model parameters and not a bit on the others. Such a case is referred to as a *mixed-determined* problem, which arises in seismic tomography studies, there may be one block that receives several rays while other blocks are devoid of a single ray coverage (completely undetermined).

For an *under-determined* problem, the pseudo inverse (Moore-Penrose generalized inverse) can be written as:

$$m^{\text{est}} = G^T (G^T G)^{-1} d, \quad (11c)$$

and for a completely *over-determined* problem, the model parameters can be obtained by the weighted least square solution, as given by

$$\mathbf{m}^{\text{est}} = (\mathbf{G}^T \mathbf{W}_e \mathbf{G})^{-1} \mathbf{G}^T \mathbf{W}_e \mathbf{d}, \quad (11d)$$

where the matrix $\mathbf{W}_e$ is known as the weighting factor, which denotes the relative influence of each error on the total prediction error.

It is important to note that the matrix $\mathbf{G}^T \mathbf{G}$ can occasionally be singular or nearly singular, so the inversion of the matrix is not possible or calculated imprecisely. The damped least squares method is proposed as a solution. For this, the solution to the inverse problem can be represented as

$$\mathbf{m}^{\text{est}} = (\mathbf{G}^T \mathbf{G} + \lambda \mathbf{I})^{-1} \mathbf{G}^T \mathbf{d}, \quad (11e)$$

where λ and $\mathbf{I}$ denote the damping parameter and identity matrix, respectively. The damping parameter (λ) plays a major role in preventing the inversion from blowing-off by increasing the magnitude of eigenvalues of the matrix (Dimri 1992; Menke 2012; Ganguli et al. 2016). The damped least square method is analogous to the regularization method (e.g., Tikhonov regularization) that minimizes the objective function (misfit) in a least square manner and ensures the smoothness of the model parameters. In practice, λ regulates the smoothness of the model which is attained in an iterative sense. For a detailed study on regularization techniques, damped least squares, weighted least squares, and the Backus-Gilbert method for ill-posed inverse problems, see Menke (1984) and Dimri (1992).

For the solution of the inverse problem of type $\mathbf{d} = \mathbf{G}\mathbf{m}$, there may be a situation that the least square inversion begets numerical inaccuracies while generating the matrix $\mathbf{G}^T \mathbf{G}$. In such a scenario, it is prudent to obtain a straightforward solution, i.e., $\mathbf{m}^{\text{est}} = \mathbf{G}^{-1} \mathbf{d}$. In general, it is not feasible to determine $\mathbf{G}^{-1}$ employing standard approaches, since $\mathbf{G}$ is not a square matrix in most of the geoscientific cases. To overcome this, singular value decomposition (SVD) of the matrix $\mathbf{G}$ can be performed to compute the pseudo inverse. SVD is an extremely effective method for factorizing the matrix $\mathbf{G}$ ($\text{ND} \times \text{NM}$) into orthogonal matrices in the following manner

$$\mathbf{G} = \mathbf{U} \mathbf{S} \mathbf{V}^T, \quad (12)$$

where $\mathbf{U}$ ($\text{ND} \times \text{ND}$) is the left orthogonal matrix, is $\mathbf{V}$ ($\text{NM} \times \text{NM}$) the right orthogonal matrix, and $\mathbf{S}$ is a singular value matrix consisting of singular values as diagonal elements in decreasing order, respectively. If $\mathbf{G}$ matrix is singular, then some of the elements in $\mathbf{S}$ will also be zero. Further, these two eigenvector matrices, i.e., $\mathbf{U}$ and $\mathbf{V}$ may be each divided into two submatrices $\mathbf{V} = [\mathbf{V}_p \mid \mathbf{V}_0]$, and $\mathbf{U} = [\mathbf{U}_p \mid \mathbf{U}_0]$, in which $\mathbf{V}_p$ and $\mathbf{U}_p$ are allied with singular values $\mathbf{S}_i \neq 0$ and

$\mathbf{V}_0$ and $\mathbf{U}_0$ are allied with the singular values $\mathbf{S}_i = 0$. The $\mathbf{V}_0$ and $\mathbf{U}_0$ vector spaces are known as “null spaces” because these are essentially the blind spots devoid of illumination by the operator $\mathbf{G}$. To cite some examples, the null space in seismic deconvolution corresponds to the frequencies that lie in the range outward of the estimated wavelet bandwidth. In crosswell seismic tomography, it primarily resembles a rapid change in horizontal velocity. In the presence of null space, the $\mathbf{G}$ matrix can be rewritten as

$$\mathbf{G} = [\mathbf{U}_p \mid \mathbf{U}_0] \begin{bmatrix} \mathbf{S}_p & 0 \\ 0 & 0 \end{bmatrix} [\mathbf{V}_p \mid \mathbf{V}_0]^T. \quad (13)$$

The uniqueness of the solution can be explained by $\mathbf{V}_p$ and $\mathbf{V}_0$ while the existence of the solution is attributed to the vector spaces $\mathbf{U}_p$ and $\mathbf{U}_0$. Therefore, we can write

$$\mathbf{G} = \mathbf{U}_p \mathbf{S}_p \mathbf{V}_p^T \text{ and } \mathbf{G}^{-1} = \mathbf{V}_p \mathbf{S}_p^{-1} \mathbf{U}_p^T.$$

Finally, we have the estimated model parameters as

$$\mathbf{m}^{\text{est}} = \mathbf{G}^{-1} \mathbf{d} (= \mathbf{G}\mathbf{m}) = \mathbf{V}_p \mathbf{S}_p^{-1} \mathbf{U}_p^T \mathbf{U}_p \mathbf{S}_p \mathbf{V}_p^T \mathbf{m} = [\mathbf{V}_p \mathbf{V}_p^T] \mathbf{m} = \mathbf{R}\mathbf{m}, \quad (14)$$

where $\mathbf{R} = \mathbf{V}_p \mathbf{V}_p^T$ is the model resolution matrix. If this matrix is an identity matrix ($\mathbf{R} = \mathbf{I}$), the solution is unique and the resolution is perfect. Otherwise, the resolution is not good with the estimated model parameters being some weighted averages of the true model parameters. Likewise, we can also write

$$\mathbf{d}^{\text{est}} = \mathbf{G}\mathbf{m}^{\text{est}} = [\mathbf{U}_p \mathbf{U}_p^T] \mathbf{d} = \mathbf{N}\mathbf{d}. \quad (15)$$

Here $\mathbf{N} = \mathbf{U}_p \mathbf{U}_p^T$ is called data resolution matrix, which measures the capability of the inverse operator to uniquely estimate the synthetic data. If $\mathbf{N} = \mathbf{I}$, the modeled data will be precisely similar to the observed data.

Iterative Linear Inversion Methods

We have now seen that the solution of the inverse problem of the type $\mathbf{d} = \mathbf{G}\mathbf{m}$ can be straightforwardly obtained by the method of least squares or the SVD method. However, such linear methods do not offer a solution if the linearity relation is not valid, especially for quasi-linear problems. These types of problems are solved by the iterative linear methods (also known as calculus-based methods), in which the minimum of the error function can be achieved iteratively by consecutive linear guesstimates at individual iteration, and refined successively until the stopping criterion is satisfied (Sen and Stoffa 1995). Most of the iterative methods do not necessitate the matrix to be explicitly well-defined. This inversion scheme updates the reference model (or a priori model) to

compute a new model and the difference between observed and synthetic data (Δd) is measured to evaluate the model performance. The process is repeated until the data misfit achieves its minimum and the relevant equation for the method is given below (Sen and Stoffa 1995):

$$m_{i+1} = m_i + G^{-1}[d_{\text{obs}} - G(m_i)], \quad (16)$$

where $G(m_i)$ generates the synthetic data for a model m_i to match iteratively the observed data, d_{obs} . This inversion method is reasonably sensitive to the initial model guess, therefore, among several minima, it prefers to choose the minimum of the error function adjacent to the starting model. In order to avoid the uncertainties due to data and prior models, both the data and model can be represented by random variables and the inversion can be solved in the context of probability theory.

Grid Search Method

This method entails a methodical search through each point (every possible) in model space to select the solution with the smallest error. For many geophysical inverse problems, it is a trying task to find “every possible” solution while computing the synthetic data for large model space. Yet a large set of trial solutions can be drawn from a regular grid in the model space. This method is most practical for problems where the total number of model parameters (NM) is small (trial solutions (L) can be proportional to L^{NM} , where the grid is M-dimensional) and the solution falls within a certain range of values which may be utilized to describe the limits of the grid. In general, several approximations are invoked to estimate the model parameters when the inputs are inadequate to use the grid search algorithm. Some applications of the grid search method to infer seismic parameters can be found in Koesoemadinata and McMechan (2003), Lodge and Helffrich (2009), and others.

Monte Carlo Method

The Monte Carlo method (MC) is an amendment of the grid search method in which the trial solutions are obtained, self-sufficiently of earlier ones, after some small and finite number of trials by random sampling of the model space. This involves a completely blind search of the model space, hence computationally expensive. In this method, theoretical values are computed based on each model parameter varying within a preset search interval (say $m_i^{\text{min}} \leq m_i \leq m_i^{\text{max}}$) and subsequently compared with the observed data. Afterward, a random number, say R_n , is sampled from a uniform distribution, $U [0,1]$, and is subsequently mapped into a model parameter. The new model parameter can be obtained by

$$m_i^{\text{new}} = m_i^{\text{min}} + R_n(m_i^{\text{max}} - m_i^{\text{min}}). \quad (17)$$

The latest model parameter is thus produced by random perturbation of a definite number of model parameters in the model vector. Model response is analyzed for the new model and compared with the observed data. If the difference between the observed data and modeled data is less than the previously accepted solution, the model is accepted. Otherwise, the model is rejected and the search continues several times until convergence is met. The most popular convergence criterion, in this case, is the total number of accepted models. With the advent of computational efficiency, millions of models can be attempted to obtain a suitable solution. One of the well-known applications of the MC algorithm in geophysics is to the inversion of seismic body wave travel times generating five million earth models with the data comprising 97 eigenperiods of free oscillations, travel times of P and S waves, and mass and moment of inertia of the earth (Press 1968). Finally, out of five million models, only six models met all the constraints. Through this work, Press (1968) addressed for the first time the problem of uniqueness through direct samples drawn from the model space.

There are, however, always two important issues with the MC method that need to be pondered and these are: (a) how to know whether one has tested sufficient models that represent the observations properly? and (b) how the successful models can be utilized to estimate the uncertainty in the estimation?

To make this method to be more practical in use, search space can be compacted, in fact, it can be defined through a pdf (preferably Gaussian with small variance) instead of a uniform distribution. It is desirable to devise random search-based algorithms that can sample a large portion of the model space more efficiently considering randomly selected several error functions. Due to the large degree of randomness, we can expect unbiased solutions. Also, it is equally important to avoid getting trapped in a local minimum of the error function. These types of issues can be well addressed by the directed Monte Carlo methods, e.g., simulated annealing and genetic algorithms. These two methods utilize random sampling to guide their search for models and results in a global minimum of the misfit function amidst several local minima, unlike the pure MC method. Both methods are suitable for large-scale geoscientific problems. Numerous pieces of literature are available that discuss the applications of MC methods and directed MC methods to solve various geoscientific problems (Sambridge and Drij koningen 1992; Sen and Stoffa 1995; Beaty et al. 2002) and many others.

Applications of Inverse Theory

Seismic Inversion to Detect Thin Layers

As an application of inverse theory in geophysics, we will provide here an example of a basis pursuit seismic inversion to detect the stratigraphic layer boundaries, especially

resolving thin layers, beneficial for reservoir characterization. Hydrocarbon exploration and development studies demand proper identification of layer boundaries and thin beds (mainly interbedded) in seismic. Although challenging, thin-bed imaging aids to develop a suitable petrophysical model that is more representative of the reservoir. Given the dominant frequency content of seismic, delineation of a bed with only a few meters thickness is still a burdensome job. Seismic inversion based on the least squares method involving Tikhonov regularization often fails to provide focused transitions between the adjacent layers. The basis pursuit inversion (BPI) algorithm based on L_1 norm optimization method can be formulated to obtain a more reliable detection of sharp boundaries between thin layers. The inverse problem is to infer the physical properties such as velocity (both V_p and V_s) and density of the stratigraphic layers from the amplitude variation with angle (AVA) seismic data. The initial model for prestack seismic data can be built in a similar way as performed in poststack case through interpolation and extrapolation of the well-log derived V_p , V_s , and density values. The BPI algorithm defines the objection function that comprises L_2 norm of the data misfit function and L_1 norm of the solution, and these functions are minimized concurrently to attain the final solution (Zhang et al. 2013). The objective function in the BPI method takes the following form

$$\min[\|d - G_w m_w\|_2 + \lambda \|m_w\|_1]. \quad (18)$$

Initially, a set of angle-dependent wavelets were identified through the calibration of individual angle datum with well log reflectivity, as a substitute for constant wavelets. The wavelet kernel matrix for the respective angle data was developed exploiting each wavelet having a length of 200 milliseconds (ms). The derivation of the density was ignored here since the maximum angle of investigation is only 20° , which is not suitable for inferring density. After testing the BPI algorithm in one well location, it was subsequently applied to a real field 2D seismic profile. The well log velocities are blocky filtered by a 10 ms window that essentially replaces the entire values within this specific window into the mean value. The inverted velocities were found to be in good agreement with the well log measured velocities, and the bed boundaries were well resolved in the seismic (Zhang et al. 2013). This also stands valid when 1D velocities derived from the well log, BPI method, and conventional method were compared.

For a proper comparison, the blocky filtered well log velocities were adapted because the well log contains high frequencies, and it is not advisable to compare the unfiltered well data with the inverted results. As a whole, the BPI algorithm could delineate the bed boundaries exceptionally well as compared to the conventional approach, offering valuable information for quantitative interpretation of elastic

properties. Of course, some discrepancies between the well log measured velocities and inversion results were detected; however, the BPI inverted results could follow the similar trend adopted by the conventional method. The BPI algorithm includes a wedge dictionary that supports an improved resolution of the single as well as double thin beds (interbedded layers) potential for hydrocarbon exploration. To get more insights on the BPI method and its detailed application for reservoir characterization, readers are referred to Zhang et al. (2013).

SVD Gravity Inversion to Identify Geological Structures

The second application of inverse theory in geoscience is presented here as SVD gravity inversion to decipher geological structures. The Bouguer gravity anomaly was analyzed in terms of eigenimages using SVD inversion. In practice, SVD decomposes the matrix G into a sequence of eigenimages, which appeared to be suitable for the separation of signal from background noise. In consonance with the nonlinear theory, singular geological processes are those which entail anomalous aggregates of energy release or accumulation within a restricted range of space or time. Any such anomalous energy aggregates due to geological features may be captured in a few particular eigenimages. Thus, to characterize the geological structure's potential for hydrocarbon or mineral prospects, it is significant to delineate those specific eigenimages employing the SVD inversion method. The SVD inferred singular values possess a unique feature of signifying diverse weighting coefficients of eigenimages. The SVD of the matrix G can be expressed as (Ganguli and Dimri 2013):

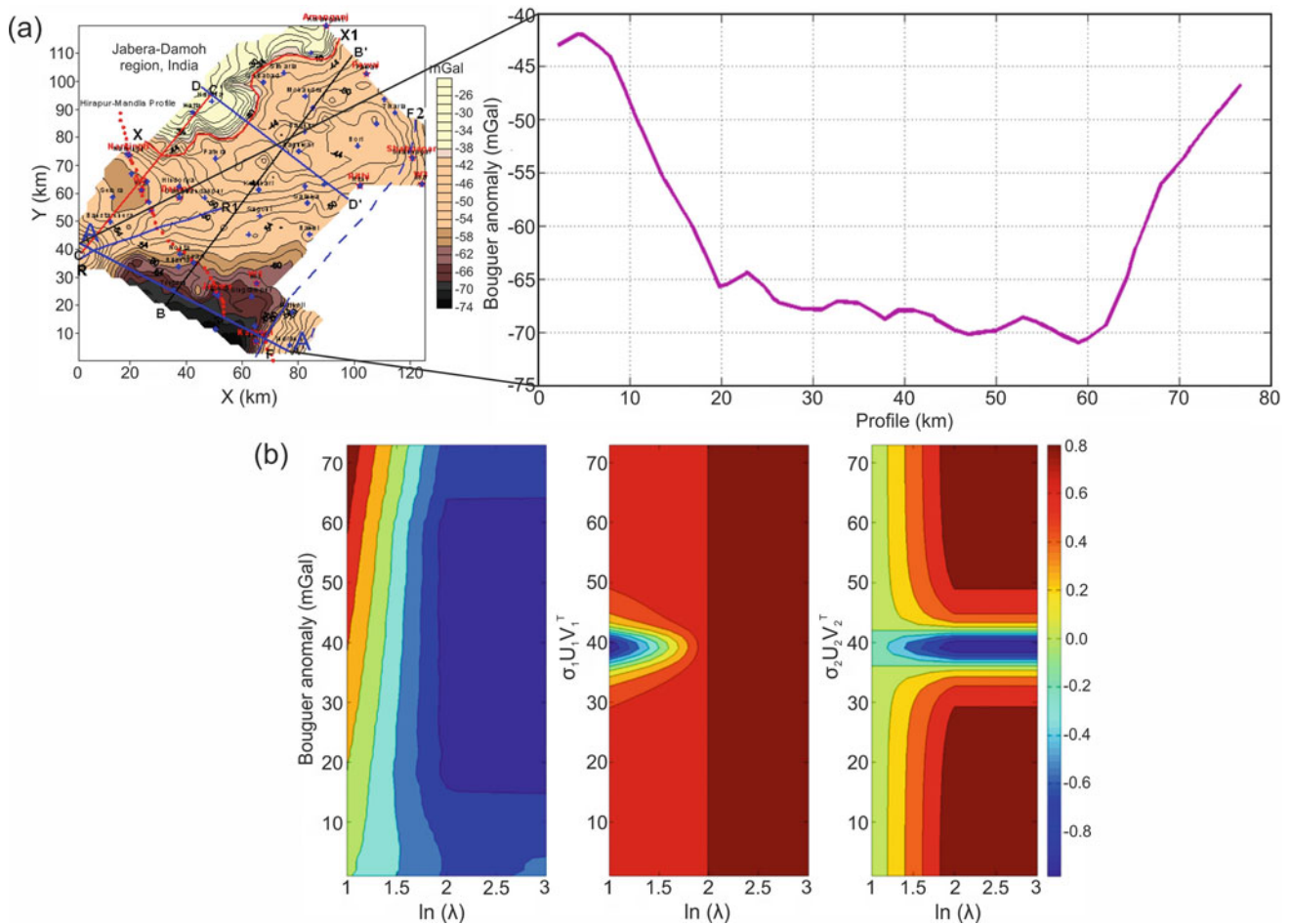
$$G = \sum_{k=1}^l \sigma_k U_k V_k^T. \quad (19)$$

According to the above equation, the G matrix can be reconstructed using the sequence of eigenimages. Note that the signal (i.e., Bouguer anomaly) can be suitably reconstructed from only those few eigenimages whose eigenvalues are not zero. In order to know which eigenimages will contribute more to the Bouguer anomaly reconstruction, the following equation can be considered:

$$P_k = \sigma_k^2 / \sum_{j=1}^l \sigma_j^2 = \lambda_k / \sum_{j=1}^l \lambda_j, \quad (20)$$

where P_k denotes the percentage of each eigenimage contributing to the reconstruction, l is the rank of the matrix, and λ_k represents the eigenvalues in descending order.

In this study, the application of eigenimage extraction using the SVD inversion is verified over a real field Bouguer anomaly from the Jabera-Damoh region, Vindhyan Basin (India). The Jabera-Damoh region has received importance



Inversion Theory in Geoscience, Fig. 3 (a) the Bouguer anomaly map along with the profile under investigation (i.e., AA', as marked by blue line); (b) the reconstructed eigenimages inferred from SVD

inversion on the Bouguer gravity data from the studied region. (Modified after Ganguli and Dimri 2013)

in recent decades due to its potential for hydrocarbon prospects; the national oil and gas company has discovered gas near a borehole at a depth of around 3 km depth. (Ganguli and Dimri 2013). A Bouguer anomaly profile, namely AA' (marked by the blue line in Fig. 3a) from this region was considered for the analysis since it passes through one of the major anomalous parts of this region. The key objective was to identify some important singular geological processes relevant for hydrocarbon prospects in the region through eigenimage analysis. It was identified that about 85% of the total anomalous energy of this gravity signal was dominantly captured by the first eigenimage and the remaining by the second eigenimage. The regional geological structure was efficiently captured by the first eigenimage, while the second eigenimage depicts the local geological characteristics together with noise from different sources (Fig. 3b). Note that the reconstructed eigenimages for the real field gravity data are akin to those derived for a buried faulted structure. In such a case, the sedimentary basin faulted on both margins can be characterized by relatively low (negative) singular values at the center

of the second eigenimage, separating the footwall and hanging wall corresponding to higher and positive singular values (Fig. 3b). The low singular values can be attributed to accumulated sediments of a low density from the studied sedimentary basin. The structure depicted in the second eigenimage may be related to the shallow features and not due to deeper geological bodies since the contribution of anomalous signal in the second eigenimage is always less as compared to the first eigenimage. This observation is further supported by other geophysical and geological studies from this basin, suggesting the study area to be dominated by faulted basin in the crystalline basement. For detailed analysis and interpretation, we refer the interested readers to the work by Ganguli and Dimri (2013).

Summary and Conclusions

The inverse theory is crucial to geoscience studies for inferring the physical properties of the earth system. Here, we

presented an overview of the inverse theory together with its fundamental concepts and two applications to demonstrate the power of inversion in solving seismic and gravity inverse problems. Most direct inversion methods are not powerful to invert the data that are inadequate to recover properties of the earth and contaminated with noise. This unlocked an opportunity for the application of model-based inversion techniques to derive rock properties of interest, which have gained considerable popularity. Much of the success of the inversion method lies in the data quality and forward modeling approach. It is important to have a nominal set of model parameters that fully characterize the system, together with the proper implementation of a priori information and a forward model that can make predictions on some observable parameters based on relevant physical laws. Most of the forward operators are nonlinear and the performance of the inversion algorithm can be arbitrated by how better we elucidate the geoscientific observations knowing their uncertainty. We can get a simple solution if the inverse problem is linearized by imposing a set of limiting conditions to make the forward operator a linear one. Otherwise, iterative methods such as gradient-based methods and random sampling-based search methods can be adapted to solve the geoscientific inverse problems. Several plausible solutions (equally valid) can be derived, owing to the non-uniqueness of the problem. Global optimization methods have been useful to select the final model that attains the global minimum of the objective (misfit) function amidst several minima. These algorithms, however, do not presume any shape of the objective function and are independent of any type of options for starting models. Often highly nonlinear problems can be solved by considering an initial model very close to the peak of the probability density function (PPD) while using these methods. In reality, of course, estimating PPD is quite challenging in a large multi-dimensional model space. We have presented only two applications of inverse theory in geoscience, particularly in seismic and gravity studies. Nevertheless, the application and advanced proposals for the solution of the inverse problems are incessantly intensifying. Physics-guided machine learning-based algorithms for solving large-scale complex systems of equations are receiving augmented attention. With the advent of fast and economical computing facilities, these applications are encouraged and will continue to grow to solve various geoscientific problems.

Cross-References

- [Forward and Inverse Stratigraphic Models](#)
- [Inversion Theory](#)
- [Monte Carlo Method](#)

Bibliography

- Beatty KS, Schmitt DR, Sacchi M (2002) Simulated annealing inversion of multimode Rayleigh wave dispersion curves for geological structure. *Geophys J Int* 151(2):622–631
- Cockett R, Kang S, Heagy LJ, Pidlisecky A, Oldenburg DW (2015) SimPEG: an open source framework for simulation and gradient based parameter estimation in geophysical applications. *Comput Geosci* 85:142–154
- Dimri VP (1992) Deconvolution and inverse theory: application to geophysical problems. Elsevier, Amsterdam
- Ganguli SS, Dimri VP (2013) Interpretation of gravity data using eigenimage with Indian case study: a SVD approach. *J Appl Geophys* 95: 23–35
- Ganguli SS, Nayak GK, Vedanti N, Dimri VP (2016) A regularized Wiener–Hopf filter for inverting models with magnetic susceptibility. *Geophys Prospect* 64(2):456–468
- Hampel FR (1968) Contributions to the theory of robust estimation. Unpublished dissertation, University of California, Berkeley
- Huber PJ (1964) A robust estimation of location parameter. *Ann Math Stat* 35:73–101
- Koesoemadinata AP, McMechan GA (2003) Petro-seismic inversion for sandstone properties. *Geophysics* 68(5):1446–1761
- Lodge A, Helffrich G (2009) Grid search inversion of teleseismic receiver functions. *Geophys J Int* 178:513–523
- Menke W (1984) *Geophysical data analysis: discrete inverse theory*. Academic, New York
- Menke W (2012) *Geophysical data analysis: discrete inverse theory*, 3rd edn. Academic, 330 pp
- Press F (1968) Earth models obtained by Monte Carlo inversion. *J Geophys Res* 73(16):5223–5234
- Richter M (2020) *Inverse problems: basics, theory and applications in geophysics*. Lecture notes in geosystems mathematics and computing, vol XIV, 2nd edn. Birkhäuser, Basel, 273 pp
- Sambridge M, Drij Koningen G (1992) Genetic algorithms in seismic waveform inversion. *Geophys J Int* 109:323–342
- Sen MK, Stoffa PL (1995) *Global optimization methods in geophysical inversion*. Elsevier, Amsterdam
- Tarantola A (1987) *Inverse problem theory, methods of data fitting and model parameter estimation*. Elsevier, Amsterdam
- Tikhonov AN, Arsenin V (1977) *Solution of ill-posed problems*. Wiley, Washington, DC
- Wang R, Wang Y, Rao Y (2020) Seismic reflectivity inversion using an L_1 -norm basis-pursuit method and GPU parallelization. *J Geophys Eng* 17:776–782
- Wilcox R (2011) *Introduction to robust estimation and hypothesis testing*, 3rd edn. Elsevier
- Zhang R, Sen MK, Srinivasan S (2013) A prestack basis pursuit seismic inversion. *Geophysics* 78(1):R1–R11

Iterative Weighted Least Squares

Sara Taskinen and Klaus Nordhausen
Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland

Definition

Iterative (re-)weighted least squares (IWLS) is a widely used algorithm for estimating regression coefficients. In the

algorithm, weighted least squares estimates are computed at each iteration step so that weights are updated at each iteration. The algorithm can be applied to various regression problems like generalized linear regression or robust regression. In this entry, we will focus, however, on its use in robust regression.

Introduction

Consider a data set consisting of n independent and identically distributed (iid) observations $(\mathbf{x}_i^\top, y_i)$, $i = 1, \dots, n$, where $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$ and y_i are the observed values of the predictor variables and the response variable, respectively. The data are assumed to follow the linear regression model

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i,$$

where the p -vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ contains unknown regression coefficients, which are to be estimated based on the data. The errors ϵ_i are iid with mean zero and variance σ^2 and independent of $\mathbf{x}_i$.

The well-known least squares (LS) estimator for $\boldsymbol{\beta}$ is the $\hat{\boldsymbol{\beta}}$ that minimizes the sum of squared residuals $\sum_{i=1}^n r_i(\boldsymbol{\beta})^2$, where $r_i(\boldsymbol{\beta}) = y_i - \mathbf{x}_i^\top \boldsymbol{\beta}$, or equivalently, solves $\sum_{i=1}^n r_i(\boldsymbol{\beta}) \mathbf{x}_i = \mathbf{0}$. To put this into a matrix form, let us collect the responses and predictors into a $n \times 1$ vector $\mathbf{y} = (y_1, \dots, y_n)^\top$ and a $n \times p$ matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$, respectively, then the LS problem is given by

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \quad (1)$$

and the estimate can be simply computed as

$$\hat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

assuming that $(\mathbf{X}^\top \mathbf{X})^{-1}$ exists.

If the assumption of constant variance of the errors ϵ_i is violated, we can use the weighted least squares method to estimate the regression coefficients $\boldsymbol{\beta}$. The weighted least squares (WLS) estimate solves $\sum_{i=1}^n w_i r_i(\boldsymbol{\beta}) \mathbf{x}_i = \mathbf{0}$, where $w_1, \dots, w_n$ are some nonnegative, fixed weights. If we let $\mathbf{W}$ be a $n \times n$ diagonal matrix with weights $w_1, \dots, w_n$ on its diagonal, then the WLS estimate can be computed by applying ordinary least squares method to $\mathbf{W}^{1/2} \mathbf{y}$ and $\mathbf{W}^{1/2} \mathbf{X}$. Thus,

$$\hat{\boldsymbol{\beta}}_{WLS} = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y}. \quad (2)$$

If we assume that the errors ϵ_i are independent with mean zero and variance σ_i^2 , then the WLS estimate with weights

$w_i = 1/\sigma_i^2$ is optimal. If we further assume that the error distribution is Gaussian, then the WLS estimate is the maximum likelihood estimate. Notice that in practice the weights are often not known and need to be estimated. Some examples of weight selection are given, for example, in Montgomery et al. (2012). Notice also that linear regression with non-constant error variance is only one application area of WLS.

Iterative Weighted Least Squares

When the weight matrix $\mathbf{W}$ in (2) is not fixed, but may, for example, depend on the regression coefficients via residuals, we can apply the iterated weighted least squares (IWLS) algorithm for estimating the parameters. In such a case, the regression coefficients and weights are updated alternately as follows:

1. Compute an initial regression estimate $\hat{\boldsymbol{\beta}}_0$.
2. For $k = 0, 1, \dots$, compute the residuals $r_{i,k}(\hat{\boldsymbol{\beta}}_k) = y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_k$ and weight matrix $\mathbf{W}_k = \operatorname{diag}(w_{1,k}, \dots, w_{n,k})$, where $w_{i,k} = w(r_{i,k}(\hat{\boldsymbol{\beta}}_k))$ with some weight function w . Then update $\hat{\boldsymbol{\beta}}_{k+1}$ using WLS as in (2) with weight matrix $\mathbf{W}_k$, that is,

$$\hat{\boldsymbol{\beta}}_{k+1} = (\mathbf{X}^\top \mathbf{W}_k \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}_k \mathbf{y}.$$

3. Stop when $\max_i |r_{i,k} - r_{i,k+1}| < \epsilon$, where ϵ may be fixed or related to the residual scale.

It is shown in Maronna et al. (2018) that the algorithm converges if the weight function $w(x)$ is non-increasing for $x > 0$, otherwise starting values should be selected with care.

IWLS and Robust Regression

Let us next illustrate how IWLS is used in the context of robust regression, which is widely used in geosciences as atypical observations should not have an impact on the parameter estimation and often should be detected. For a recent comparison of nonrobust regression with robust regression applied to geochemical data, see, for example, van den Boogaart et al. (2021). For a given estimate $\hat{\sigma}$ of the scale parameter σ , a robust M estimate of regression coefficient $\boldsymbol{\beta}$ can be obtained by minimizing

$$\sum_i \rho\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\right), \quad (3)$$

where ρ is a robust, symmetric ($\rho(-r) = \rho(r)$) loss function with a minimum at zero (Huber 1981), or equivalently, by solving

$$\sum_i \psi\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\right) \mathbf{x}_i = \mathbf{0}, \quad (4)$$

where $\psi = \rho'$. Here the scale estimate $\hat{\sigma}$ is needed in order to guarantee the scale equivariance of $\hat{\boldsymbol{\beta}}$. If the estimate is not known, it can be estimated simultaneously with $\hat{\boldsymbol{\beta}}$. For possible robust scale estimators, see, for example, Maronna et al. (2018).

The IWLS algorithm can be used for estimating robust M estimates as follows. Write $w(x) = \psi(x)/x$ and further $w_i = w(r_i(\boldsymbol{\beta})/\hat{\sigma})$. Then the M estimation equation in (4) reduces to

$$\sum_i w_i r_i(\boldsymbol{\beta}) \mathbf{x}_i = \mathbf{0},$$

and the robust M estimate of regression can be computed using the IWLS algorithm. For the comparison of robust M estimates computed using different algorithms, see Holland and Welsch (1977). For monotone $\psi(x)$, the IWLS algorithm converges to a unique solution given any starting value. Starting values, however, affect the number of iterations and should therefore be chosen carefully. Maronna et al. (2018) advice to use the least absolute value (LAV) estimate as $\boldsymbol{\beta}_0$. If the robust scale is estimated simultaneously with the regression coefficients, it is updated at each iteration step. The median absolute deviation (MAD) estimate can then be used as a starting value for the scale. For a discussion on

convergence in the case of simultaneous estimation of scale and regression coefficients, see Holland and Welsch (1977).

Some widely used robust loss functions include the Huber loss function

$$\rho_H = \begin{cases} \frac{1}{2}x^2, & |x| \leq c \\ c\left(|x| - \frac{c}{2}\right), & |x| > c \end{cases}$$

yielding

$$\psi_H(x) = \begin{cases} x, & |x| \leq c \\ \text{sign}(x), c & |x| > c, \end{cases}$$

and the Tukey biweight function

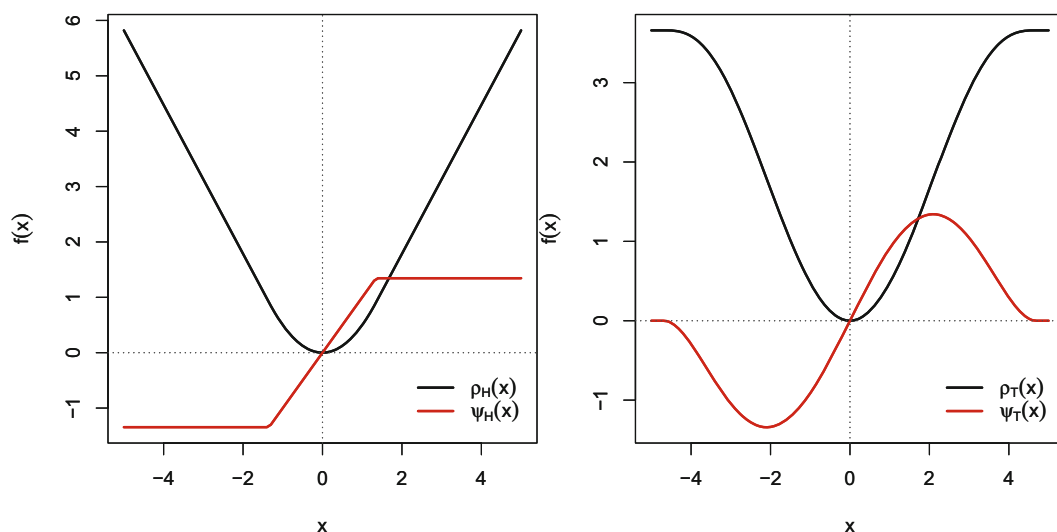
$$\rho_T = \begin{cases} 1 - \left[1 - \left(\frac{x}{c}\right)^2\right]^3, & |x| \leq c \\ 1, & |x| > c, \end{cases}$$

yielding

$$\psi_T(x) = x \left[1 - \left(\frac{x}{c}\right)^2\right]^2 I(|x| \leq c).$$

The tuning constant c provides a trade-off between robustness and efficiency at the normal model. Popular choices are $c_H = 1.345$ and $c_T = 4.685$ which yield an efficiency of 95% for the corresponding estimates. The two loss functions and their derivatives are shown in Figure 1.

Notice that as the Tukey biweight function is not monotone, the IWLS algorithm may converge to multiple solutions. Good starting values are thus needed in order to ensure convergence to a good solution. The asymptotic normality



Iterative Weighted Least Squares, Fig. 1 Huber's and Tukey's biweight ρ and ψ functions with c chosen such that the efficiency at the normal model is 95%.

of robust M regression estimates is discussed, for example, in Huber (1981).

Other Usages of IWLS

Generalized linear models (GLM, McCullagh and Nelder 1989), a generalization of the linear model described above, allows the response variable to come from the family of exponential distributions to which also the normal distribution belongs to. The regression coefficients in GLM are usually estimated via the maximum likelihood method which coincides with the LS method for normal responses. However, for other members from the exponential family, there are no closed form expressions for the regression estimates. The standard way of obtaining the regression estimates is the Fisher scoring algorithm which can be expressed as an IWLS problem (McCullagh and Nelder 1989). The Fisher scoring algorithm creates working responses at each iteration step. Therefore, when using IWLS for estimating coefficients of GLMs, not only are the weights in $\mathbf{W}$ updated at each iteration but also the (working) responses $\mathbf{y}$. How $\mathbf{W}_k$ and $\mathbf{y}_k$ are updated depends on the distribution of the response. For details, see, for example, McCullagh and Nelder (1989). The use of GLMs in soil science is, for example, discussed in Lane (2002).

The idea in robust M estimation described above is to down-weight large residuals, whereas in the LS method, where the L_2 norm is used in the minimization problem (1), these get a large weight. Another way of approaching the estimation is to consider another norm, such as a general L_p norm, which then leads to

$$\arg \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p^p,$$

which, for example, for $p = 1$ corresponds to the least absolute value (LAV) regression. It turns out that to solve the L_p norm minimization problem, also IWLS can be used with weights $\mathbf{W}_k = \text{diag}(r_1(\boldsymbol{\beta}_k)^{2-p}, \dots, r_n(\boldsymbol{\beta}_k)^{2-p})$. Uniqueness of the solution and behavior of the algorithm depend on the choice of p . Also residuals, which are too close to zero often, need to be replaced by a threshold value. For more details, see, for example, Gentle (2007) and Burrus (2012). The motivation for using L_p norm in regression is that it may produce sparse solutions for $\boldsymbol{\beta}$ and thus can be used in

image compression such as in hyperspectral imagery (see, e.g., Zhao et al. 2020, for details).

Summary and Conclusion

The iterative weighted least squares algorithm is a simple and powerful algorithm, which iteratively solves a least squares estimation problem. The algorithm is extensively employed in many areas of statistics such as robust regression, heteroscedastic regression, generalized linear models, and L_p norm approximations.

Cross-References

- Least Absolute Value
- Least Mean Squares
- Least Squares
- Locally Weighted Scatterplot Smoother
- Ordinary Least Squares
- Regression

References

- Burrus CS (2012) Iterative reweighted least squares. OpenStax CNX
- Gentle JE (2007) Matrix algebra. Theory, computations, and applications in statistics. Springer, New York
- Holland PW, Welsch RE (1977) Robust regression using iteratively reweighted least-squares. *Commun Stat Theor Methods* 6(9): 813–827. <https://doi.org/10.1080/03610927708827533>
- Huber PJ (1981) Robust statistics. Wiley, New York
- Lane PW (2002) Generalized linear models in soil science. *Eur J Soil Sci* 53(2):241–251. <https://doi.org/10.1046/j.1365-2389.2002.00440.x>
- Maronna RA, Martin RD, Yohai VJ, Salibián-Barrera M (2018) Robust statistics: theory and methods (with R), 2nd edn. Wiley, Hoboken
- McCullagh P, Nelder J (1989) Generalized linear models, 2nd edn. Chapman & Hall, Boca Raton
- Montgomery DC, Peck EA, Vining GG (2012) Introduction to linear regression analysis, 5th edn. Wiley & Sons, Hoboken
- van den Boogaart KG, Filzmoser P, Hron K, Templ M, Tolosana-Delgado R (2021) Classical and robust regression analysis with compositional data. *Math Geosci* 53(3):823–858. <https://doi.org/10.1007/s11004-020-09895-w>
- Zhao X, Li W, Zhang M, Tao R, Ma P (2020) Adaptive iterated shrinkage thresholding-based lp-norm sparse representation for hyperspectral imagery target detection. *Remote Sens* 12(23):3991. <https://doi.org/10.3390/rs12233991>

Journel, André

R. Mohan Srivastava
TriStar Gold Inc, Toronto, ON, Canada

Biography

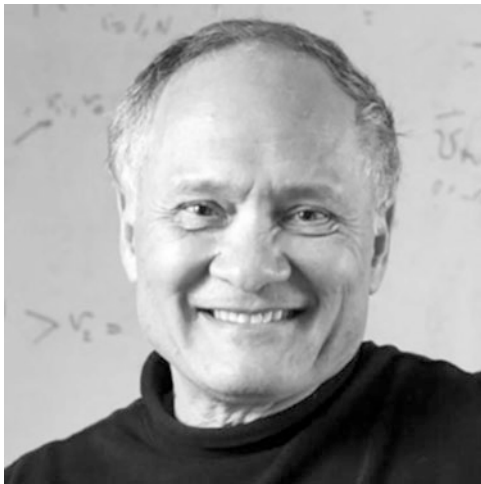


Fig. 1 André Journel (Courtesy Mohan Srivastava)

André Journel is a geostatistician, one of the first group of students to whom Georges Matheron taught his theory of regionalized variables. Educated as a mining engineer at the École Nationale Supérieure de Géologie in Nancy in the late 1960s, he joined the geostatistics research group at Fontainebleau in 1969 and worked there for a decade as a researcher, teacher, and consultant to mining companies around the world. His book *Mining Geostatistics*, co-authored with Charles Huijbregts in 1978, was the first comprehensive and rigorous reference book on the subject, the “bible” for geostatistics for many who explored the mysteries of the new science of spatial interpolation grounded in statistical theory.

In 1978, he became a professor at Stanford University, where he established one of the world’s preeminent geostatistics research programs. In the early 1980s, he moved the application of geostatistics into environmental sciences and championed the theory and practice of risk-qualified mapping. He worked with the US Environmental Protection Agency to create Geo-EAS, the first widely used public-domain software toolkit for geostatistical data analysis and interpolation. In the late 1980s, he and his students at Stanford pioneered many new methods for stochastic simulation, conditioned by hard and soft data. These new algorithms were quickly adopted by the international petroleum industry as the accepted approach for modeling rock and fluid properties in oil and gas reservoirs where correctly representing spatial heterogeneity in computer models of the subsurface is critical for accurate flow modeling and where risk analysis calls for a family of alternate scenarios, all honoring the same input data and user parameter choices.

The principal vehicle for André Journel’s prolific research through the last half of his active academic career was the Stanford Center for Reservoir Forecasting (SCRF), a consortium of industry, government research organizations, and software companies that he founded in 1986 and directed for more than 20 years.

By the time he retired from Stanford in 2010, he had served as the advisor and mentor to more than 50 graduate students and had been the catalyst for two more major public-domain software packages: GSLIB in the 1990s and SGeMS in the early 2000s. He guided the “Applied Geostatistics Series” published by Oxford University Press from its inception in the late 1980s through the early 2000s.

He has been the recipient of many awards and honors, including the Krumbein Medal from the International Association for Mathematical Geosciences (1989), the Earth Sciences Teaching Award at Stanford (1995), the Lucas Gold Medal from the Society of Petroleum Engineers (1998), and the Erasmus Award from the European Association of Geoscientists and Engineers (2012). In 1998, he was elected to the US National Academy of Engineering.

K-Means Clustering

Jaya Sreevalsan-Nair

Graphics-Visualization-Computing Lab, International
Institute of Information Technology Bangalore, Electronic
City, Bangalore, Karnataka, India

Definition

Clustering is a process of grouping n observations into k groups, where $k \leq n$, and these groups are commonly referred to as clusters. *k-means clustering* is a method which ensures that the observations in a cluster are the closest to the representative observation of the cluster. The representative observation is given by the centroid, i.e., the mean, of the observations belonging to the cluster. This is akin to Voronoi tessellation of observation in a d -dimensional space, if the observations are represented using d -dimensional feature vector. An optimal k -means clustering is one where it achieves minimum intra-cluster, i.e., within-cluster, variance that also implies maximum inter-cluster, i.e., between-cluster, variance. Given a set of observations $(x_1, x_2, \dots, x_n)$ that are distributed/partitioned into k clusters, $\mathbf{C} = \{C_1, C_2, \dots, C_k\}$, such that $k \leq n$, then the objective function for optimization is given by:

$$\arg \min_{\mathbf{C}} \sum_{i=1}^k \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg \min_{\mathbf{C}} \sum_{i=1}^k |C_i| \sigma^2(C_i),$$

where $\boldsymbol{\mu}_i$ is the mean of the observations in the cluster C_i and $\sigma^2(C_i)$ is its intra-cluster variance. Intra-cluster variance is defined as the variance of the position coordinates of the observations in a cluster, i.e., given by the sum of squared Euclidean distances of all observations in the cluster with all the others within the same cluster. Similarly, inter-cluster variance refers to the sum of the distance of the observations belonging to two different clusters.

Euclidean distance or the L^2 norm of the distance vector is conventionally used in k -means clustering. The use of Euclidean distances manifest as nearly spherical clusters. Thus, k -means clustering works the best for observations which are inclined to have spherical clusters. Other distance measures, such as Mahalanobis distance and cosine similarity, can also be used (Morissette and Chartier 2013). If the distance measure used is the Mahalanobis distance, then the covariance between the vectors of observations is considered. Similarly, if the distance measure is the cosine similarity, then correlation is used as the distance measure. Thus, the characterization of the distance measure influences the clustering outcomes.

Overview

The term *k-means* was first coined in the context of qualitative and quantitative analysis of large-scale multivariate data (MacQueen et al. 1967). The Lloyd-Forgy algorithm (Forgy 1965; Lloyd 1982) is the classical implementation that is widely used today and is also referred to as the naïve k -means. The algorithm, in both Lloyd-Forgy and MacQueen variants, comprises six key steps: (i) choose k , (ii) choose distance metric, (iii) choose method to pick centroids of k clusters, (iv) initialize centroids, (v) update assignment of membership of observation to closest centroid, and update centroids. Step (v) is implemented iteratively until convergence, which is determined by no change from previous iteration. Apart from convergence, the termination condition could also be imposed by forcing the number of iterations.

Now, the difference in both variants is in the manner of implementation of step (v). In the MacQueen algorithm, only observation is updated at a time, causing centroids to be updated, whereas in Lloyd-Forgy algorithm, the membership of all observations to the clusters is updated, and then the centroids are updated. Thus, MacQueen algorithm is an *online/incremental algorithm*, and Lloyd-Forgy algorithm is

an *offline/batch algorithm*. The former is efficient owing to frequent updates of centroids, whereas the latter is better suited for analyzing large datasets. The Lloyd and Forgy algorithms themselves differ in a minor point of considering the data distribution to be discrete and continuous, respectively. The Lloyd-Forgy algorithm is considered to be analogous to a spatial partition algorithm, specifically Voronoi tessellation.

Hartigan-Wong implementation, unlike the Macqueen and Lloyd-Forgy variants, focuses on achieving the objective function for optimization using the local minima as opposed to the global minimum (Hartigan and Wong 1979). It follows the aforementioned steps (i)–(v), followed by an additional step (vi) compute the within-cluster sum of squares of errors (SSE) to locally optimize. It means that when a centroid of a cluster, say C_i , is updated in the last iteration, the within-cluster SSE is compared with the within-cluster SSE if an observation currently in C_i were to be in another cluster C_j , such that ($i \neq j$). It does not guarantee a global optimum and hence, is a heuristic. Hartigan-Wong algorithm, like the Macqueen one, is sensitive to the order of change of cluster membership of the observations.

The pros of the basic iterative implementation of the k-means clustering algorithm include its relatively simple implementation and guaranteed convergence. But, its cons include the requirement of the number of clusters k as an input and the initialization of centroids prior to the iterative implementation. While the latter is conventionally done using random values, there are several methods which improve the initialization, e.g., using progressive methods, structured partitioning, a priori information, hierarchical clustering (Ward's method), etc. The k-means algorithm also suffers from the curse of dimensionality, as the ratio of the standard deviation to the mean of distance between observations decreases with increase in the number of dimensions. As the ratio decreases, the k-means clustering algorithm becomes increasingly ineffective in distinguishing between the observations. The curse of dimensionality can be alleviated using spectral clustering by projecting the observations to a lower dimensional subspace of k -dimensions, using principal component analysis, and then performing k-means in the k -dimensional space.

Owing to the popularity of the algorithm, several improvements have been implemented on the k-means algorithm. The effectiveness of the algorithm can be measured using internal and external criteria. Internal criterion is a metric that measures if the global optimum has been reached, and the external criterion measures the similarity of the current partition with a known partition, e.g., ground truth. Dunn index and Jaccard similarity are examples of internal and external criteria, respectively. k-means algorithm is an unsupervised learning method, as it does not require training data for pattern matching.

Applications

The pros of the algorithm satisfy its two significant goals, namely, exploratory data analysis and reduction of its complexity (Morissette and Chartier 2013). Geoscientific and related datasets have the characteristics of unknowns in the data and of being large-scale and complex. Thus, k-means clustering is widely used in different geoscientific domains and for different types of data analysis. The examples in seismology, flood extent change detection from synthetic aperture radar (SAR) images, and land cover change detection (LCCD) from satellite images are covered as samples of the specific problems in different domains that use k-means clustering. The examples in clustering for machine learning and visualization tool show how k-means clustering is a tool or a method integrated with a data analysis toolkit or application. Overall, these examples demonstrate how k-means clustering is one of the standard data analysis methods used in geoscience.

The k-means algorithm has been used in seismology for probabilistic seismic hazard analysis, where k-means is used to obtain a partition of earthquake hypocenters or alternatively a partition of fault ruptures (Weatherill and Burton 2009). The initialization of cluster centroids is done using ensemble analyses for identifying best choices. The k-means partitions of seismicity are then used as source models, whose representation in the regional seismotectonics is studied. The assessment shows that the k-means algorithm has given clusters that provide the most appropriate spatial variation in hypocentral distribution and fault type in the region. Similarly, k-means can be applied to features representing earthquake waveforms to study temporal patterns in low-magnitude earthquakes in geothermal fields. This indicates the usefulness of the k-means algorithm in solid Earth geosciences.

Change detection is done from multitemporal data, e.g., SAR and satellite images. Flood extent mapping has been done by change detection in SAR by using a combination of techniques using a two-step process (Zheng et al. 2013). Firstly, the difference image of SAR images of a region from two consecutive time-stamps is computed. After applying median and mean filters on the difference image, and the outcomes are combined, k-means algorithm is applied as a second step to cluster changed and unchanged regions. k-means algorithm has been found to be effective in capturing spatial variations in the difference image. Similar method can be applied for LCCD in remote sensed images, i.e., Landsat images (Lv et al. 2019). In addition to the aforementioned two steps, there is an additional step of adaptive majority voting (AMV) in local neighborhoods to identify the class type for land cover to analyze urban expansion, urban build-up changing, deforestation, etc. Majority voting is an ensemble analysis method. Overall, such integrated methods that include

efficient methods, such as the k-means algorithm, provide flexibility of applications and generality of change.

In addition to unsupervised learning applications, k-means algorithm has been used to reduce complexity in geoscientific data that is further visualized using appropriate metaphors (Li et al. 2015). For instance, geoscientific observation stations can be clustered for providing overview of the data, and these clusters are represented using simplified data visualization layouts, e.g., radial layout.

Future Scope

Geoscience and related datasets are now classified as big data, owing to the volume, velocity, variety, value, and veracity, i.e., the five Vs. Such complex data requires high-performance computing for its analysis. Hence, graphics processing unit (GPU)-based large-scale computations are implemented which entails hardware accelerated implementation of the algorithms, including k-means algorithm (Yu et al. 2019). The use of GPUs enable automatic hyperparameter tuning which allows the scalability of k-means algorithms to large datasets and also improving the potential of the algorithm for various applications. Parallel implementation of k-means has been used with real-world geographical datasets. Given the increasing complexity of the geoscientific and geographical data and metadata, there is potential for both hardware and software enhancements to be fulfilled, for the k-means algorithm.

Overall, k-means algorithm is an effective and efficient clustering algorithm that has been successfully used on geoscientific datasets.

Bibliography

- Forgy EW (1965) Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics* 21:768–769
- Hartigan JA, Wong MA (1979) Algorithm AS 136: a k-means clustering algorithm. *J R Stat Soc Ser C (Appl Stat)* 28(1):100–108
- Li J, Meng ZP, Huang ML, Zhang K (2015) An interactive radial visualization of geoscience observation data. In: *Proceedings of the 8th international symposium on visual information communication and interaction*, pp 93–102
- Lloyd S (1982) Least squares quantization in PCM. *IEEE Trans Inf Theory* 28(2):129–137
- Lv Z, Liu T, Shi C, Benediktsson JA, Du H (2019) Novel land cover change detection method based on K-means clustering and adaptive majority voting using bitemporal remote sensing images. *IEEE Access* 7:34,425–34,437
- MacQueen J et al (1967) Some methods for classification and analysis of multivariate observations. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Oakland, vol 1, pp 281–297
- Morissette L, Chartier S (2013) The k-means clustering technique: general considerations and implementation in Mathematica. *Tutor Quant Methods Psychol* 9(1):15–24
- Weatherill G, Burton PW (2009) Delineation of shallow seismic source zones using K-means cluster analysis, with application to the Aegean region. *Geophys J Int* 176(2):565–588
- Yu T, Zhao W, Liu P, Janjic V, Yan X, Wang S, Fu H, Yang G, Thomson J (2019) Large-scale automatic K-means clustering for heterogeneous many-core supercomputer. *IEEE Trans Parallel Distrib Syst* 31(5):997–1008
- Zheng Y, Zhang X, Hou B, Liu G (2013) Using combined difference image and k-means clustering for SAR image change detection. *IEEE Geosci Remote Sens Lett* 11(3):691–695

K-Medoids

Jaya Sreevalsan-Nair

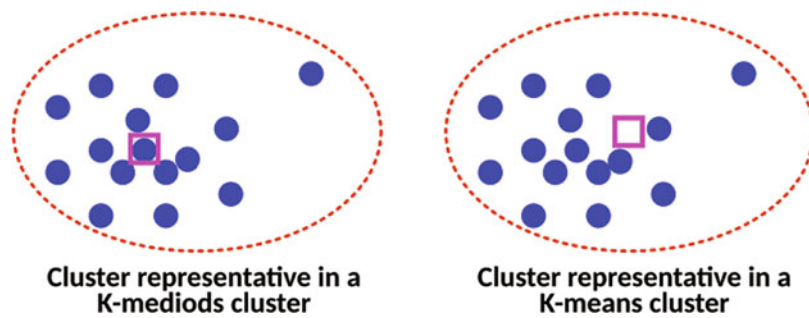
Graphics-Visualization-Computing Lab, International Institute of Information Technology Bangalore, Electronic City, Bangalore, Karnataka, India

Definition

A *medoid* is defined as a representative item in a dataset or its subset (or cluster), which is *centrally located* and has the least sum of dissimilarities with other items in the group (Kaufman and Rousseeuw 1987). Medoids are identified in a dataset to implement partitioning around medoids (PAM), which is a clustering method. Since PAM is used to generate K clusters, where K is a user-defined positive integer, such that $K > 1$, this method is also referred to as K-medoids clustering method (Kaufman and Rousseeuw 1987). A medoid is not equivalent to other statistical descriptors of median, centroid, or mean, but is the closest to median by virtue of an item in the dataset serving as a representative, as opposed to a computed entity. A median is identified by sorting items with respect to a single variable and picking the “middle” value of the variable, whereas the medoid is the item itself. Hence, medoids can be considered to be more generalized since it represents multivariate or multidimensional space, with flexibility in identifying them, based on the usage of appropriate dissimilarity or distance measures. Given that the rest of the article impresses upon geoscientific application, “item” will be referred to as “point” hereafter.

Overview

K-medoids clustering is often compared with the K-means one (*see the article on “K-Means Clustering”*), owing to the similarity in the algorithms. A key difference in both these partitioning algorithms is that the cluster representative or center is an actual point in the case of K-medoids method, i.e., the medoid itself (Kaufman and Rousseeuw 1987),



K-Medoids, Fig. 1 A graphical representation of the key difference in the cluster representative or the cluster center identified in (left) K-medoids and (right) K-means clustering algorithms. The magenta box indicates the cluster center, for a cluster of sites (blue circles)

indicated by its boundary (red dotted curve). (Image courtesy: Author's own, but adapted from a diagram in the article by Jin and Han (2011) https://doi.org/10.1007/978-0-387-30164-8_426)

whereas a centroid is computed from the points present in the K-means cluster. Traditionally, by design, another important difference between the two methods existed that revolved around the constraint of use of Euclidean distance in K-means, which was generalized to any similarity measure in K-medoids. Thus, the first implementation of PAM (Kaufman and Rousseeuw 1987) focused on introducing the use of dissimilarity measures, e.g., one minus Pearson correlation, to be a governing factor in the partitioning algorithm. Over time, Euclidean distance, one minus absolute correlation, and one minus cosine similarity have also become widely used dissimilarity measures (Van der Laan et al. 2003). K-medoids clustering problem is known to be NP-hard and hence has approximate solutions (Fig. 1).

The implementation of the original PAM has two steps, INITIAL and SWAP. INITIAL starts with D , a $n \times n$, dissimilarity matrix for n points, and M is a set of initial seed points of size K , used as medoids, for K clusters. Each point is associated with a seed point by virtue of its lowest dissimilarity with the seed point, i.e., its corresponding medoid. Then, SWAP is done iteratively which minimizes the optimizes M by minimizing the loss function, i.e., the sum of dissimilarities of each point and its corresponding medoid. In the SWAP step, each medoid is considered for a *swap of roles* with a non-medoid point, and the swap is implemented only when the recomputed loss function is minimized. This greedy algorithm terminates when SWAP does not find any more changes for implementation and is thus more efficient than an exhaustive search method.

The original PAM algorithm recommends determining a data-driven K value by maximizing the average silhouette. The silhouette is computed based on the outcome of the implementation of PAM with a random K value (Kaufman and Rousseeuw 1990). PAM has been alternatively solved using the loss function that is the sum of maximum dissimilarity between a point and its corresponding medoid (Kaufman and Rousseeuw 1990). This methodology has been referred to as the K-center model.

PAM has been extended to cluster large datasets using a sampling strategy. This variant of PAM is called CLARA (clustering large applications) (Kaufman and Rousseeuw 1990). CLARA involves two steps, where the first step involves finding a sample in the large dataset and implementing PAM on this sample or subset, and the second step entails adding the remaining points (outside of the sample) to the current set of medoids. After the clustering of the entire dataset is completed, the medoids can be recomputed, as required.

CLARANS (clustering large applications based on randomized search) has been a further improvement over PAM and CLARA, using an abstraction of a hypergraph (Ng and Han 2002). A graph is constructed in such a way that each node is a set of K objects, such that they can be potentially medoids, and two nodes are neighbors only when the cardinality of their set intersection is $K - 1$. Thus, a node is an abstraction of an instance of the partition or clustering itself. PAM is now implemented on this graph as a search for the node with the minimum cost by going from an initial random node to a neighbor with the deepest descent in costs. CLARANS has been proven to run efficiently on large databases too.

The attractive characteristic of PAM is in its flexibility of distance or similarity measures that can be used as per the semantics needed in the clustering. Medoids are specific cluster representatives, unlike the cluster centers used in similar partitioning methods are less sensitive to outliers (Van der Laan et al. 2003). In hierarchical clustering methods, the representative points used as medoids can be effectively used. This is owing to the completeness of data in all dimensions for the medoids, just as any other point in the data, whereas the computed cluster centers usually have the attributes only for the subset of dimensions used for the clustering process. While PAM does not find good clusters when the cluster sizes are unbalanced or skewed, this can be resolved by using the silhouette function as the cost function in the PAM algorithm (Van der Laan et al. 2003).

The implementation of PAM is still inefficient and has time complexity of $O(K^3 \cdot n^2)$ (Xu and Tian 2015), which is improved by implementing the distance matrix only once and iteratively finding new medoids, by running several methods to identify the initial medoids (Park and Jun 2009). This algorithm is referred to as a “K-means like” algorithm, which improves the efficiency of PAM. This faster algorithm has a time complexity of $O(K(n - K)^2)$, which is high compared to the time complexity of CLARA, which is $O(K^3 + nK)$, and to that of CLARANS, which is $O(n)$, which also has to include its characterization of a randomized algorithm.

Applications

This method is applicable to all datasets involving scattered points, which includes both geospatial locations and large collections or databases of definite objects, e.g., images, vector data (cartographic maps), and videos.

One such example of clustering spatial points is that of ARGO buoys in the ocean done using an improved K-medoids method for anomaly detection in the ocean data (Jiang et al. 2019). This entails including a density criterion in identifying the initial seed points for running the K-medoids algorithm. For each cluster, the points that do not belong to the dense regions of the cluster are identified as outliers. In such an application, it has been found that the density-based approach in the K-medoids algorithm has improved the results by a higher accuracy score and a lower false detection rate. Similar applications have been found in the spatial clustering of traffic accidents.

As an example of clustering of objects, different from geospatial locations, K-medoids have been used for clustering and classification of remotely sensed images of snowflakes acquired using a multi-angle snowflake camera (MASC) (Leinonen and Berne 2020). In this specific example, a generative adversarial network (GAN) is used for feature extraction from the images, and then, the feature vectors are clustered and classified using K-medoids and hierarchical clustering.

Future Scope

As observed, one of the key criticisms against the K-medoids algorithm is its high run-time complexity, which has been the scope of improvement in the current and future work on this method. There are several improvements that have already been proposed. One such example is in the parallel implementation of the K-medoids algorithm, PA-MAE (parallel k-medoids clustering with high accuracy and efficiency) (Song et al. 2017). PAMAE implements a two-phase method, namely, parallel seeding and parallel refinement, for improving accuracy and efficiency. These phases are implemented in

the form of MapReduce, which enables parallelism. Parallel seeding involves identifying a sample of points in the large dataset and conducting a global search over this subset of the points. The parallel refinement is in MapReduce that involves performing a local search in the entire dataset. The difference between local and global searches is in finding a new medoid within a considered group vs outside of the group, respectively. This method is designed to be implemented on Spark and Hadoop, which are both horizontally scalable big data frameworks.

Faster implementation of PAM has been achieved by relaxing the number and choice of swaps in the SWAP step, e.g., FastPAM (Schubert and Rousseeuw 2019). This strategy is flexible to be incorporated with variants of PAM implementations, such as CLARA and CLARANS algorithms. FastPAM works by removing redundancy in computations for finding the best swap in PAM and swapping for multiple medoids simultaneously, owing to the independence of clusters.

Another line of work on the K-medoid method is in incorporating it as a step in data science workflows, such as that for the classification of snowflake images (Leinonen and Berne 2020). Such integration is usually challenging when the method has to be coupled with another data mining method. In that regard, the flexibility built into the K-medoids clustering both in terms of choice of dissimilarity measure and initialization method for the seed points is useful in coupling it with other methods, such as feature extraction, feature ranking, etc.

In summary, K-medoids clustering is a partitioning algorithm that provides robust cluster representatives which have proven to be useful across several geoscientific applications. Thus, this method has been successfully used for more than three decades.

Cross-References

- [K-Means Clustering](#)
- [Neural Networks](#)

Bibliography

- Jiang H, Wu Y, Lyu K, Wang H (2019) Ocean data anomaly detection algorithm based on improved K-medoids. In: 2019 Eleventh international conference on advanced computational intelligence (ICACI). IEEE, pp 196–201. <https://doi.org/10.1109/ICACI.2019.8778515>
- Kaufman L, Rousseeuw PJ (1987) Clustering by means of medoids. In: Proceedings of the statistical data analysis based on the L1 norm conference, Neuchatel, Switzerland, pp 405–416
- Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. Wiley, New York
- Leinonen J, Berne A (2020) Unsupervised classification of snowflake images using a generative adversarial network and K-medoids classification. Atmos Meas Tech 13(6):2949–2964. <https://doi.org/10.5194/amt-13-2949-2020>

- 5194/amt-13-2949-2020. <https://amt.copernicus.org/articles/13/2949/2020/>
- Ng RT, Han J (2002) CLARANS: a method for clustering objects for spatial data mining. *IEEE Trans Knowl Data Eng* 14(5):1003–1016
- Park HS, Jun CH (2009) A simple and fast algorithm for K-medoids clustering. *Expert Syst Appl* 36(2):3336–3341
- Schubert E, Rousseeuw PJ (2019) Faster K-medoids clustering: improving the PAM, CLARA, and CLARANS algorithms. In: *Similarity search and applications (SISAP)*. Springer, Cham, pp 171–187. https://doi.org/10.1007/978-3-030-32047-8_16
- Song H, Lee JG, Han WS (2017) PAMAE: parallel K-medoids clustering with high accuracy and efficiency. In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, pp 1087–1096. <https://doi.org/10.1145/3097983.3098098>
- Van der Laan M, Pollard K, Bryan J (2003) A new partitioning around medoids algorithm. *J Stat Comput Simul* 73(8):575–584
- Xu D, Tian Y (2015) A comprehensive survey of clustering algorithms. *Ann Data Sci* 2(2):165–193. <https://doi.org/10.1007/s40745-015-0040-1>

K-nearest Neighbors

Jaya Sreevalsan-Nair
Graphics-Visualization-Computing Lab, International
Institute of Information Technology Bangalore, Electronic
City, Bangalore, Karnataka, India

Definition

The k-nearest neighbors (kNNs) generally refer to the kNN algorithm that was initially proposed as a nonparametric solution to the *discrimination problem* (Fix and Hodges 1951). The discrimination problem refers to that of determining if a random variable Z with an observed value z is distributed over a p -dimensional space according to parametric distributions F or G . There are three variants of the problem, where both F and G are (i) completely known; (ii) known, but parameters are unknown; and (iii) completely unknown. The seminal solution proposed for the variant (iii), now known as the kNN algorithm, uses a nonparametric approach. It is given that there are m samples, $\{X_i \mid 0 \leq i < m\}$ and n samples, $\{Y_i \mid 0 \leq i < n\}$, of F and G , respectively. The solution says that if the k -nearest neighbors of z include M X s and $N = (k - M)$ Y s, where $(k, M, N \in \mathbb{Z}^+)$, $(k < m)$, $(k < n)$, then using likelihood ratio discrimination, Z , can be assigned to F iff $\left(\frac{M}{m} > c \frac{N}{n}\right)$, where c is an appropriate positive constant. This is the nonparametric counterpart of the parametric condition used to solve variants (i) and (ii), i.e., $\left(\frac{f(z)}{g(z)} > c\right)$, using density functions f and g , of F and G , respectively. This proposed approach is now known as the *majority voting*.

Over time, the kNN algorithm has been reframed as a *classification method*, where, in the absence of the parametric class distributions of a set of labeled objects S , a new object

X is assigned a class label, i.e., the label of a majority of the objects in k objects from S , closest to X . Thus, a majority voting approach is implemented on a predetermined set of labeled local neighbors of X , $\mathcal{N}(X)$, where its set cardinality is k , for determining the class label of X . The closeness between objects is defined using pairwise *distance functions*, e.g., Euclidean distance. Given two objects X and Y , defined using attribute/feature vector $a \in \mathbb{R}^p$, the Euclidean distance is given by

$$d(X, Y) = \|a(X) - a(Y)\|_2 = \sqrt{\sum_{i=1}^p (a_i(X) - a_i(Y))^2}.$$

Given $\mathcal{N}(X) \subset S$, such that $\mathcal{N}(X) = \{Y_1, \dots, Y_k\}$ and the class labels set C , the majority voting gives the class label of X as

$$\begin{aligned} c(X) &= \arg \max_{c \in C} \sum_{i=1}^k \delta(c, c(Y_i)), \text{ where } \delta(c, c(Y_i)) \\ &= \begin{cases} 1, & \text{if } c = c(Y_i), \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Overview

The basic kNN algorithm using majority voting has been improved using several strategies (Jiang et al. 2007). The use of Euclidean distance leads to inaccuracies in the presence of irrelevant features, which is referred to as the *curse of dimensionality*. Hence, strategies such as the use of optimal feature subsets based on feature ranking or filtering, attribute weighting using mutual information, frequency-based distance measure, etc. have been successfully used to overcome the curse of dimensionality. Adaptive or dynamic kNN involves identification of an optimal k value for each object prior to running the kNN algorithm or other related methods.

The kNN algorithm is now extended beyond classification owing to the versatility of applications of the k-nearest neighbors (kNNs) themselves. The set of kNNs of an object X , $\mathcal{N}(X)$, can now be defined as its local neighborhood that can have different variants based on the choice of the distance function, which includes other distance measures, e.g., Hamming, Chebyshev distances, etc. The use of kNNs now include feature extraction, segmentation, regression, and other data mining processes. Generalizing the use of kNNs across all these applications entails applying appropriate (function) *map* on a specific property, q , of each of the local neighbors of X and subsequent *reduce* operation, i.e., accumulation functions, to determine the value of q at X . q is now generalized as class label, value of a feature in its hand-crafted feature vector, global or local descriptor, etc. In the basic kNN algorithm, the map is an identity function on class label, as q , and the majority function is used for accumulation. As an

example of generalized application, in LiDAR (light detection and ranging) point cloud processing, the position coordinates of 3D points serve as q , the covariance matrix between a 3D point and its neighbor is the map, and the matrix sum is a reduce operation, which gives the design matrix (Filin and Pfeifer 2005). This matrix serves as the local geometric descriptor of the 3D point, required for feature extraction for point cloud classification. The local neighborhood used for the LiDAR points can be of kNNs or of spherical, cuboidal, or cylindrical shape.

Applications

There are several geoscientific applications for which both kNN algorithm and kNNs have been successfully used. There has been a long history of classification problems pertaining to digital maps and satellite remote sensing image data for both pixel-based and image classification, for which kNN algorithm and its variants are useful. These examples here demonstrate that the kNN algorithm works effectively on different types of remote sensed images for domain-specific classification problems.

- As an example of pixel-based classification, delineation of forest and non-forest land use classes has been done using kNN algorithm to classify plot pixels on digital map data, routinely used in Forest Inventory and Analysis (FIA) monitoring (Tomppo and Katila 1991). Satellite imagery and field data from FIA have been used to improve the weighting distances used in kNN method. The field data also serves as ground truth. The classification is used for constructing maps for basal area, volume, and cover type of the forests. kNN has been widely used for forest map delineation as it is effective for mapping continuous variables, e.g., basal area, volume, and cover type. Another reason for quick adoption of the kNN algorithm by the FIA community has been its easy integration with existing systems and methods.
- In another example of pixel-based classification involving hyperspectral images (HSI), kNN algorithm with guided filter has been effective in extraction of the information of the spatial context as well as denoising classification outcomes using edge-preserving filtering (Guo et al. 2018). HSI classification involves using spectral features of pixels to classify them into different classes, e.g., different types of soil types, plant species, etc. The joint representation of kNN and guided filter has been found to work the best with the full feature space, without performing any dimensionality reduction.
- As an example of image classification, for that of infrared satellite images to six different cloud types, kNN classifier has been found to be more accurate than the neural network using self-organizing feature map (SOFM) classifier, when applied to different texture features (Christodoulou et al. 2003). kNN classifier with weighted averaging has

also performed better than the traditional majority voting for cloud satellite image classification.

Apart from classification, kNN regression has been widely used in geoscientific applications for prediction and estimation. As an example of regression, estimating ground vibrations in rock fragmentation is considered. Blasting is routinely used for rock fragmentation, and estimating the blast-induced ground vibration is an important problem to study the effect of blasting in the surrounding environment. A combination of kNNs and particle swarm optimization (PSO) has been effectively used for estimating the peak particle velocity (ppv) that is a measure of the ground vibration (Bui et al. 2019). Here, the kNNs are used as the specific observations that influence the output predictions in a regression model. The optimization of particle positions using the PSO algorithm enables improving the precision of the kNN regression.

kNNs are effective for value replacement in several geoscientific applications. An example is in the use of nearest neighbors in improving the analysis of the geochemical composition of soil. For $k = 1$, kNN approach is used to replace values screened of several elements (e.g., Al, C, Ca, Na, P, etc.) at sample sites (Grunsky et al. 2018). Usually, there are observations for several elements where the values are less than the lower limit of detection ($<LLD$), which is undesirable. For such cases, the data is adjusted using the values observed in the neighbors. This adjusted data is further used for multivariate statistical analysis of the soil in the study area.

Overall, the kNN algorithm and the data models using kNNs have helped in creating powerful data processing and mining tools.

Case Study: LiDAR Point Cloud Processing – For LiDAR point clouds, the shape of the local neighborhood of a point is a significant differentiator in classifying the point. The local neighborhood itself is determined by using specific constraints on search for the neighbors. The search query includes finding points present in a search volume or *support region* of a specific size, e.g., kNNs, or of a specific shape, e.g., spherical, cylindrical, cuboidal, etc. (Filin and Pfeifer 2005). A combination of constraints can be used for support region, e.g., kNNs of a specific k value within a spherical neighborhood of a specific radius. The search query itself is implemented using efficient spatial data structures, such as kd-trees, octrees, etc. While the direct use of the kNN algorithm is rare in LiDAR point cloud classification, kNNs are used for feature extraction (Weinmann et al. 2014).

Prior to feature extraction for classification, a local geometric descriptor is computed at each point using the selected local neighborhood. The descriptor upon eigenvalue decomposition contributes to the eigenvalue-based features for classification. It is now known that a single support region is not sufficient to obtain an optimal descriptor for classification, owing to the inherent uncertainty in the data. Hence, increasingly multiscale methods are used in LiDAR point cloud processing. The size of the support region is considered to

be the *scale*. Thus, for kNNs, the different k values are used as scales. Identifying an *optimal scale* from multiple scales and using the features computed at the optimal scale for classification have been found to improve classification accuracy (Weinmann et al. 2014). Identifying multiple scales is the most straightforward when using kNNs, as opposed to other types of support regions, as the different scales can be identified by using all positive integer values. The use of optimal neighborhood using kNNs, where optimality is identified based on global minimum of eigen-entropy, has been found to improve classification accuracy.

Segmentation and classification of 3D point cloud can be largely seen as variants of the clustering or the grouping problem. Thus, segmentation of LiDAR point clouds to extract localized objects, e.g., buildings, use kNNs to group the specific set of points for geometric reconstruction (Wang and Shan 2009). There are two types of edges that can be reconstructed in the buildings from the LiDAR point clouds, namely, the jump edges and the crease edges. The jump edge is a discontinuity in height values, and the crease edge is where two surfaces intersect. Jump edges are critical in segmenting the point cloud into surface patches that identify objects such as the ground, buildings, etc. kNNs are instrumental in identifying points on the jump edges and in the jump edge extraction. Thus, kNNs play a central role in different point cloud processing methods.

Future Scope

Given the advent of supervised learning and deep learning for classification, the use of kNN algorithm and kNNs in general has become indispensable in data mining and knowledge discovery in geoscientific datasets. Given the ease of implementation, integration, and enhancement of kNN algorithm, it will continue to be used for classification, regression, and several other data mining methods, applied on images, point clouds, grids, and other data types used in geoscience.

Bibliography

- Bui XN, Jaroontattanapong P, Nguyen H, Tran QH, Long NQ (2019) A novel hybrid model for predicting blast-induced ground vibration based on k-nearest neighbors and particle swarm optimization. *Sci Rep* 9(1):1–14
- Christodoulou CI, Michaelides SC, Pattichis CS (2003) Multifeature texture analysis for the classification of clouds in satellite imagery. *IEEE Trans Geosci Remote Sens* 41(11):2662–2668
- Filin S, Pfeifer N (2005) Neighborhood systems for airborne laser data. *Photogramm Eng Remote Sens* 71(6):743–755
- Fix E, Hodges JL (1951) Discriminatory analysis. Nonparametric discrimination: consistency properties. Project number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, pp 1–24
- Grunsky E, Drew L, Smith D (2018) Analysis of the United States portion of the North American soil geochemical landscapes project –

a compositional framework approach. In: *Handbook of mathematical geosciences*. Springer, Cham, pp 313–346

- Guo Y, Cao H, Han S, Sun Y, Bai Y (2018) Spectral-spatial hyperspectral image classification with k-nearest neighbor and guided filter. *IEEE Access* 6:18,582–18,591
- Jiang L, Cai Z, Wang D, Jiang S (2007) Survey of improving k-nearest neighbor for classification. In: *Fourth international conference on fuzzy systems and knowledge discovery (FSKD 2007)*, IEEE, vol 1, pp 679–683
- Tomppo E, Katila M (1991) Satellite image-based national forest inventory of Finland. *Int Arch Photogramm Remote Sens* 28(7–1):419–424
- Wang J, Shan J (2009) Segmentation of LiDAR point clouds for building extraction. In: *American Society for Photogrammetry and Remote Sensing annual conference*, Baltimore, pp 9–13
- Weinmann M, Jutzi B, Mallet C (2014) Semantic 3D scene interpretation: a framework combining optimal neighborhood size selection with relevant features. *ISPRS Ann Photogramm Remote Sens Spat Inf Sci* 2(3):181

Kolmogorov, Andrey N.

Frits Agterberg
Central Canada Division, Geomathematics Section,
Geological Survey of Canada, Ottawa, ON, Canada



Fig. 1 Andrey N. Kolmogorov (Author: Konrad Jacobs. Source: <https://opc.mfo.de/detail?photoID=7493>. CC BY-SA 2.0 de)

Biography

Andrey Nikolayevich Kolmogorov was born in 1903 in Tambov, 50 km from Moscow. His talent for mathematics showed at the age of 5 when he discovered a remarkable property of the integer numbers: $1=1^2$, $1+3=2^2$, $1+3+5=3^2$, $1+3+5+7=4^2$, etc. In 1920, he graduated from high school and went to Moscow State University. As a student, he constructed a Fourier series that diverges almost everywhere (Kolmogorov 1923). He graduated in 1925, obtained his

Doctor of Philosophy Degree in 1929, and became a professor at Moscow State University in 1931. Later, in addition to performing research and teaching, he occupied various administrative positions at this university. This included serving as Dean of the Department of Mechanics and Mathematics. In 1939, he became a full member of the USSR Academy of Sciences. From 1930 onward, he made several trips abroad to visit colleagues. He also participated in several scientific conferences organized outside the Soviet Union. His fruitful life ended in 1987.

Kolmogorov is probably best known for his book “Foundations of the Theory of Probability” that contains the widely accepted axioms which provided mathematical statistics with its sound basis (Kolmogorov 1933). He made numerous other contributions to mathematics and science. For example, Kolmogorov (1941a) contains the original theory on turbulence in fluids for large Reynolds’ numbers that later was seen as the earliest multifractal cascade model. A complete list of his 518 publications can be found in the *Annals of Probability* (1989, Vol. 17, No. 3). They include 119 contributions to the second edition of the *Large Soviet Encyclopedia* (in Russian). After 1964, relatively many of Kolmogorov’s publications were concerned with the teaching of mathematics and statistics in Russian high schools. His name is remembered in many theorems and statistical methods including the Chapman-Kolmogorov and Fisher-Kolmogorov equations and the Kolmogorov-Smirnov test.

Several times, Kolmogorov contributed to mathematical geoscience: Kolmogorov (1941b) discusses a modification of the central limit theorem to explain lognormality of grain-size distributions in sedimentary petrology. He was a mentor of Andrew Vistelius who is considered to be a father of mathematical geology (cf. Henley 2018). In 1975, I visited Vistelius when he was the director of the “Laboratory of Mathematical Geology” in Leningrad. During discussions in his large private apartment, Vistelius enumerated the invaluable help he had received from Kolmogorov who helped him in 1948 to become the youngest Doctor of Science in the USSR. Vistelius had arranged for me to meet with Kolmogorov in Moscow, but this visit could not take place for medical reasons. Two publications of Kolmogorov are related to projects on which Vistelius had been working as well. These are the geometric configuration of crystals in igneous rocks (Kolmogorov 1949a) and the deposition of strata in sedimentary basins (Kolmogorov 1949b).

Bibliography

Henley S (2018) Andrey Borisovich Vistelius. In: Daya Sagar BS, Cheng Q, Agterberg F (eds) *Handbook of mathematical geosciences – fifty years of IAMG*. Springer Nature, Cham

- Kolmogorov AN (1923) Une série de Fourier-Lebesgue divergente Presque partout. *Fundam Math* 4:324–328s
- Kolmogorov AN (1933) *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin; English ed (1950), Chelsea, New York
- Kolmogorov AN (1941a) Local structure of turbulence I incompressible viscous fluid for very large Reynolds’ numbers. *C R Acad Sci URSS (N S)* 30:301–305
- Kolmogorov AN (1941b) Über das logarithmisch normale Verteilungsgesetz der Dimensionen der Teilschen bei Zerstückelung. *C R Acad Sci URSS (N S)* 31:99–101
- Kolmogorov AN (1949a) On the problem of “geometrical selection” of crystals. *Dokl Akad Nauk SSSR* 65:681–684
- Kolmogorov AN (1949b) The solution of a probability problem related to the question of the formation of strata. *Dokl Akad Nauk SSSR* 65: 793–779. (*Amer Math Soc Transl No 53*, 1951)

Korvin, Gabor

B. S. Daya Sagar

Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, Karnataka, India



Fig. 1 Gabor Korvin, courtesy of Prof. Korvin

Biography

Professor Gabor Korvin, Applied Mathematician, Geophysicist, Petrophysicist, Historian. He was born in Hungary (1942), has M. Sc. in Applied Mathematics (1966, Univ. Nat. Sciences, Budapest, Hungary); Ph.D. in Geophysics (1978, Univ. Heavy Industries, Miskolc, Hungary); Graduate

Diploma in Islamic Studies (1998, Univ. New England, Armidale, Australia).

Between 1966 and 1985 he was exploration seismologist and software developer in the Hungarian Geophysical Institute, Budapest; in 1986–1991 Senior Lecturer, University of Adelaide, Australia; 1994–2016 (when he retired) Professor of Geophysics, King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia. At KFUPM he was Coordinator of the *Reservoir Characterization Research Group*. As Professor, he taught *Reservoir Characterization, Seismic Stratigraphy, Petrophysics & Well logging, Solid Earth Geophysics, Geoelectric Exploration, Reflection Seismology, Inverse Problems, Geostatistics*.

He has more than 80 papers. His book *Fractal Models in the Earth Sciences* (Amsterdam, Elsevier, 1992a) was internationally acclaimed. His publication on resistivity anisotropy in thinly laminated sand-shale received the *Best Petrophysics Paper in 2012 Award*. His main Research Interest lies in finding stochastic mathematical and physical models to describe sedimentary rocks. He used the most diverse fields of mathematics and physics to solve petrophysical problems, such as: Gaussian random fields (1973, 1977, 1981); effective field theory (1982, 2012); fractals (1992a); and Percolation Theory (1992b). In 2014 he worked out a new rock model to connect different petrophysical properties. For some 20 years he was Earth Sciences Editor of the *Arabian Journal of Science and Engineering* (AJSE, Dhahran, Saudi Arabia). On the occasion of his retirement he wrote a Review Paper for this Journal on his dreams about the stochastic approach to Petrophysics (2016). He also publishes in Theoretical Mathematics (*Number Theory, Combinatorics*) and on Cultural and Religious History.

Bibliography

- Korvin G (1973, 1977, 1981) Certain problems of seismic and ultrasonic wave propagation in a medium with inhomogeneities of random distribution. Parts I–III. *Geophysical Trans* 21:5–34, 24(Suppl 2):1–38, 28(1):8–19
- Korvin G (1982) Axiomatic characterization of the general mixture rule. *Geoexploration* 19(4):267–276
- Korvin G (1992a) *Fractal models in the earth sciences*. Elsevier, Amsterdam
- Korvin G (1992b) A percolation model for the permeability of kaolinite-bearing sandstones. *Geophysical Trans* 37(2–3):177–209
- Korvin G (2012) Bounds for the resistivity anisotropy in thinly-laminated sand-shale. *Petrophysics* 53(1):14–21
- Korvin G (2014) A simple geometric model of sedimentary rock to connect transfer and acoustic properties. *Arab J Geosci* 7(3): 1127–1138
- Korvin G (2016) Permeability from microscopy: review of a dream. *Arab J Sci Eng* 41(6):2045–2065

Krige, Daniel Gerhardus

Richard Minnitt

School of Mining Engineering, University of the Witwatersrand, Johannesburg, South Africa



Fig. 1 Professor Daniel Gerhardus Krige, 26 August 1919–12 March 2013 © International Association for Mathematical Geosciences 2013, Courtesy of Ansie Krige

Biography

Daniel G. Krige has global recognition in the mining industry as the pioneer of modern statistical methods of ore evaluation known as geostatistics. His surname describes geostatistical techniques of “kriging,” applied worldwide in exploration, ore evaluation, environmental, petroleum, hydrology, agriculture, and other disciplines. Born in Bothaville on 26 August 1919, Danie graduated as a mining engineer in 1938 at the University of the Witwatersrand. He worked on various gold mines until 1943 after which he joined the Government Mining Engineer’s Department where he pioneered statistical applications to mineral valuation and ore reserve estimation of gold mines in the 1950s (Krige 1951, 1952). The rapid development of the Free State and Klerksdorp goldfields meant that decisions about mining investments based on limited numbers of boreholes were made without assessing the risk of failure. It was this that stimulated Danie’s research into ore evaluation. Apart from the lognormal frequency distribution model devised by Herbert Sichel in South Africa (Sichel 1947, 1952), approaches to statistical mine evaluation were unknown. The statistical explanation of the notorious problem of conditional bias in routine ore block and reserve valuations, given by Krige (1951), led several gold mines to use elementary regression correction techniques based on what is now known as kriging. Basic geostatistical concepts

of “support,” “spatial structure,” “selective mining units,” and “grade-tonnage curves” were introduced by Krige (1952).

Following the translation of his 1951 publication on the statistical approach to mine evaluation, interest among French mining engineers in the new field of geostatistics grew, leading to the establishment of Le Centre de Géostatistique de l'École des Mines de Paris in Fontainebleau. The term “kriging,” coined by Georges Matheron, describes the mineral evaluation process leading to improved estimates and reduced financial uncertainty about mining investments. Danie implemented the first routine application of the kriging of ore reserves on two large gold mines belonging to Anglovaal in the early 1960s. His research and publication of some 90 technical papers on the practical application of regionalised variables and spatial evaluation of mineral resources and reserves, is as a result, deeply entrenched in the mining industry.

Danie's work in geostatistics continued until he retired from Anglovaal in 1981. Following retirement, he held a position as Professor of Mineral Economics at the University of the Witwatersrand until 1991 and continued to consult to the mining industry until 2010. His contributions were recognized by the University through the award of the D.Sc. (Eng.) degree in 1963, by three honorary degrees (from the universities of Pretoria, UNISA, and Moscow State Mining University), by many awards from the South African Institute of Mining and Metallurgy, the International Association for Mathematical Geosciences, Die Suid Afrikaanse Akademie vir Wetenskap en Kuns, the SME of the American Institute of Mining Engineers, the International APCOM Council in 1999, the University of Antofagasta in Chile, and by the South African President (Order for Meritorious Service, Class 1, Gold). He was a fellow and honorary life member of various technical societies locally and overseas, including the Royal Society of South Africa. In recognition of Danie Krige's distinguished contributions to engineering, he was elected a Foreign Associate of the United States National Academy of Engineering (NAE) Section 11, Earth Resources Engineering, in February 2010. Danie is the first South African to ever receive this award from the NAE.

Bibliography

- Krige DG (1951) A statistical approach to some basic mine valuation problems on the Witwatersrand. *J Chem Metall Min Soc S Afr* 52:119–139
- Krige DG (1952) A statistical analysis of some of the borehole values in the Orange Free State goldfield. *J Chem Metall Min Soc S Afr* 53:47–64
- Sichel HS (1947) An experimental and theoretical investigation of bias error in mine sampling with special reference to narrow gold reefs,

vol 56. Transactions of the Institution of Mining and Metallurgy, London, pp 403–473

Sichel HS (1952) New methods in the statistical evaluation of mine sampling data. *Bulletin of the Institution of Mining and Metallurgy*, London, pp 261–288

Kriging

Xavier Freulon and Nicolas Desassis
Mines ParisTech, PSL University, Centre de Géosciences,
Fontainebleau, France

Synonyms

Gaussian process regression; Kolmogorov-Wiener prediction; Spatial best linear unbiased prediction

Definition

Kriging is a geostatistical procedure to estimate the value of a variable of interest from its measurements at some sampling locations. Kriging weights are determined by two conditions: the value to be estimated and its estimator share the same average value; the variance of the estimation error, also called prediction variance, takes the smallest possible value.

Prediction based on kriging is equivalent to spatial best linear unbiased predictor (or BLUP) where, for a given covariance function, the predictor minimizes the mean-squared prediction error over all linear combinations of observation values. The support of the target value can be punctual or volumic (e.g., the grade of a block); in this latter case, the estimated value is the average value over the support and is computed from samples located inside or outside this volume.

Some History

In geosciences, most of the observed variables vary continuously in space showing high irregularity at the local scale and some trends with a larger scale. Viewed mathematically, the value of the property of interest is a function of its position that cannot be expressed by some simple and closed expression. A practicable approach is to regard such a property as a random variable and to treat statistically its variation in space. Matheron (1962) coined such properties as regionalized variables and geostatistics, as the application of probabilistic methods to such variables.

Geostatistics stems largely from mining, and its terminology retains a strong mining flavor: one of the tasks of the mining engineer is to select the blocks to be exploited. To simplify, a block deserves to be exploited only if the cost of its extraction and processing does not exceed the value of its metal content. In practice, the true grade of a block is not known before its exploitation, so that the selection is made on the basis of an estimated grade. Thus improving the estimation of ore concentration in rock and of recoverable mining resources from fragmentary information (i.e., scarce drill holes) is a key operational issue. At the beginning of the 1950s, this estimate was simply the average grade of the data belonging to the block or situated at its border. Krige (1951), studying exploitation data in goldfields of South Africa, observed that for high cutoffs the blocks selected by this way were on average less rich than expected. Mathematically, this expresses a conditional bias. To avoid this bias, Krige gave a weight to the average grade of the data situated in the block and the complementary weight to the average grade of the ore body, these weights being determined by linear regression. Matheron (1962) generalized Krige's regression by assigning a proper weight to each sample, these weights being determined so as to minimize the estimation variance under the condition that the weight sum to 1 (this condition simply expresses that the estimator is a weighted average of the data), and he called this method kriging in honor to Danie G. Krige.

The principle of building an estimator by minimization the estimation variance has been well known to statisticians since the works of Wiener (1950). The novelty of kriging was a practical day-to-day application of this principle in geosciences. Cressie (1990) gives a detailed discussion of the historical development of kriging, not only in geostatistics but also within other scientific disciplines, and proposes three key components of kriging: (i) the variable of interest is interpreted as a realization of a second-order random field; (ii) the mean is estimated (ordinary kriging and universal kriging) rather than assumed to be known (simple kriging); and (iii) the covariance function, or the variogram, is known and used to compute the kriging weights. In other words, kriging is a spatial best linear unbiased predictor.

Linear Geostatistics

The theory of kriging as it is usually presented appears in (Matheron 1962, 2019): a natural variable (e.g., a grade) distributed in space and whose behavior presents a large complexity of detail is interpreted as a realization of a random function. The regionalized variable and its random model are noted, respectively, $z(x)$ and $Z(x)$, where x denotes a point in a d -dimensional space, with d generally equal to 2 or 3. The aim

of kriging is to predict the target z_0 , which is the value $z(x_0)$ of the regionalized variable z at an unobserved point x_0 or more generally the average value $z(v)$ of z in a given cell or block v

$$z(v) = \frac{1}{|v|} \int_v z(x) \, dx.$$

The kriging value of z_0 given N observations of the variable at locations x_α , $\alpha = 1, 2, \dots, N$, with values $z_\alpha = z(x_\alpha)$, is by definition a weighted average of the observations

$$z_0^* = \sum_{\alpha=1}^N \lambda_\alpha z_\alpha.$$

The values of the regionalized variable are unknown, except at the data locations. The estimator properties are thus evaluated on all the possible outcomes of the random model in order to take into account the uncertainties resulting from a partial observation of the actual phenomenon. Later on, the kriging estimator or predictor is expressed in terms of the probability model

$$Z_0^* = \sum_{\alpha=1}^N \lambda_\alpha Z_\alpha.$$

Working in the framework of a second-order random function model, the random field is characterized by the mean (also called the drift), $m(x) = E\{Z(x)\}$, and the covariance function $C(x, x') = \text{Cov}(Z(x), Z(x'))$ or the variogram $\gamma(x, x') = \frac{1}{2}E\{(Z(x) - Z(x'))^2\}$. The weights are chosen so as to ensure that:

- The estimated value and its estimator have the same average value $E\{Z_0^* - Z_0\} = 0$ (i.e., the condition for an unbiased predictor).
- The variance of the estimation error $\text{Var}(Z_0^* - Z_0)$ be minimal (i.e., the condition for an optimal predictor).

The covariance function or the variogram is assumed known, but the stationarity of the model is not required to derive the estimators. Assumptions on the mean and the covariance function define various methods of kriging, and they have been extended in several directions. One deals with broader forms of nonstationarities (Matheron 1973; Lindgren et al. 2011), and another one considers vector-valued or even function-valued random fields (Wackernagel 2013; Menafoglio et al. 2013).

Ordinary Kriging (OK)

Ordinary kriging assumes that the mean of the random function m is constant but unknown. To ensure the unbiased condition, the kriging weights λ_α^{OK} have to sum to 1. Hence choosing the weights minimizing the variance of the estimation error $Z_0^* - Z_0$ with the condition on their sum leads to the ordinary kriging system (Chilès and Delfiner 2012, p. 163). It is a linear system of $N + 1$ equations with $N + 1$ unknowns (the N weights λ_α^{OK} and a Lagrange parameter μ)

$$\begin{cases} \sum_{\beta=1}^N \lambda_\beta^{OK} \sigma_{\alpha\beta} + \mu = \sigma_{\alpha 0} \\ \sum_{\beta=1}^N \lambda_\beta^{OK} = 1 \end{cases}$$

where $\sigma_{\alpha\beta}$ is the covariance of the observations Z_α and Z_β and $\sigma_{\alpha 0}$ is the covariance of Z_α and the target Z_0 . The variance of the estimation error, called the ordinary kriging variance, is given by the expression

$$\sigma_{OK}^2 = E\{Z_0^* - Z_0\} = \sigma_{00} - \sum_{\alpha=1}^N \lambda_\alpha^{OK} \sigma_{\alpha 0} - \mu$$

where σ_{00} is the variance of Z_0 . Because the mean m is not involved in ordinary kriging, it is possible to extend ordinary kriging to a more general random function model, the intrinsic random function model, characterized by

$$\begin{aligned} E\{Z(x+h) - Z(x)\} &= 0 \\ E\{(Z(x+h) - Z(x))^2\} &= \gamma(h) \end{aligned}$$

The variogram $\gamma(h)$ summarizes the spatial variability of the random function. Geostatistics provides a set of consistent tools for choosing the variogram model adapted to a particular situation (Chilès and Delfiner 2012). The above OK system and OK variance remain valid provided that the covariance function is formally replaced by the opposite of the variogram, $-\gamma(h)$, in the equations defining the kriging system. This is the framework where kriging is widely used, especially in mining applications.

Drifts and Residuals

Universal Kriging (UK)

The assumption of a constant mean even if unknown can be a limitation for the application of kriging to phenomena displaying a trend. However, kriging can be generalized to random functions whose drift $m(x)$ is a linear combination of deterministic known functions

$$m(x) = \sum_{l=0}^L a_l f^l(x)$$

where the a_l are unknown coefficients and the $f^l(x)$ are the $L + 1$ drift functions. The kriging estimator, renamed as universal kriging, remains of the form $Z_0^* = \sum_\alpha \lambda_\alpha^{UK} Z_\alpha$, but, because the a_l are not known, unbiasedness is ensured only under the $L + 1$ constraints

$$\sum_{\alpha=1}^N \lambda_\alpha^{UK} f_\alpha^l = f_0^l, \quad l = 0, \dots, L$$

where $f_\alpha^l = f^l(x_\alpha)$ and f_0^l is $f^l(x_0)$. The minimization of the estimation variance leads to a UK system similar to the OK system except that there are now $L + 1$ constraints instead of a single one and as many Lagrange parameters.

Simple Kriging

The simple kriging assumes that the drift of the random function, $m(x)$, is known (it can be either constant in the stationary case or any deterministic spatial function), and the unknown variable to be predicted is the residual between the field value and its mean $Z(x_0) - m(x_0)$ (when a block value is to be estimated the residual is computed with the average of the mean over a block, $Z(v) - m(v)$). The estimator of the residual is

$$(Z(x_0) - m(x_0))^* = \sum_{\alpha=1}^N \lambda_\alpha^{SK} (Z_\alpha - m_\alpha)$$

and the simple kriging estimator of the variable of interest writes

$$Z(x_0)^* = m(x_0) + \sum_{\alpha=1}^N \lambda_\alpha^{SK} (Z_\alpha - m_\alpha).$$

The condition on the mean value is fulfilled by construction, and the minimization of the estimation variance leads to the simple kriging system

$$\sum_{\beta=1}^N \lambda_\beta^{SK} \sigma_{\alpha\beta} = \sigma_{\alpha 0}$$

for $\alpha = 1, \dots, N$ and to the simple kriging variance

$$\sigma_{SK}^2 = E\{Z_0^* - Z_0\} = \sigma_{00} - \sum_{\alpha=1}^N \lambda_\alpha^{SK} \sigma_{\alpha 0}.$$

In practical applications, to know the drift is a strong assumption, hence simple kriging receives limited applications. From a theoretical point of view, however, it is important because it has nice properties not shared by ordinary or universal krigings.

Additivity Relationship

Ordinary kriging, and more generally universal kriging, provides a predictor of the target Z_0 without having to estimate the drift. However, it does implicitly estimate the drift, and it is equivalent to optimum estimation of the drift followed by a simple kriging of the residuals. For this, a similar approach of kriging can be applied to estimate the drift: a linear estimator is formed, $m^*(x_0) = \sum_{\alpha=1}^N \lambda'_\alpha Z_\alpha$, with the weights selected so that

- $E\{m^*(x_0) - m(x_0)\} = 0$ (unbiased condition).
- $Var(m^*(x_0) - m(x_0))$ is minimal (optimal condition).

The resulting system coincides with a universal kriging system in which all the covariances between the data and the target on the right-hand side of the system are null. This estimate of the drift can be regarded as generalized least squares with correlated residuals. Using this optimal estimate of the drift, m^* , universal kriging can be rewritten as

$$Z^{UK}(x_0) = m^*(x_0) + \sum_{\alpha=1}^N \lambda_\alpha^{SK} (Z(x_\alpha) - m^*(x_\alpha)).$$

The additivity relation on the kriging value extends to the kriging variance with

$$\sigma_{UK}^2 = \sigma_{SK}^2 + Var(m^*).$$

This decomposition can be exploited from a numerical point of view to efficiently solve the kriging system: the universal kriging can be expressed as simple kriging, followed by a drift correction, which appears as the solution of a linear system of $L + 1$ equations with $L + 1$ unknowns, whose matrix is positive definite. It is thus advantageous to exploit this result to solve the simple kriging system and the drift correction system both by the Cholesky method. In matrix notation, the universal kriging system is

$$\begin{Bmatrix} \Sigma & \mathbf{F} \\ \mathbf{F}^t & 0 \end{Bmatrix} \times \begin{Bmatrix} \lambda^{UK} \\ \mu \end{Bmatrix} = \begin{Bmatrix} \sigma_0 \\ f_0 \end{Bmatrix},$$

where Σ is the $N \times N$ matrix whose elements are the covariances between the observations, σ_0 is the vector of covariances between the N observations and the target, $\mathbf{F}$ is the

$N \times (L + 1)$ matrix whose (α, l) element is $f^l(x_\alpha)$, and f_0 is the vector of the $L + 1$ drift function values at the target location. Hence, the kriging system can be solved to compute the kriging weights λ^{UK} and the Lagrange parameters μ with the following procedure, which requires the resolution of two subsystems with positive definite matrices (Chilès and Delfiner 2012, p. 170):

1. Solve $\Sigma \lambda^{SK} = \sigma_0$ and $\Sigma w = F$.
2. Compute $S = F^t w$ and $R = F^t \lambda^{SK} - f_0$.
3. Solve $S \mu = R$ (S is positive definite).
4. $\lambda^{UK} = \lambda^{SK} - w \mu$.

Some Properties of Kriging

Consistency with Data Points

The kriging is an exact estimator: if the target coincides with an observation, the observed value is obviously a linear estimator, and its kriging variance is equal to zero. Thus the kriging is simply the observed value.

Smoothing Relationship

As kriging minimizes the variance of estimation error, it can be interpreted as a projection on a subset of the linear combinations of the observations fulfilling the unbiased conditions (Journel 1977). When the mean is known, the simple kriging and its kriging error are orthogonal, i.e., noncorrelated, and the Pythagorean theorem implies the smoothing relationship

$$Var(Z_0) = Var(Z_0^*) + \sigma_{SK}^2.$$

The simple kriging estimate is therefore less dispersed than the data. When the mean is unknown, the smoothing relationship can be generalized to

$$Var(Z_0 - m^*(x_0)) = Var(Z_0^* - m^*(x_0)) + \sigma_{UK}^2.$$

When observations are sparse, the estimated maps using kriging are smooth and may look unrealistic. These negative aspects of kriging have fostered the development of stochastic simulations. Incidentally, the orthogonality between the simple kriging and the estimation error is used to derive an algorithm to turn a nonconditional simulation into a conditional one.

Kriging as an Interpolator

The kriging value can be expressed as the sum of the optimal drift and a linear combination of the covariance functions (Matheron 2019)

$$Z^*(x) = \sum_{\alpha=1}^N b_{\alpha} C(x_{\alpha} - x) + \sum_{l=0}^L c_l f^l(x)$$

where the coefficients $\mathbf{b} = [b_{\alpha}]_{\alpha \in 1 : N}$ and $\mathbf{c} = [c_l]_{l \in 0 : L}$ are the solution of the dual kriging system

$$\begin{Bmatrix} \Sigma & \mathbf{F} \\ \mathbf{F}^t & 0 \end{Bmatrix} \times \begin{Bmatrix} \mathbf{b} \\ \mathbf{c} \end{Bmatrix} = \begin{Bmatrix} \mathbf{z} \\ \mathbf{0} \end{Bmatrix}$$

Block Kriging and Area-to-Point Kriging

The original development of kriging for mining aimed at estimating the average grades over the volume of selection, $Z(V)$, as the optimal linear combination of the measured grade of assays or cores, $(Z(v_{\alpha}))_{\alpha = 1, \dots, N}$. In the mining context, the support of the observations is usually constant and smaller than the support of the variable to be estimated (e.g., cores are smaller than the mining blocks); hence observations are assumed to be punctual and their covariance function, or variogram, known. In this case, the change of support from the points to block affects only the target and defines the block kriging. As kriging is a linear estimator and the transform from punctual grade to average grade is linear too, the block kriging system is very similar to the punctual kriging, only the right-hand side of the system has to be changed using the covariance values between the data points and the volume of the target

$$C_{V,\alpha} = \text{Cov}\left(Z(V), Z(x_{\alpha})\right) = \frac{1}{|V|} \int_V \text{Cov}(Z(x), Z(x_{\alpha})) dx$$

and the average value of the drift functions over the target block satisfies

$$f_V^l = \frac{1}{|V|} \int_V f^l(x) dx, \quad l = 0, \dots, L.$$

The expression of the kriging variance has to be adjusted too with the variance of the block variable. In the context of geographical applications, Kyriakidis (2004) formulates the reverse procedure of area-to-point interpolation where kriging with areal data is used to predict point values. The linear properties of kriging imply to modify only the left-hand side of the kriging system using the covariance between the observed areal data. Both are special cases of kriging where both the data and the target are defined on their own support. Kriging as a linear predictor is designed to handle such configurations if the covariance function, or the variogram, of the punctual random function is known. It only requires the computation of the average covariance between observations defined on different supports (for instance via quadrature rule)

$$\begin{aligned} C_{v_{\alpha}, v_{\beta}} &= \text{Cov}(Z(v_{\alpha}), Z(v_{\beta})) \\ &= \frac{1}{|v_{\alpha}| \times |v_{\beta}|} \int_{v_{\alpha} \times v_{\beta}} \text{Cov}(Z(x), Z(x')) dx dx' \end{aligned}$$

and the average value of the drift functions over the observation support

$$f_{v_{\alpha}}^l = \frac{1}{|v_{\alpha}|} \int_{v_{\alpha}} f^l(x) dx, \quad \alpha = 1, \dots, N.$$

Filtering with Kriging

The factorial kriging (Matheron 1982) considers an orthogonal decomposition of the variable of interest of the form $Z(x) = m(x) + \sum_k a_k Y_k(x)$, where the a_k are known coefficients and the random fields $Y_k(x)$ are orthogonal with zero mean and covariance functions C_k . In practice, the experimental variogram of the variable Z is fitted by a positive combination of covariance functions $C = \sum_k a_k^2 C_k$ with the assumption that each structure represents some heuristic factors (e.g., acquisition or processing artifacts in the case of a seismic survey). This model defines a filtering procedure extracting each factor by (co-)kriging, $Y_k^*(x_0)$, with the kriging system derived by minimizing the variance of the estimation error $\text{Var}(Y_k^*(x_0) - Y_k(x_0))$ with the no-bias condition $E\{Y_k^*(x_0)\} = 0$

$$\begin{cases} \sum_{\beta=1}^N \lambda_{\beta} C(x_{\alpha} - x_{\beta}) + \sum_{l=0}^L \mu_l f^l(x_{\alpha}) &= a_k C_k(x_0 - x_{\beta}) \\ \sum_{\beta=1}^N \lambda_{\beta} f^l(x_{\beta}) &= 0 \end{cases}$$

These kriging estimates automatically satisfy the consistency relation

$$Z^*(x_0) = m^*(x_0) + \sum_k a_k Y_k^*(x_0)$$

where $m^*(x_0)$ is the optimal estimate of the drift $m(x_0)$. This approach is commonly used to filter seismic and remote sensing surveys with multivariate data sets. A linear model of coregionalization has to be fitted using a least squares method (Desassis and Renard 2013) or via spatial factor analysis, such as min/max autocorrelation factors (Switzer and Green 1984).

Kriging for Large Data Sets

As seen in previous sections, solving the kriging system with N observations involves the inversion of a square covariance

matrix, which requires $O(N^3)$ operations and $O(N^2)$ memory. This computational intractability when N is only moderately large constitutes hurdles that geostatistics had to face since the origins of kriging, even with expensive and scarce spatial data (e.g., drill holes for mining resource or oil reserve estimations). These direct data are now usually complemented by remote sensing and indirect geophysical measurements, creating massive and heterogeneous spatial databases. As any increase of computing resources is usually balanced by some increase of the size of the data sets, strategies had to be developed in order to compute kriging in operational conditions.

Local Approach and Moving Neighborhoods

A local approach, or kriging with a moving neighborhood, was first developed, considering only a relatively small number of points that lie close to the target point. Selecting neighboring data and ignoring other data is justified in the case where these other data would have received a zero or negligible kriging weight. Sophisticated algorithms have been designed for the selection of the neighbors (Renard and Yancey 1984); these strategies try to reach a compromise between near and far sample points and data in various directions around the target location, taking into account two competing effects: the screen effect with data being screened out by the presence of closer data and the relay effect induced by the mutual correlation between data. The relay effect may be exacerbated in the case of ordinary or universal krigings, as significant weights can be attributed to data points far from the target by the implicit estimation of the drift. In such a case, when the target moves, the neighborhood moves too, and the outermost data of the neighborhood may suddenly leave the neighborhood. This is likely to create discontinuities, in particular when the dropped value is extreme. Modification of the kriging system has been proposed to ensure the continuity with moving neighborhood (Rivoirard and Romary 2011). However large data sets allow for complex and nonstationary modeling, whose complex correlation structure is not readily captured by a local neighborhood. Therefore, specific geostatistical methods have been developed for global analysis of large data sets.

Global Approach

Global approaches either approximate or select a covariance model so that computations be tractable. They rely on sparse representation of the covariance matrix or of its inverse, the precision matrix. Once the kriging system is sparse, it can be solved using the efficient algorithms of the sparse linear algebra. A first option to make the covariance matrix sparse is the covariance tapering introduced by Furrer et al. (2006). The initial covariance function is tapered beyond a certain range using a positive definite but compactly supported

function, introducing a lot of zeros into the kriging matrix. The kriging predictor can then be calculated efficiently using sparse matrix inversion algorithms. Another option to reduce this complexity is the low-rank kriging proposed by Cressie and Johannesson (2008). It represents $Z(x)$ as a linear combination of L given basis functions $f^l(x)$ with random coefficients η_l , plus a white noise $\varepsilon(x)$, whose constant variance equals to σ^2

$$Z(x) = \sum_{l=1}^L \eta_l f^l(x) + \varepsilon(x).$$

The basic functions are usually chosen so as to represent several scales of variation and, for each scale, to cover the whole study domain. Denoting by $\mathbf{F}$ the $N \times L$ matrix whose (α, l) element is $f^l(x_\alpha)$, by Σ_0 the covariance matrix of the random coefficients η_l , the covariance matrix of the kriging system is

$$\Sigma = \sigma^2 \mathbf{I} + \mathbf{F} \Sigma_0 \mathbf{F}'$$

It can be inverted using the Woodbury matrix identity

$$\Sigma^{-1} = \frac{1}{\sigma^2} \left[\mathbf{I} - \frac{1}{\sigma^2} \mathbf{F} \left(\Sigma_0^{-1} + \frac{1}{\sigma^2} \mathbf{F}' \mathbf{F} \right)^{-1} \mathbf{F}' \right].$$

Within this model, kriging becomes tractable even with a very large number of data.

Finally the Markov random field approximation (Rue and Held 2005) is the opposite of that of covariance tapering in the sense that it seeks to make the inverse of the covariance matrix and not the covariance matrix itself sparse. This strategy to work directly on the precision matrix can be extended further by rephrasing the kriging problem in terms of stochastic partial differential equations.

Stochastic Partial Differential Equations (SPDE)

The SPDE framework proposed by Lindgren et al. (2011) allows to efficiently compute kriging with a very large number of observations while taking into account non-stationarities of the second-order structure. The starting point is a result of Whittle (1954) stating that a stationary random function Z with Matern covariance function can be written as the solution of the following SPDE

$$(\kappa^2 - \Delta)^{\alpha/2} Z(x) = \mathcal{W}(x)$$

where Δ is the Laplacian in $\mathbb{R}^d$, $\mathcal{W}$ is a white-noise with suitable variance, and κ^2 and α are positive values

respectively related to the scale parameter and the regularity parameter of the Matern covariance function. The principle of the SPDE approach is to compute a discretized version of the solution on the vertices of a meshing by using the finite element method. By this way, a random vector $\mathbf{Z}$ with covariance matrix Σ is obtained. Σ is a good approximation of the matrix which would be obtained by applying the covariance function to the distances between the vertices. When α is an integer, the precision matrix $Q = \Sigma^{-1}$ can directly be computed from the finite element method, and it is a very sparse matrix. By considering a meshing with vertices including the data locations, the kriging estimator at the other vertices can be obtained by the following way. Let us first consider the vector

$$\mathbf{Z} = \begin{Bmatrix} Z_T \\ Z_D \end{Bmatrix},$$

where Z_D is the sub-vector corresponding to the data locations and Z_T the one corresponding to the target ones. Σ and the associated precision matrix Q can be written by blocks:

$$\Sigma = \begin{Bmatrix} \Sigma_{TT} & \Sigma_{TD} \\ \Sigma_{DT} & \Sigma_{DD} \end{Bmatrix} \quad Q = \Sigma^{-1} = \begin{Bmatrix} Q_{TT} & Q_{TD} \\ Q_{DT} & Q_{DD} \end{Bmatrix}$$

where $\Sigma_{DD} = \text{Cov}(Z_D, Z_D)$ is the covariance matrix corresponding to the data, $\Sigma_{TD} = \Sigma_{DT}^T = \text{Cov}(Z_T, Z_D)$ are the covariances between data and targets, and $\Sigma_{TT} = \text{Cov}(Z_T, Z_T)$. With these matrix notations, simple kriging (with mean 0) is obtained by

$$Z_T^* = \Sigma_{TD} \Sigma_{DD}^{-1} Z_D,$$

which can be rewritten with the precision matrix blocks

$$Z_T^* = -Q_{TT}^{-1} Q_{TD} Z_D.$$

The kriging over the meshing can be obtained with the inversion of the sparse matrix Q_{TT} . Its size does not depend on the number of observations but only on the number of unknown vertices. This task can be efficiently performed by sparse linear algebra routines developed to solve partial differential equations. For instance, the Cholesky decomposition of Q_{TT} can be used for moderate size problems (mostly in 2D). For larger problems, the system becomes too large for direct approaches, and some iterative methods have to be used, for instance, the conjugate gradient algorithm. The nonstationary case can easily be handled by replacing the Laplacian operator Δ with the spatially varying operator $\nabla \cdot H(x) \nabla$ where ∇ is the gradient operator of $\mathbb{R}^d$ and $H(x)$ is a continuously varying

tensor taking into account the local anisotropies (Fuglstad et al. 2015). These random fields on $\mathbb{R}^d$ with locally varying anisotropies can also be interpreted as random field defined on a Riemann manifold. Finally time is becoming pervasive in geosciences applications (e.g., remote sensing provides regular snapshots, and automatic ground stations register fine resolution chronicles). From kriging's point of view, these space-time data sets combine three challenges: the modeling of the spatio-temporal dependencies, the modeling the evolution of a phenomenon with trends, and the handling of huge data sets with peculiar structures. The SPDE approach has the flexibility to handle these issues and is an active field of research (Cameletti et al. 2013).

Limitations of Kriging and Alternative Techniques

Practitioners have argued that kriging suffers from limitations, such as its smoothing effect, its sensitivity to extreme values, its inability to take into account constraints on the variables, and its weights depending only on the covariance and the data location but not on the actual observed values. From a probabilistic point of view, the conditional expectation is the optimal predictor defined for a given spatial model. As kriging computes the conditional expectation of the Gaussian field, it is optimal as long as the spatial phenomenon of interest is adequately described by such a model. In other cases (e.g., indicator or positive variables), kriging becomes less relevant. An alternate model should be considered first; then its actual conditional expectation computed. These real issues have induced fruitful developments of more advanced techniques such as nonlinear geostatistics and stochastic simulations for which kriging often remains a key ingredient.

Nonlinear Geostatistics and Multi-Gaussian Kriging

Nonlinear methods like indicator kriging, disjunctive kriging, or conditional expectation have been developed to handle heavy-tailed distributions or to map the risks to exceed a given threshold and more generally any nonlinear transform of the observed variable (Chilès and Delfiner 2012). A model widely used in practice is the transformed Gaussian model as its conditional expectation has a closed form for any transform of point or block value. The observed variable is supposed to be a transform of a stationary Gaussian field, with a null mean and a unit variance, $Z = \phi(Y)$. This model is completely specified by the anamorphosis function, ϕ , and the correlation function of the Gaussian values, $p(h) = E\{Y(x)Y(x+h)\}$. These parameters can be easily inferred from the original data when the distribution is continuous: firstly the anamorphosis function, strictly increasing, is inverted to assign to each observed value its Gaussian counterpart; then

the empirical variogram of the Gaussian variable is computed by the method of moments and then fitted with a model by nonlinear least squares.

Gaussian fields enjoy two useful and unique properties: They are stable by linear combinations (i.e., any linear combinations of the Gaussian values is a Gaussian variable), and the absence of correlation and independence is equivalent. This property implies that the simple kriging $Y^{SK}(x_0)$ and its kriging error $Y(x_0) - Y^{SK}(x_0)$, both linear combinations of Gaussian values, form a Gaussian pair without correlation, thus independent. Hence the target variable can be rewritten as the sum of two independent Gaussian fields

$$Y(x) = Y^{SK}(x) + [Y(x) - Y^{SK}(x)].$$

This decomposition is used to compute the conditional expectation, also called the multi-Gaussian kriging (Verly 1983), of any transform of the original variable, ψ . At a target point, the Gaussian variable is expressed as the sum of the kriging value and a normal residual scaled by the standard deviation of the kriging error, $Y(x_0) = Y^{SK}(x_0) + \sigma_{SK}U$. Then nonlinear predictor is calculated using the following formula (Emery 2005)

$$\{\psi \circ \phi(Y(x_0))\}^{CE} = \int_{-\infty}^{+\infty} \psi \circ \phi(Y^{SK}(x_0) + \sigma_{SK}u) g(u) du.$$

Efficient methods are available to compute the integral value such as Gauss quadrature rules or the Hermite decomposition.

Conditional Simulations

When the variable of interest T is a nonlinear functional of the random field $\phi(Y)$ over the domain (e.g., the maximum of the field in an area, the volume of a component connected to a point, or the probability that the average over a block exceeds a threshold), the conditional expectation is still the optimal predictor, but it has generally not a closed form. However, it can be approximated by the mean of evaluations of the functional over a set of conditional simulations

$$\{T \circ \phi(Y)\}^{CE} \approx \frac{1}{N} \sum_{i=1}^N T \circ \phi(Y^{(i)}),$$

where $(Y^{(i)})_{0 < i \leq N}$ are conditional simulations of the Gaussian random field. To compute these conditional simulations, the orthogonal decomposition of the field proposed above is used to transform nonconditional simulations $\tilde{Y}^{(i)}$ of the Gaussian

field into conditional ones, by adding to the nonconditional simulation the simple kriging of the discrepancy at the conditioning points (Chilès and Delfiner 2012, p. 495)

$$Y^{(i)}(x) \equiv \tilde{Y}^{(i)}(x) + \sum_{\alpha} \lambda_{\alpha}^{SK} [Y(x_{\alpha}) - \tilde{Y}^{(i)}(x_{\alpha})].$$

Conclusions

Kriging is an optimal linear estimator that aims at estimating the value of a regionalized variable, either at a punctual location or over a larger support such as a block or a domain (Cressie 1991; Chilès and Delfiner 2012). The kriged value at a target point is a weighted average of the data values. The kriging weights are computed by solving the kriging system which takes into account the spatial continuity of the phenomenon (modeled by a covariance function or a variogram) and the relative location of the observation and the target. Kriging is unbiased and optimal and provides a measure of the accuracy of its predicted value.

Since Krige's initial regression, which took into account a local average grade and a global one to avoid bias in the estimation of a block, kriging techniques have constantly evolved to answer to new estimation problems encountered in the numerous operational situations of geosciences. Kriging is also used as a basic ingredient for more advanced techniques to handle non-Gaussian models constrained by nonlinear observations (e.g., multi-Gaussian kriging and Monte-Carlo simulations).

Due to its simplicity, kriging is robust and flexible, and it can be extended to handle new applications. Among the recent examples, SPDE techniques are now available for global application of kriging and conditional geostatistical simulations to large data sets and nonstationary random function models; the multivariate kriging or co-kriging (Wackernagel 2013) has been extended to functional data (Menafoglio et al. 2013) in order to analyze infinite-dimensional data, such as curves, surfaces, and images; the integration of complex processes governing the spatial variability of the variable of interest can be achieved using physically informed kriging (Yang et al. 2019).

Bibliography

- Cameletti M, Lindgren F, Simpson D, Rue H (2013) Spatiotemporal modeling of particulate matter concentration through the SPDE approach. *ASTA Adv Stat Anal* 97(2):109–131
- Chilès J-P, Delfiner P (2012) *Geostatistics: modeling spatial uncertainty*, 2nd edn. Wiley, New York/Toronto
- Cressie N (1990) The origins of kriging. *Math Geol* 22(3):239–252
- Cressie N (1991) Geostatistical analysis of spatial data. In: *Spatial statistics and digital image analysis*, pp 87–108

- Cressie N, Johannesson G (2008) Fixed rank kriging for very large spatial data sets. *J R Stat Soc: Ser B (Stat Methodol)* 70(1): 209–226
- Desassis N, Renard D (2013) Automatic variogram modeling by iterative least squares: univariate and multivariate cases. *Math Geosci* 45(4): 453–470
- Emery X (2005) Simple and ordinary multigaussian kriging for estimating recoverable reserves. *Math Geol* 37(3):295–319
- Fuglstad GA, Lindgren F, Simpson D, Rue H (2015) Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy. *Stat Sin* 25:115–133
- Furrer R, Genton MG, Nychka D (2006) Covariance tapering for interpolation of large spatial datasets. *J Comput Graph Stat* 15(3): 502–523
- Journel AG (1977) Kriging in terms of projections. *J Int Assoc Math Geol* 9(6):563–586
- Krige D (1951) A statistical approach to some basic mine valuation problems on the Witwatersrand. *J South Afr Inst Min Metall* 52(6): 119–139
- Kyriakidis P (2004) A geostatistical framework for area-to-point spatial interpolation. *Geogr Anal* 36(3):259–289
- Lindgren F, Rue H, Lindström J (2011) An explicit link between Gaussian fields 670 and Gaussian Markov random fields: the SPDE approach (with discussion). *J R Stat Soc, Ser B* 73:423–498
- Matheron G (1962) *Traité de géostatistique appliquée* tome 1, *memoires*, 14. Bureau de Recherches Géologiques et Minières
- Matheron G (1973) The intrinsic random functions and their applications. *Adv Appl Probab* 5(3):439–468
- Matheron G (1982) Pour une analyse krigéante des données régionalisées. Tech Rep N-732, Centre de Géostatistique, Fontainebleau
- Matheron G (2019) Matheron's theory of regionalized variables. International Association for Mathematical Geology Studies in mathematical geology. Oxford University Press, Oxford, UK
- Menafoglio A, Secchi P, Rosa MD et al (2013) A universal kriging predictor for spatially dependent functional data of a Hilbert space. *Electron J Stat* 7:2209–2240
- Renard D, Yancey SD (1984) Smoothing discontinuities when extrapolating using moving neighbourhood. In: *Geostatistics for natural resources characterization*. NATO advanced Study Institute, pp 679–690
- Rivoirard J, Romary T (2011) Continuity for kriging with moving neighborhood. *Math Geosci* 43(4):469–481
- Rue H, Held L (2005) *Gaussian Markov random fields: theory and applications*. CRC Press, Boca Raton, FL
- Switzer P, Green AA (1984) Min/max autocorrelation factors for multivariate spatial imagery. Tech Rep 6, Department of Statistics, Stanford University
- Verly G (1983) The multigaussian approach and its applications to the estimation of local reserves. *J Int Assoc Math Geol* 15(2): 259–286
- Wackernagel H (2013) *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media, Berlin/Heidelberg, Germany
- Whittle P (1954) On stationary processes in the plane. *Biometrika* 41(3/4):434–449
- Wiener N (1950) *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. MIT Press, Cambridge, MA
- Yang X, Barajas-Solano D, Tartakovsky G, Tartakovsky AM (2019) Physics-informed cokriging: a Gaussian-process-regression-based multifidelity method for data-model convergence. *J Comput Phys* 395:410–431. <https://doi.org/10.1016/j.jcp.2019.06.041>. ISSN 0021-9991

Krumbein, William Christian

E. H. Timothy Whitten

Riverside, Widescombe-in-the-Moor, Devon, UK



Fig. 1 The portrait of Krumbein was made in 1954 by the Stuart-Rodgers Studio, Evanston, Illinois, USA and I published it in 1975 in © Geological Society of America Memoir 142

Biography

William Christian KRUMBEIN, born on 28 January 1902 in Beaver Falls, Pennsylvania, USA, died 18 August 1979 in Los Angeles. He is widely recognized as “the father of mathematical geology.”

Gaining a University of Chicago PhD degree (business administration, with one introductory geology course, 1926), Krumbein worked on insurance adjustments for a finance company. Returning to University of Chicago (1929), J Harland Bretz supervised his PhD thesis (1932) on the mechanical analysis of glacial till samples. At that time, geological science was largely descriptive, but Krumbein recognized the potential of applying business statistical methods to sediments. He advanced from instructor to associate professor at the University of Chicago (1933–1942). During World War II, he served (1942–1945) with the Beach Erosion Board, US Army Corps of Engineers (military beach-landing intelligence/planning); the Board

funded much of Krumbein's later research. Krumbein became Professor of Geology at Northwestern University (1946), being named William Deering Professor of Geological Sciences (1960), and Emeritus Professor in 1970. The University of Syracuse awarded an honorary DSc (1979).

Krumbein's early work on sediment analysis and sedimentation included mechanical analyses, sampling problems, and data representation; he introduced the important *phi* scale (logarithmic transformation) for sediment particle-size distributions. Such work was tedious and difficult with hand and electrical calculators; an early IBM360 mainframe computer, installed in Northwestern University's Astronomy Department, soon had Krumbein processing data with IBM punch cards. He (with L. L. Sloss) introduced the computer into geology with their 1958 AAPG paper "High-speed digital computers in stratigraphic and facies analysis."

Krumbein was a masterful, stimulating teacher in the small Northwestern Geology Department, where close colleagues included Dapples and Sloss. Constitutionally rejecting "conventional wisdom", Krumbein pursued innovative methods whereby geological phenomena could be expressed with mathematical rigor. He taught students to question established theories and paradigms, encouraging them to view statistics as fun. He demonstrated the many ways to gather, process, and interpret the meaning of information. Krumbein paid particular attention to appropriate sampling and to each variable's variance; students were encouraged to recognize effects of normalization, transformations, etc., on calculations by writing their own computer programs.

By the 1970s, probabilistic sampling, properties, and classification of modern and ancient sediments, process-response

models, stochastic simulation, regression analysis, Markov chains, deterministic modeling, and spatial "geostatistics" were important in his research.

Manual of Sedimentary Petrography (1938, with Pettijohn) and *Stratigraphy and Sedimentation* (1951, 1963, with Sloss) were standard textbooks for several decades; *Lithofacies Maps, an atlas . . .* (1960, with Dapples and Sloss) was a compilation of their students' research. His book *An Introduction to Statistical Models in Geology* (1965, with Graybill) confirmed Krumbein at the forefront of quantitative geology. He co-authored several papers with leading statisticians including John Tukey, Geoffrey Watson, and Graybill. Krumbein published over 140 papers dealing with sediment analysis (39 publications), stratigraphic analysis (27), and statistical methods in geology (62).

The founding meeting of IAMG (1968, Prague) named Krumbein its Past President. The Krumbein Medal, IAMG's premier award initiated in 1976, is awarded annually for career achievement of a senior scientist.

Bibliography

- Krumbein WC, Dapples EC, Sloss LL (1960) *Lithofacies maps, an atlas of the United States and Southern Canada*. Wiley, New York
- Krumbein WC, Graybill FA (1965) *An introduction to statistical models in geology*. McGraw-Hill, New York
- Krumbein WC, Pettijohn FJ (1938) *Manual of sedimentary petrography*. Appleton-Century, New York
- Krumbein WC, Sloss LL (1963) *Stratigraphy and sedimentation*, 2nd edn. Freeman, San Francisco

Laplace Transform

Jaya Sreevalsan-Nair

Graphics-Visualization-Computing Lab, International
Institute of Information Technology Bangalore, Electronic
City, Bangalore, Karnataka, India

Definition

Several physical phenomena, such as heat conduction and wave propagation, are modeled as ordinary and partial differential equations (ODEs and PDEs), along with initial and boundary conditions. These equations are usually differentiated with respect to time and have several approaches for finding its solutions. Laplace transform is one such tool that is available for solving these equations (Spiegel 1965; Schiff 1999). If $F(t)$ is a function of time t , such that $t > 0$, then the Laplace transform of a real values function $F(t)$ is given by $\mathcal{L}(F(t))$, which is given by:

$$\mathcal{L}(F(t)) = f(s) = \lim_{\tau \rightarrow \infty} \int_{t=0}^{\tau} e^{-st} \cdot F(t) \cdot dt, \text{ for parameter } s \in \mathbb{R}.$$

If $F(t)$ is a complex-valued function, then s is complex. The Laplace transform $\mathcal{L}(F(t))$ exists if and only if the limit of the integral exists. If the limit exists, then the integral is said to converge, and in the absence of the limit, the integral is said to diverge. The integral is the ordinary Riemann improper integral (Schiff 1999). The Laplace inverse transform on $f(s)$ gives back the original function. As an example, for $F(t) = t$ for $t > 0$, its Laplace transform is $f(s) = \frac{1}{s^2}$, where $s > 0$, and thus, the Laplace inverse of $f(s) = \frac{1}{s^2}$ is $F(t) = t$, as shown in Fig. 1. If $s \leq 0$, then the integral diverges. Thus, Laplace transformation, denoted by the symbol $\mathcal{L}$, is a map applied to a function $F(t)$ to generate a new function $f(s)$, and the same is applicable to its inverse, $\mathcal{L}^{-1}$.

Theorem 1 If $F(t)$ is piecewise continuous on $[0, \infty)$ and of exponential form with order α , then the transform $\mathcal{L}(F(t))$ exists for $\text{Re}(s) > \alpha$ and the integral converges.

For example, $\mathcal{L}(e^{\alpha \cdot t}) = \frac{1}{(s-\alpha)}$, where $\text{Re}(s) > \alpha$. Thus, this theorem defines the class of functions, L , with existing Laplace transform (Schiff 1999).

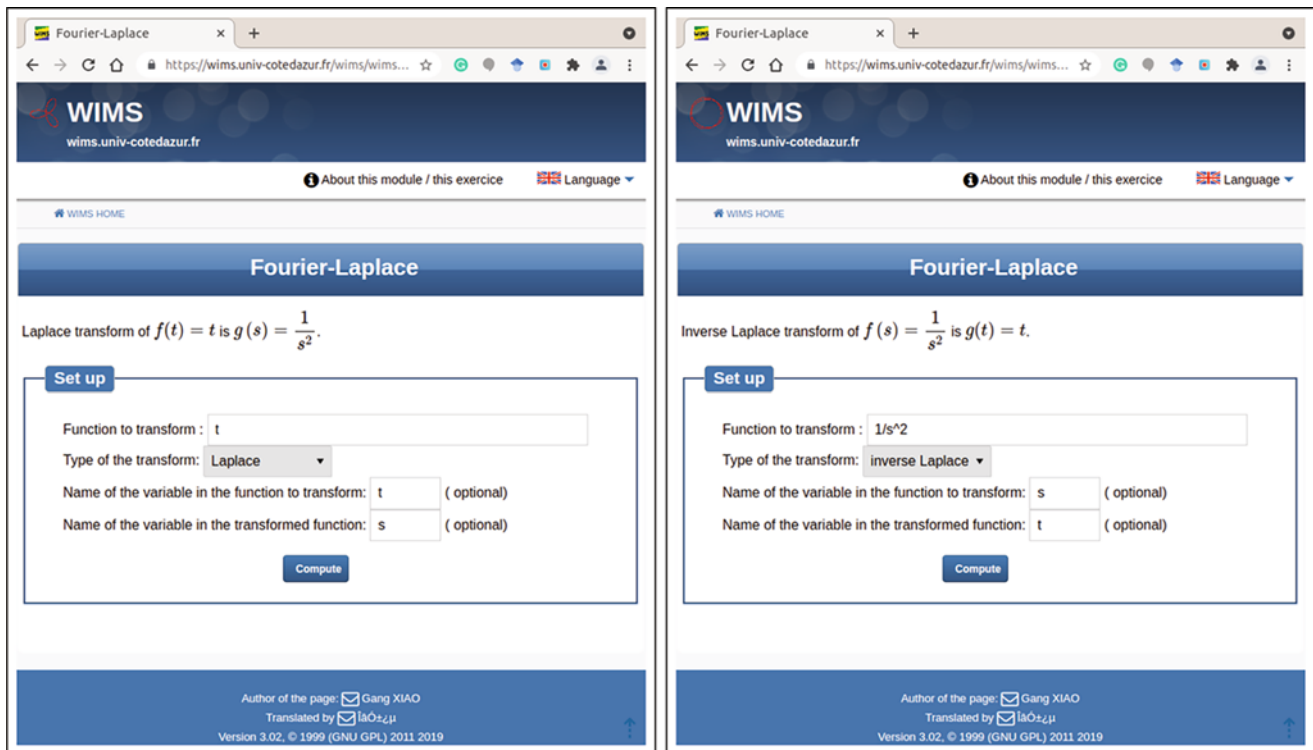
Some of the useful properties of Laplace transform include:

- **Linearity:** If $F_1 \in L$ for $\text{Re}(s) > \alpha$, and $F_2 \in L$ for $\text{Re}(s) > \beta$, then $F_1 + F_2 \in L$ for $\text{Re}(s) > \max(\alpha, \beta)$, and $\mathcal{L}(c_1 F_1(t) + c_2 F_2(t)) = c_1 \mathcal{L}(F_1(t)) + c_2 \mathcal{L}(F_2(t))$, for constants c_1 and c_2 .
- **Infinite Series:** The Laplace transform of an infinite series in t can be done term-by-term computation if the series converges for $t \geq 0$, exponential order $\alpha > 0$, and a constant term $K > 0$, where the coefficients of the powers of t in the power series have absolute values, given by $|a_n| \leq \frac{K \alpha^n}{n!}$.
- **Uniform Convergence:** For functions in L , the integral in Laplace transform converges uniformly.
- **Translation Property:** Also, known as shifting property, the first and second translations imply that shifting in the time domain (t) or Laplace domain (s), provides *closed form* solutions. For $\mathcal{L}(F(t)) = f(s)$:

$$\text{First translation : } \mathcal{L}(e^{\alpha \cdot t} F(t)) = f(s - \alpha).$$

$$\begin{aligned} \text{Second translation : } \mathcal{L}(G(t)) \\ &= e^{-\alpha \cdot s} f(s), \text{ where } G(t) \\ &= \begin{cases} F(t - \alpha), & \text{if } t > \alpha \\ 0, & \text{if } t < \alpha \end{cases} \end{aligned}$$

- **Differentiation and Integration:** There exist analytical solutions of derivatives of functions using its Laplace



Laplace Transform, Fig. 1 The Laplace transform of $F(t) = t$ and its inverse, as shown on an online tool for computing Fourier-Laplace transforms on the WIMS (WWW Interactive Multipurpose Server),

transforms and, similarly, of integrals using its inverse Laplace transforms.

- **Change of Scale Property:** The analytical solutions for scalar multiplication of the functions give the following Laplace transform, i.e., if $\mathcal{L}(F(t)) = f(s)$, then:

$$\mathcal{L}(F(\alpha \cdot t)) = \frac{1}{\alpha} f\left(\frac{s}{\alpha}\right).$$

For a more exhaustive list of properties, the textbooks on Laplace theorem (Spiegel 1965; Schiff 1999) are highly recommended references. The direct or inverse Laplace transforms of some of the special functions have closed form solutions. Such special functions include Gamma function (Γ), Bessel function of order n (J_n), Heaviside's unit step function (U), Dirac delta function, or unit impulse function (δ).

Laplace transform can be seen similar to Fourier transform, in the context of transforming signals in the time domain into a different domain. It can also have overlap with Z-transform, where the one-sided Z-transform and the Laplace transform of an ideally sampled signal are equivalent, when substituting $z \stackrel{\text{def}}{=} e^{sT}$, where s is the Laplace variable, $T = \frac{1}{f_s}$, where f_s is the sampling frequency. Additionally, the Laplace transform can be extended and unified with

authored by Gang Xiao (1999). (Image source: <https://wims.univ-cotedazur.fr/wims>)

Z-transform, which has potential applications with higher-order linear dynamic equations (Bohner and Peterson 2002).

Overview

The history of the modern Laplace transform is known to be not too old, as it started in 1937 with Bateman using it in 1910 and then was popularized by Bernstein and Doetsch in the 1920s and 1930s (Deakin 1992). The modern method was formalized in 1937, but prior to that, it was known as *Heaviside operational calculus*. The older *Laplace transformation* is closely related to Laplace transform and was then considered a solution to the PDE, $t.X''(t) - t.X'(t) + X(t) = 0$. Heaviside calculus was popular among electrical engineers, and Heaviside was a contemporary of Koppelman and Boole. The Laplace transformation itself was introduced by Euler and then used and studied by Laplace, Abel, and Poincaré. In 1937 and later, the modern Laplace transform slowly replaced the Laplace transformation, when the solutions to the aforementioned PDE began to be used in the integral form, $X(t) = \int_C e^{st} x_1(s) ds$ for $x_1(s) = a.s^{-2}$. Unlike

the earlier solutions, which used fixed limits and precomputed tables, the new solution allowed for the application/problem to define the contour C in the integral equation.

In practice, there are several numerical methods used for Laplace inversion (Davies and Martin 1979). There are broadly five classes of methods, namely, one which (i) computes a sample, (ii) expands $F(t)$ in exponential functions, (iii) uses Gaussian quadrature, (iv) uses bilinear transformation, and (v) uses Fourier series. In (i), there are methods such as Widder and Gaver-Stehfast. In (ii), methods using Legendre polynomials, trigonometric functions, and methods by Bellman et al. and Schapery are included. In (iii), Piessens' Gaussian quadrature and Schmittroth's methods are considered. In (iv), methods by Laguerre-Weeks, Laguerre-Piessens-Branders, and Chebyshev are included. In (v), methods by Dubner-Abate, Silverberg-Durbin, and Crump are included. In current practice, these methods are included in several mathematical packages, e.g., Mathematica.

Applications

There are several applications where the Laplace operator is used in geoscientific modeling and analysis. As an example, in differential geometry, generalizations of Laplace operator are widely used, such as Laplace-Beltrami operator and Laplace-de Rham operator. Each generalization has applications of its own, e.g., Laplace-Beltrami operator is widely used for shape analysis in computational geometry, for any twice-differentiable real-valued function in Euclidean space. The Laplace-Beltrami operator can be applied as the divergence of covariant tensor derivative and hence can be used on second-order tensor fields, such as stress, strain, etc., commonly used in geological or geophysical applications (Ken-Ichi 1984). Yet another example is in the use of waveform inversion using Laplace-transformed wavefields to delineate structures of the Earth, in three-dimensional (3D) seismic inversion (Shin and Cha 2008).

In practice, a class of applications where the Laplace operator is widely used is in hydrological flows (Chen et al. 2003). For convergent flows, say for that in sand and gravel aquifer, Laplace transformed power series is used for solving the radially scale-dependent advection-dispersion equation (Chen et al. 2003). Advection-dispersion equation is a special case of advection-dispersion-reaction (ADR) equation. ADR equation is used as a transport model, e.g., contaminant transport in groundwater, where the advection part of the PDE models the bulk movement of the solute (contaminant) in the flow, the dispersion part models the spreading of the solute, and the reaction part models changes in the solute mass owing to biotic and abiotic processes. In groundwater flow modeling, inverse Laplace transform has been routinely used for studying the late-time behavior of the hydraulic head in the flow (Mathias and Zimmerman 2003). In shallow water equations, Laplace inversion has been used in a contour integral to simulate low-frequency components and at the same time,

denoise, i.e., eliminate high-frequency components in advection flow in atmospheric waves, such as Kelvin waves (Clancy and Lynch 2011).

Conclusions

In summary, several physical phenomena require Laplace transforms and their inversions for their modeling, simulation, and analysis. Hence, it continues to be a powerful tool used in geoscientific applications, even in conjunction with modern deep learning solutions.

Cross-References

► [Fast Fourier Transform](#)

References

- Bohner M, Peterson A (2002) Laplace transform and Z-transform: unification and extension. *Meth Appl Anal* 9(1):151–158
- Chen JS, Liu CW, Hsu HT, Liao CM (2003) A Laplace transform power series solution for solute transport in a convergent flow field with scale-dependent dispersion. *Water Resour Res* 39(8)
- Clancy C, Lynch P (2011) Laplace transform integration of the shallow-water equations. Part I: Eulerian formulation and Kelvin waves. *Q J R Meteorol Soc* 137(656):792–799
- Davies B, Martin B (1979) Numerical inversion of the Laplace transform: a survey and comparison of methods. *J Comput Phys* 33(1):1–32
- Deakin MA (1992) The ascendancy of the Laplace transform and how it came about. *Arch Hist Exact Sci* 44(3):265–286
- Ken-Ichi K (1984) Distribution of directional data and fabric tensors. *Int J Eng Sci* 22(2):149–164
- Mathias SA, Zimmerman RW (2003) Laplace transform inversion for late-time behavior of groundwater flow problems. *Water Resour Res* 39(10)
- Schiff JL (1999) *The Laplace transform: theory and applications*. Springer Science & Business Media, New York
- Shin C, Cha YH (2008) Waveform inversion in the Laplace domain. *Geophys J Int* 173(3):922–931
- Spiegel MR (1965) *Laplace transforms*. McGraw-Hill, New York

Least Absolute Deviation

A. Erhan Tercan

Department of Mining Engineering, Hacettepe University, Ankara, Turkey

Synonyms

L1-norm; Least Absolute Error; Least Absolute Value; Minimum Sum of Absolute Errors

Definition

Least absolute deviation (LAD) is an optimization criterion based on minimization of sum of absolute deviations between observed and predicted values.

The following sections discuss the problem within the scope of estimation. Let x_{ik} be the i th observation on the k th independent variable for $i = 1, \dots, n$ and $k = 1, \dots, p$ and y_i be the i th value of the dependent variable. We want to find a function f such that $f(x_{i1}, x_{i2}, \dots, x_{ip}) = y_i$. To attain this objective, suppose that the function f is of the linear form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i \quad (1)$$

where $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ are parameters whose values are not known but which we would like to estimate and ε_i is the random disturbance. We now seek estimated values of the unknown parameters, denoted by $b_0, b_1, b_2, \dots, b_p$ that minimize the sum of the absolute values of the residuals

$$\sum_{i=1}^n |e_i| = \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2)$$

where $\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + b_p x_{ip}$ is the i th predicted value, n is the number of observations, and p is the number of the independent variables.

This type of minimization problem can arise in many fields of science and engineering, including regression analysis, function approximation, signal processing, image restoration, parameter estimation, filter design, robot control, and speech enhancement.

LAD estimation is commonly used in regression analysis as an alternative measure to least squares deviation (LSD) that uses minimum sum of squared errors. It is more attractive than LSD when the errors follow non-Gaussian distributions with longer tails.

Historical Material

Roger Joseph Boscovich introduced LAD criterion in 1757 to attempt to reconcile inconsistent measurements for the purpose of estimating the shape of the Earth. Surprisingly, it was nearly a half-century before Legendre published his “Principle of Least Squares” in 1805 (Farebrother 1999). Computational difficulties associated with LAD effectively prevented its use until the second half of the twentieth century. Charnes et al. (1955) indicated that solving least absolute deviation problem is essentially equivalent to solving a linear programming problem via the simplex method. This was probably the birth of automated LAD fitting

(Bloomfield and Steiger, 1983). Starting in 1987 a number of congresses were dedicated to the problems related to least absolute deviation, and the last one was held in 2002 (Dodge 2002).

LAD did not attract Earth scientists very much due to computational and theoretical difficulties. Dougherty and Smith (1966) proposed the use of a simple trend surface method in which the sum of absolute values of deviations is minimized. While looking for resistant and robust methods of kriging, Henley (1981) devoted his book to robust methods of spatial estimation and suggested a kriging estimator based on the minimization of absolute error. As a linear estimator, kriging minimizes squared difference (error) between true and estimated value and therefore is not resistant to outliers. Dowd (1984) discussed several resistant kriging methods, reviewed robust and resistant variogram estimations, and introduced alternative methods. It is well known that the traditional estimator of the variogram is neither robust nor resistant. Walker and Loftis (1997) developed a median-type estimator based on distance-weighted LAD regression, minimizing the sum of absolute errors to optimally fit a polynomial regression to the observations.

Essential Concepts and Applications in More Detail

When $p = 0$, Eq. 2 is reduced to minimization of $\sum_{i=1}^n |y_i - b_0|$ so that the optimal value of b_0 is the median of the values y_i , $i = 1, \dots, n$. For the same estimation problem but with LSD, the solution for b_0 is the arithmetic mean of the values of the dependent variable. It is easy to see that for an even number m of the values, the optimum value of b_0 for LAD lies anywhere between the two central values of y_i . Thus, the minimization of Eq. 2 with $p = 0$ often leads to an infinity of solutions. In contrast, LSD produces unique solutions.

When $p \geq 1$, there are no simple formulas for minimization of Eq. 2 because it is non-differentiable. On the other hand, it is well known that it can be transformed into solving the linear programming problem as follows:

$$\text{Minimize } \sum_{i=1}^n (a_i + z_i) \quad (3)$$

subject to

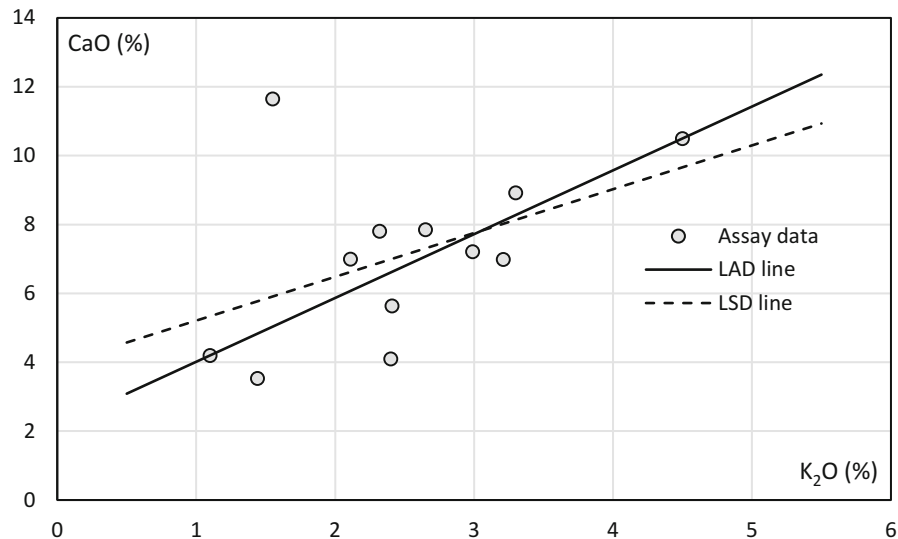
$$b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + b_p x_{ip} + z_i - a_i = y_i$$

$$a_i \geq 0, z_i \geq 0, i = 1, \dots, n$$

Least Absolute Deviation, Table 1 The assay data for 12 samples collected from the iron deposit

K ₂ O (%)	4.50	2.11	1.44	3.30	2.99	2.32	1.55	3.21	1.10	2.40	2.65	2.41
CaO (%)	10.50	7.00	3.53	8.92	7.21	7.81	11.65	6.99	4.20	4.10	7.85	5.64

Least Absolute Deviation, Fig. 1 Scatter plot of K₂O versus CaO for assay data from the iron deposit. The fitted LAD line (solid) is $\text{CaO} = 1.85 \times \text{K}_2\text{O} + 2.16$, while it is $\text{CaO} = 1.27 \times \text{K}_2\text{O} + 3.94$ for LSD (dashed)



where a_i and z_i are, respectively, the positive and negative deviation associated with the i th observation (Armstrong and Kung 1978). This problem can be solved by the simplex method.

To illustrate the method, consider the data in Table 1 representing the variation of K₂O and CaO for 12 samples from a volcanic syn-sedimentary iron deposit.

Assume that K₂O is the independent variable and CaO is the dependent variable. Figure 1 shows a scatter plot of K₂O versus CaO together with the LAD and the LSD lines. In the plot, the solid line is the LAD fit to the data, while the dashed line is for the LSD fit. The intercept and the slope for LAD are $b_0=2.16$ and $b_1=1.85$, respectively. The LAD line was computed using the algorithm given in Armstrong and Kung (1978).

Note that the data contains an outlier at the point (1.55% K₂O, 11.65% CaO). As long as a sample remains on the same side of the LAD line, the estimates, b_0 and b_1 , do not change. This is obviously not the case for LSD. LAD criterion is thus less sensitive to outliers than that LSD; the LAD estimator is called a robust estimator.

Note also that the LAD line passes through two of the data points. This is always the case produced by the LAD algorithm for a nondegenerate data set with $p = 1$. If the data set is subject to degeneracy, then the LAD line passes through more than two data points. Birkes and Dodge (1993) suggest simple solutions for such cases.

Conclusions

Least absolute deviation is a robust criterion when the deviations follow distributions that are non-normal and subject to outliers. Since the objective function based on the LAD criterion is non-differentiable, there are no formulas for its solution; instead, the iterative algorithms such as the simplex method are used. Perhaps because of these computational and theoretical difficulties, robust methods developed on the LAD criterion have not gained wide acceptance among geoscience practitioners.

Cross-References

- Iterative Weighted Least Squares
- Least Absolute Value
- Least Squares
- Ordinary Least Squares
- Regression

Bibliography

Armstrong RD, Kung MT (1978) Algorithm AS 132: least absolute estimates for a simple linear regression problem. J R Stat Soc Series C 27(3):363–366

Birkes D, Dodge Y (1993) Alternative methods of regression. Wiley, New York

Bloomfield P, Steiger WL (1983) Least absolute deviations: theory, applications and algorithms. Birkhäuser, Boston

- Charnes A, Cooper WW, Ferguson RO (1955) Optimal estimation of executive compensation by linear programming. *Manag Sci* 1: 138–151
- Dodge Y (ed) (2002) Statistical data analysis based on the L_1 -norm and related methods. Birkhäuser, Basel
- Dougherty EL, Smith ST (1966) The use of linear programming to filter digitized map data. *Geophysics* 31:253–259
- Dowd PA (1984) The variogram and kriging: robust and resistant estimators. In: Verly G, David M, Journel AG, Marechal A (eds) *Geostatistics for natural resources characterization*. Springer, Dordrecht, pp 91–106
- Farebrother RW (1999) Fitting linear relationships: a history of the calculus of observations 1750–1900. Springer, New York
- Henley S (1981) *Nonparametric Geostatistics*. Elsevier, London
- Walker DD, Loftis JC (1997) Alternative spatial estimators for ground-water and soil measurements. *Ground Water* 35(4):593–601

Least Absolute Value

U. Radojčić¹ and Klaus Nordhausen²

¹Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria

²Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland

Definition

The least absolute value (LAV) method is a statistical optimization method which has, as a minimization criterion, the absolute errors between the data and the statistic of interest. In the one sample location problem, the solution is the sample median, while in a regression context it is an alternative to the least squares regression and is also known as least absolute deviations (LAD) regression, least absolute errors (LAE) regression, or L_1 -norm regression.

Introduction

Consider a random sample $\mathbf{y} = (y_1, \dots, y_n)^\top$ of size n . When one is interested in estimating the statistic $t(\mathbf{y})$, then the least absolute value (LAV) method solves the given problem by minimizing the criterion:

$$\sum_{i=1}^n |y_i - t(\mathbf{y})| \quad \text{or} \quad \sum_{i=1}^n |e_i|,$$

with $e_i = y_i - t(\mathbf{y})$. Thus, it minimizes the sum of absolute deviations between the observed samples and the statistic. For example, in the one sample location case the solution to the given problem is the median of the given sample.

LAV is however mostly considered in a regression framework. Let $\mathbf{y}$ denote the vector of responses and $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$ be the matrix containing the p predictors for each response for the regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (1)$$

where $\boldsymbol{\beta}$ is the p -variate unknown vector of coefficients one needs to estimate and $\boldsymbol{\varepsilon}$ is an n -vector of random disturbances. For the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, the residuals are defined as:

$$\mathbf{e} = (e_1, \dots, e_n)^\top = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}.$$

The least absolute value (LAV) estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ minimizes thus

$$\sum_{i=1}^n |e_i|. \quad (2)$$

An early formulation of LAV dates back to R. J. Boscovich in 1757 but the form mentioned above is due to F. Y. Edgeworth who considered it in 1887 for the simple linear regression model.

LAV regression was however often overshadowed by ordinary least squares regression (OLS) which is formulated as the minimization problem:

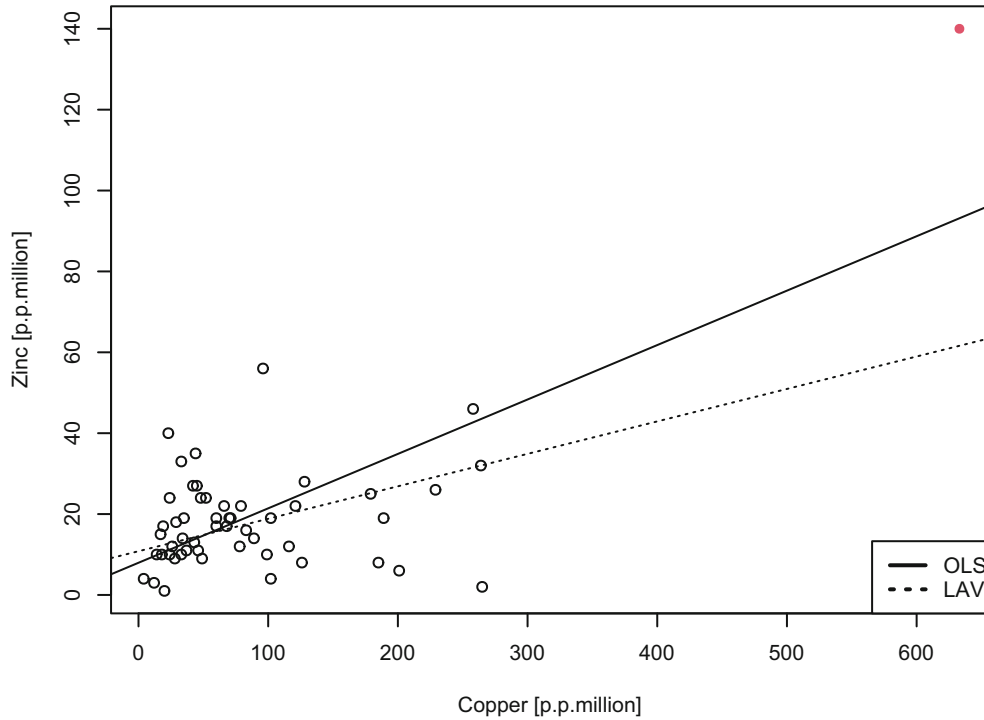
$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^n e_i^2 \quad (3)$$

and thus minimizes the squared deviations which can be seen as L_2 -norm regression.

Motivation

There is a large variety of reasons why OLS regression is probably the most common regression method, one of which is widely available efficient and easy-to-use computer routines which have made the application of OLS rather effortless and inexpensive. Also, for independent and identically distributed (iid) disturbances with finite variances, the OLS has many optimality properties, while for iid Gaussian disturbances it corresponds to the maximum likelihood estimator of model (1). However, it is well known that the OLS suffers if the disturbances have heavy tails and already a single outlier can ruin the estimates completely.

The lack of robustness is due to the optimization criterion which squares the deviations and therefore tries to avoid large residual values. This might even cause outliers in the data to



Least Absolute Value, Fig. 1 Mineral dataset with OLS and LAV model regression lines when modeling zinc content versus copper content. The suspicious observation is marked in red

be masked making outlier identification, based on residuals, difficult.

Figure 1 visualizes this problem. The data shown there comes from 53 rock samples collected in Western Australia and gives the copper (Cu) and zinc (Zn) contents (in p.p. million) and can be found in the R package **RobStatTM** (Yohai et al. 2020) as the dataset **minerals**. The figure shows the OLS fit and the LAV fit and reveals that the OLS fit is much more attracted by the observation marked in red, which can be considered an outlier, than the LAV fit. Figure 2 shows the corresponding residuals for both fits and the outlier is much less extreme in the OLS residuals than in the LAV residuals.

To address the lack of robustness and efficiency of OLS in such situations, many robust regression methods have been suggested. See, for example, Maronna et al. (2018) for a general overview and Wilson (1978) for a detailed comparison of OLS and LAV. The general idea is that LAV might be preferable over OLS whenever the median would be also more appropriate than the mean for the data at hand.

Computation of LAV Estimate

One of the things that make LAV regression appealing when compared to other robust alternatives to OLS is that the LAV criterion is convex. However, although the idea of LAV

regression is as intuitive as the one of OLS, unlike the OLS cost function, the LAV cost function is not smooth, and hence the computation of LAV is not that straightforward and the explicit solution to it does not exist. Hence, an iterative approach, which is guaranteed by the convexity of the LAV cost function to converge to the (possibly not unique) global minimum, is needed to find an optimal LAV solution.

The most common approach to obtain LAV estimates is based on the fact that the optimization problem can be reformulated as a linear program. Denote for that purpose $\hat{y}_i = x_i^\top \hat{\beta}$ as the fitted values, then the optimization problem with cost function as in Eq. (2) can be stated as the linear program

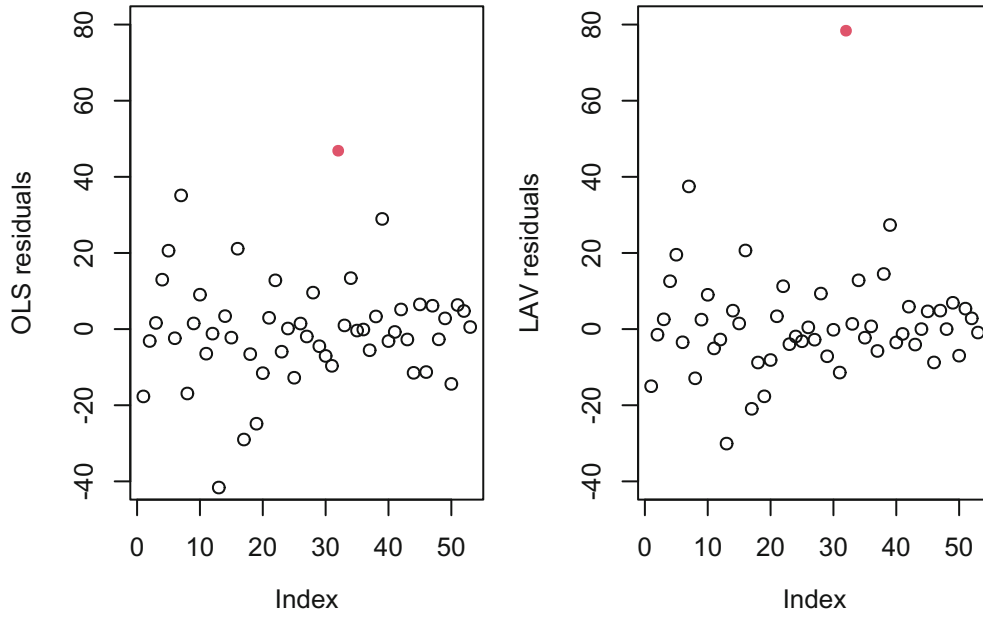
$$\min_{e_i, \hat{\beta}} \sum_{i=1}^n e_i,$$

under the constraints

$$e_i \geq y_i - \hat{y}_i, \quad e_i \geq -(y_i - \hat{y}_i), \quad \text{for } i = 1, \dots, n.$$

Thus, the constraints directly yield $e_i \geq |y_i - \hat{y}_i|$, implying that the function is being minimized for $e_i = |y_i - \hat{y}_i|$, $i = 1, \dots, n$, and for the suitable choice of $\hat{\beta}$.

Hence, the LAV estimate can be obtained with one of the many techniques to solve a linear program, like, for example,



Least Absolute Value, Fig. 2 Residuals of OLS (left) and LAV (right) models of the mineral dataset when modeling zinc content versus copper content. The residual of the suspicious observation is again marked red

the well-known simplex method (Barrodale and Roberts 1974). Other approaches to obtain the LAV estimates can be based on gradient descent (Money et al. 1982) or iterated least squares approach (Armstrong et al. 1979). A detailed list of various optimization methods can be found in Dielman (2005). A property already quite early recognized for LAV regression is that there is a solution which passes through p observations. However, going through all possible hyperplanes defined by p observations to see which is the minimizer of Eq. (2) is only feasible for small sample sizes. Another property of the LAV estimator worth mentioning is that it can, as the OLS, be seen as a maximum likelihood estimator. This is the case when the disturbances in Eq. (1) follow a Laplace distribution.

To conclude, while no closed form solution for LAV regression exists, the problem is well formulated with many algorithms available and implemented, for example, in R in the packages **quantreg** (Koenker 2021) and **L1pack** (Osorio and Wolodzko 2020). Detailed comparison of the performance of the various algorithms can be found in Dielman (2005) and the references therein.

Properties of LAV Estimator

In this section, we give some of the asymptotic properties of LAV estimator $\hat{\beta}$ of β under certain “natural” assumptions. Let the vector of observations y follow model (1). Assume furthermore that the disturbances are independent and identically distributed (iid) with distribution function F and continuous

and positive density f at the median. Furthermore, assume that $\frac{1}{n} \mathbf{X}^T \mathbf{X} \rightarrow \mathbf{Q}$ for $n \rightarrow \infty$, where $\mathbf{Q}$ is a positive definite, symmetric matrix. Then

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow \mathbf{Z},$$

where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, w^2 \mathbf{Q}^{-1})$ comes from a multivariate normal distribution and w^2/n is the asymptotic variance of the ordinary sample median of samples with distribution F , that is, $w^2 = (2f(m))^{-2}$, where m is the median of F . Moreover, under these general conditions, it has been proven that $\hat{\beta}$ is consistent for β .

This asymptotic result clearly shows that for disturbance distributions for which the median is a more efficient estimator of the location than the mean, the LAV estimator $\hat{\beta}$ of β is more efficient than the corresponding OLS estimator (Bassett and Koenker 1978).

Based on this key result it is straightforward to develop inferential tools for β . Let $\beta = (\beta_1^T, \beta_2^T)^T$, where $\beta_1 \in \mathbb{R}^{p_1}$ and $\beta_2 \in \mathbb{R}^{p_2}$, $p_1 + p_2 = p$. Consider then the tests for the nested hypothesis of the form

$$H_0 : \beta_2 = \mathbf{0};$$

- Likelihood ratio test

$$LR = 2 \frac{LAV_1 - LAV_0}{w},$$

where LAV_1 and LAV_0 are sum of absolute values of residuals obtained at the restricted and unrestricted models, respectively.

- Wald test

$$\text{WALD} = \hat{\beta}_2^\top \mathbf{D} \hat{\beta}_2 / w^2,$$

where $\hat{\beta}_2$ is the LAV estimate of β_2 and $\mathbf{D}$ is the appropriate diagonal block of $\mathbf{X}^\top \mathbf{X}$.

- Score test

$$U = \mathbf{g}_2^\top \mathbf{D} \mathbf{g}_2,$$

where $\mathbf{g}_2$ is the appropriate part of the normalized gradient of the LAV cost function, evaluated at the restricted estimate, while $\mathbf{D}$ is as discussed in the Wald test.

All three presented test statistics have, asymptotically, a chi-squared χ_{p2}^2 distribution. Unlike the score test, both the likelihood ratio and the Wald tests however require estimation of the residual scale parameter w . Details on the tests and estimation of w can be found in Dielman (2005).

Unlike most other robust alternatives to OLS, the LAV estimator possesses a number of equivariance properties. It has been proven that the LAV estimator is affine equivariant, shift and scale equivariant, and is also being equivariant to design reparameterizations. For details, see Bassett and Koenker (1978). These properties are also shared by the OLS.

LAV possess another interesting property that has been inherited from a univariate median. Namely, for regression in $\mathbb{R}^2$, the estimated LAV regression line does not change if we vary the responses, as long as they stay at the same side of the regression line. Although it is rather obvious in the case of the one-dimensional median, as nicely stated in Dasgupta and Mishra (2004), it gives us intuition to LAVs “median-type” robustness which transfers to certain insensitivity to outliers. It is obvious that this property of LAV is not being shared by OLS. However, it makes also clear that the LAV does not tolerate well outliers in the x values.

LAV Regression with Corrected Errors

Linear regression is often used to model a linear trend in the data that is serially dependent. In such cases, the disturbances in the model are correlated and the key result about the limiting distribution from above cannot be directly applied. For simplicity consider the case $(\mathbf{x}_t, y_t)$, $t = 1, \dots, n$, following model (1), where the disturbances follow an auto-regressive $AR(1)$ process:

$$\varepsilon_t = \rho \varepsilon_{t-1} + z_t,$$

where ρ , $|\rho| < 1$ is the autocorrelation coefficient and the z_t are iid random errors. There are several two-staged procedures used to correct the autocorrelation in the error terms and are commonly used in OLS regression as well. The first stage is to linearly transform the data using the autocorrelation coefficient ρ , after which the regression is done on the transformed data. The goal of the first stage is to obtain transformed data for the linear model that has uncorrelated errors. One of such two-stage procedures is Prais–Winsten (PW) (Prais and Winsten 1954). The PW transformation matrix for model (1) with disturbances from an auto-regressive $AR(1)$ process with parameter ρ can be written as:

$$\mathbf{M} = \begin{pmatrix} \sqrt{1-\rho^2} & 0 \dots & 0 & 0 \\ -\rho & 1 \dots & 0 & 0 \\ \vdots & \vdots \ddots & \vdots & \vdots \\ 0 & 0 \dots & -\rho & 1 \end{pmatrix}.$$

The first stage of PW procedure is to pre-multiply model (1) by $\mathbf{M}$ yielding

$$\mathbf{M}\mathbf{Y} = \mathbf{M}\mathbf{X}\beta + \mathbf{M}\varepsilon, \text{ i.e., } \mathbf{Y}^* = \mathbf{X}^*\beta + \varepsilon^*, \quad (4)$$

where the vector of transformed responses $\mathbf{Y}^*$ is

$$\mathbf{Y}^* = \mathbf{M}\mathbf{Y} = (\sqrt{1-\rho^2}y_1, y_2 - \rho y_1, \dots, y_n - \rho y_{n-1}),$$

and the matrix of transformed predictors (including a transformed intercept) is

$$\mathbf{X}^* = \mathbf{M}\mathbf{X} = \begin{pmatrix} \sqrt{1-\rho^2} & \sqrt{1-\rho^2}x_{1,1} & \dots & \sqrt{1-\rho^2}x_{1,p-1} \\ 1-\rho & x_{2,1} - \rho x_{1,1} & \dots & x_{2,p-1} - \rho x_{1,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1-\rho & x_{n,1} - \rho x_{n-1,1} & \dots & x_{n,p-1} - \rho x_{n-1,p-1} \end{pmatrix}.$$

In the transformed model (4), ε^* is the vector of uncorrelated errors. These transformations are possible due to the equivariance properties of the LVA estimator. Other two-staged methods tackle the problem of correlated errors in a similar manner. Discussion on the performance of various methods can be found in Dielman (2005).

Other Extensions of LAV

In recent datasets the dimension p is increasing dramatically, often having $p \gg n$ and then sparse solutions $\hat{\beta}$, of the regression problem, are required. Wang et al. (2007) suggested LAV Lasso which adds an L_1 penalty on the β

coefficients. It is remarkable that the LAV Lasso can be reformulated as a regular LAV regression by augmenting the predictor matrix and can therefore be solved using the regular algorithms mentioned above.

LAV regression is also sometimes called median regression as when there is only an intercept, i.e., in the one-sample problem, the estimator of the responses is the “median”. The median gives however only one view of the conditional distribution $y_i | x_i$. Quantile regression addresses these issues and can model any quantile of $y_i | x_i$, and therefore median regression is a special case of quantile regression. For a general overview of quantile regression, see, for example, Koenker (2005).

The least absolute values methodology was originally developed in a univariate framework. However, seeing the minimization criterion as a minimization of an L_1 -norm of the model residuals allows a natural generalization to multivariate problems. General overviews over multivariate L_1 -norm methods are, for example, Oja (2010) and Nordhausen and Oja (2011).

Summary

LAV is a general statistical approach, however, most known in the context of regression, where it is well developed and a serious competitor to the ordinary least squares regression. Compared to OLS it is more robust and has better efficiencies if the disturbances have a heavy-tailed distribution and is therefore, for example, recommended for Cauchy distributed disturbances and is optimal for Laplace distributed disturbances.

Cross-References

- Iterative Weighted Least Squares
- Least Absolute Deviation
- Least Mean Squares
- Least Squares
- Ordinary Least Squares
- Regression

Bibliography

- Armstrong RD, Ftome EL, Kung DS (1979) A revised simplex algorithm for the absolute deviation curve fitting problem. *Commun Stat Simul Comput* 8(2):175–190. <https://doi.org/10.1080/03610917908812113>
- Barrodale I, Roberts FDK (1974) Solution of an overdetermined system of equations in the l_1 norm. *Commun ACM* 17(6):319–320. <https://doi.org/10.1145/355616.361024>

- Bassett G, Koenker R (1978) Asymptotic theory of least absolute error regression. *J Am Stat Assoc* 73(363):618–622. <https://doi.org/10.1080/01621459.1978.10480065>
- Dasgupta M, Mishra SK (2004) Least absolute deviation estimation of linear econometric models: a literature review. *SSRN Electron J*. <https://doi.org/10.2139/ssrn.552502>
- Dielman TE (2005) Least absolute value regression: recent contributions. *J Stat Comput Simul* 75(4):263–286. <https://doi.org/10.1080/0094965042000223680>
- Koenker R (2005) Quantile regression. Cambridge University Press, Cambridge, UK. <https://doi.org/10.1017/CBO9780511754098>
- Koenker R (2021) quantreg: quantile regression. <https://CRAN.R-project.org/package=quantreg>. R package version 5.85
- Maronna RA, Martin RD, Yohai VJ, Salibián-Barrera M (2018) Robust statistics: theory and methods (with R), 2nd edn. Wiley, Hoboken
- Money AH, Affleck-Graves JF, Hart ML, Barr GDI (1982) The linear regression model: L_p norm estimation and the choice of p . *Commun Stat Simul Comput* 11(1):89–109. <https://doi.org/10.1080/03610918208812247>
- Nordhausen K, Oja H (2011) Multivariate L_1 statistical methods: the package MNM. *J Stat Softw* 43:1–28. <https://doi.org/10.18637/jss.v043.i05>
- Oja H (2010) Multivariate nonparametric methods with R: an approach based on spatial signs and ranks. Springer, New York
- Osorio F, Wolodko T (2020) L1pack: Routines for L_1 estimation. <http://l1pack.mat.utfsm.cl>. R package version 0.38.196
- Prais SJ, Winsten CB (1954) Trend estimators and serial correlation. Technical report, Cowles Commission discussion paper, Chicago
- Wang H, Li G, Jiang G (2007) Robust regression shrinkage and consistent variable selection through the LAD-lasso. *J Bus Econ Stat* 25: 347–355. <https://doi.org/10.1198/073500106000000251>
- Wilson HG (1978) Least squares versus minimum absolute deviations estimation in linear models. *Decis Sci* 9(2):322–335. <https://doi.org/10.1111/j.1540-5915.1978.tb01388.x>
- Yohai V, Maronna R, Martin D, Brownson G, Konis K, Salibián-Barrera M (2020) RobStatTM: robust statistics: theory and methods. <https://CRAN.R-project.org/package=RobStatTM>. R package version 1.0.2

Least Mean Squares

Mark A. Engle

Department of Earth, Environmental and Resource Sciences,
University of Texas at El Paso, El Paso, Texas, USA

Definition

Least mean squares is a numerical algorithm that iteratively estimates ideal data-fitting parameters or weights by attempting to minimize the mean of the squared distance of the errors between each data point and the corresponding fitted estimate. The minimization of least mean squares has no analytical solution and often relies on stochastic gradient descent, optimizing the first-order derivative of the least mean squares function to incrementally adjust and update model parameters or weights towards a global minimum. Typically, least mean squares makes updates to the weights after a

randomly chosen sample or a subsample, rather than the entire batch, making it a form of adaptive learning. The method is used across a number of Earth science applications including regression and geophysical data filtering and is the basis of early machine learning methods.

Introduction

Consider a dataset composed of a real, independent variable (y) that is a function of n real, dependent variables ($\mathbf{x}$). One could assume that the relationship between y and $\mathbf{x}$ can be reasonably approximated by applying n weights ($\mathbf{w}$) to the dependent variables, such as in linear regression:

$$y = w_1x_1 + w_2x_2 + \dots + w_nx_n,$$

or in vector form:

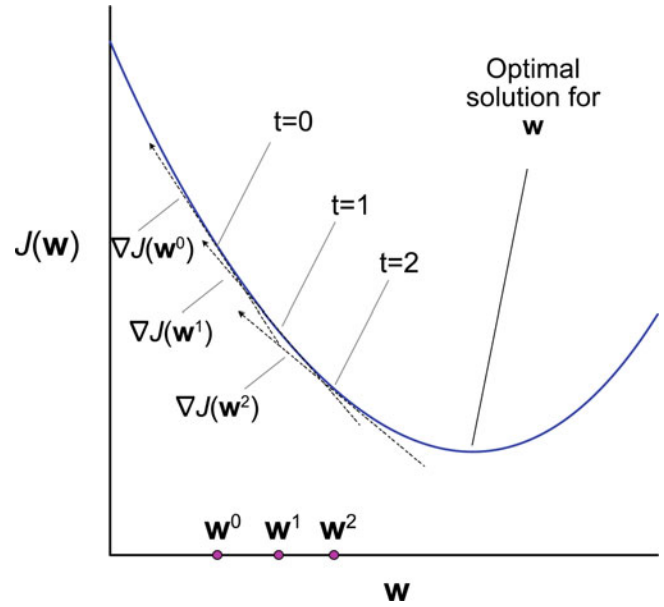
$$y = \mathbf{w}^T \mathbf{x}_i,$$

where $i=1, \dots, n$. For a given input ($y_i, \mathbf{x}_i$), the least mean squares algorithm quantifies the goodness of fit, error, or the cost ($J(\mathbf{w})$) between y_i and predicted estimates for a given $\mathbf{w}$:

$$J(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^m (y_i - \mathbf{w}^T \mathbf{x}_i)^2.$$

The function $J(\mathbf{w})$ calculates the sum of squared errors or residuals between y_i and the predicted estimates. The set of weights $\mathbf{w}$ that produce the smallest value of the function are considered optimal. There is no analytical solution for the minimization of $J(\mathbf{w})$ so other numerical methods must be considered. While stochastic methods to find optimal solutions for $\mathbf{w}$ could be applied at this point, least mean squares often utilizes a gradient descent approach. If we make an initial guess ($\mathbf{w}_0$) for the weights $\mathbf{w}$, using either then entire dataset or a subset thereof (referred to as the training dataset), we can calculate the loss function $J(\mathbf{w})$ but would not be sure how to adjust the values of $\mathbf{w}$ for our next guess. Instead, we can calculate the gradient (∇) of the loss function and determine the direction in which $J(\mathbf{w})$ decreases (Fig. 1). With that information, we can update or generate a new guess for $\mathbf{w}$ ($\mathbf{w}_1$) which makes a step in the direction of the steepest descent of $J(\mathbf{w})$. With this new estimate for $\mathbf{w}$, we can calculate a new $\nabla J(\mathbf{w})$, determine the direction of the steepest gradient, and take the step in this direction to generate yet another updated estimate of $\mathbf{w}$ ($\mathbf{w}_2$). Thus, the general algorithm goes through the following steps:

1. Select an initial estimate for $\mathbf{w}$.
2. Determine the value of the loss function $J(\mathbf{w})$, given $\mathbf{w}$ and the complete set, a subset, or a single random sample drawn from $y, \mathbf{x}$.



Least Mean Squares, Fig. 1 Gradient descent method for approaching the least mean squares (smallest value of $J(\mathbf{w})$) by taking steps in the opposite direction of the gradient $\nabla J(\mathbf{w})$ to provide updated weights.

3. If $J(\mathbf{w})$ is smaller than the convergence criteria, stop and accept the current values for $\mathbf{w}$ as estimates of the optimal solution. Else, move to step 4.
4. Calculate $\nabla J(\mathbf{w})$ using the current estimate of $\mathbf{w}$ ($\mathbf{w}^t$), and create an updated guess ($\mathbf{w}^{t+1}$), based upon a learning rate constant r :

$$\mathbf{w}^{t+1} = \mathbf{w}^t - r \nabla J(\mathbf{w}^t)$$

5. Start back at step 2

The $\nabla J(\mathbf{w})$ function has the form:

$$\nabla J(\mathbf{w}^t) = \left[\frac{\partial J}{\partial w_1}, \frac{\partial J}{\partial w_2}, \dots, \frac{\partial J}{\partial w_n} \right].$$

The partial differential of the least squares algorithm is solvable, allowing for calculation of the elements of $\nabla J(\mathbf{w}^t)$:

$$\frac{\partial J}{\partial w_j} = \frac{1}{2} \sum_{i=1}^m \frac{\partial}{\partial w_j} (y_i - \mathbf{w}^T \mathbf{x}_i)^2.$$

After application of the chain rule and algebra, the equation simplifies to:

$$\frac{\partial J}{\partial w_j} = - \sum_{i=1}^m (y_i - \mathbf{w}^T \mathbf{x}_i) x_{ij},$$

which represents the sum of the products of residual between y_i and the predicted value based on $\mathbf{w}^t$ and the complete set or a subset of $\mathbf{x}$. Thus, while there is no analytical solution to minimize the loss least mean squares function directly, there is an analytical solution of its gradient.

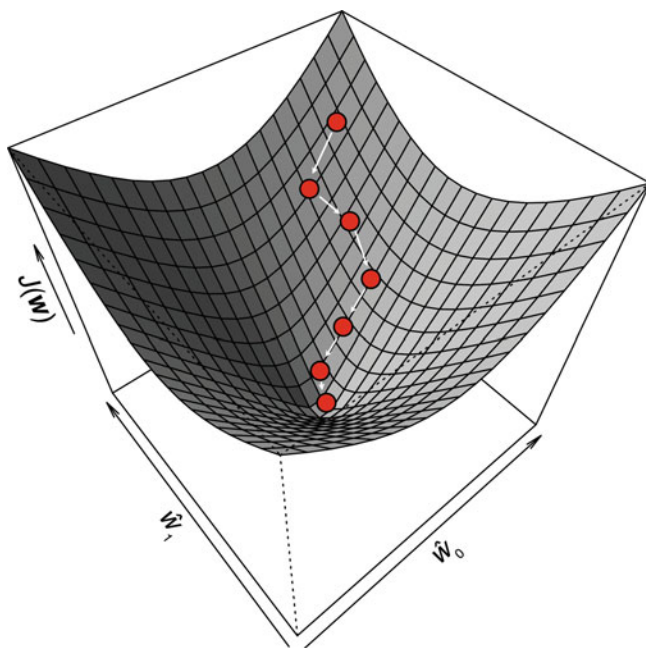
The learning rate r is generally chosen to be a relatively small constant. Generally, if r is chosen to be a large number, the algorithm will converge more quickly than if a small number is used because the size of the steps taken towards the steepest descent of $\nabla J(\mathbf{w})$ will be greater. Because the surface of least mean squares is a quadratic function, it is convex (no local minima) and has a global minimum value. However, if r is chosen to be very large, the iteration could repeatedly overshoot the global minimum leading to difficulty in convergence.

There are several variations in gradient descent solutions for least mean squares minimization. In a batch method the update process is only completed after gradient $\nabla J(\mathbf{w})$ has been calculated for all of the samples in the complete dataset of $y, \mathbf{x}$. At the other extreme, one can calculate $\nabla J(\mathbf{w})$ and update the weights for single, randomly selected samples of $y, \mathbf{x}$. This latter method, known as the incremental or true stochastic gradient descent, provides more accurate updates than the batch method but is numerically expensive. Many authors consider the incremental stochastic gradient descent to be synonymous with the least mean squares method. Conceptually, because the gradient of the loss function is updated after

every single individual sample is drawn, the method does not necessarily follow the steepest descent and produces a more random path towards the global minimum (Fig. 2). This approach allows for decisions about updating the weights to be made on the fly, which has implications for machine learning. A popular trade-off between the batch and incremental stochastic gradient descent methods is that of the mini-batch stochastic gradient descent, where a subset or a fraction of $y, \mathbf{x}$, is randomly drawn and the weights $\mathbf{w}$ are updated only after calculating $\nabla J(\mathbf{w})$ for the entire subset. One can conceptualize the mini-batch stochastic gradient descent as a path towards a minimum that it does not necessarily follow the steepest path, but not as circuitous as that of the incremental gradient descent.

Of note, the example discussed above examines the linear regression between a single output or independent variable (y) and one or more input or independent variables ($\mathbf{x}$). However, the least mean squares algorithm is extremely flexible and can be applied to solving higher-order relationships and/or scenarios with multiple independent variables. As discussed below, geophysical applications of least mean squares have been successful in estimating ideal solutions for nonlinear adaptive signal processing filters.

One serious consideration in the utilization of least mean squares methods is that estimates of optimal solutions can be strongly impacted by outliers and erroneous data because of the squaring of the error in the cost function. That is, highly anomalous or erroneous data can produce outsized loss values, as described for the entry on least squares. The effect of anomalous data can be further compounded when the method is solved using incremental stochastic gradient descent methods or mini-batch stochastic gradient descent with very small subsamples, as the impact of even a single significant outlier could cause $\nabla J(\mathbf{w})$ to be spurious and update the model weights to move away from a local minima. For this reason, datasets being used with least mean squares methods should be thoroughly examined in advance for errors or exceptional univariate and multivariate outliers.



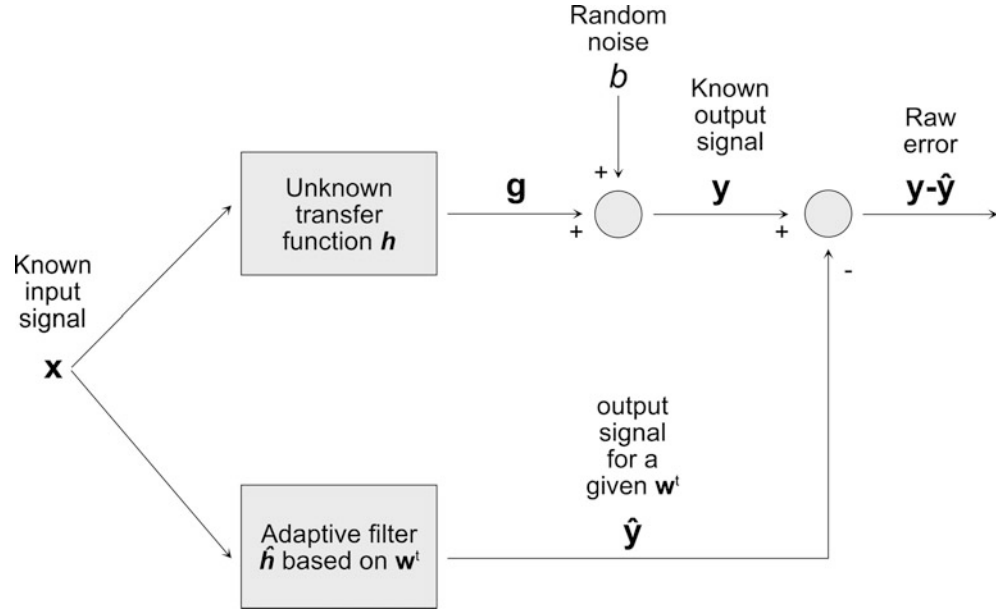
Least Mean Squares, Fig. 2 Cartoon depicting steps taken that update weights $\hat{w}_0$ and $\hat{w}_1$ and corresponding reduction of $J(\mathbf{w})$ during a stochastic descent to the global minimum value. Note that the path is not necessarily the steepest descent, opposite the local gradient, but is rather an adaptive approach based on incremental analysis of input data.

Least Mean Squares as a Geophysical Data Filter

Least mean squares is a method for adaptive data filtering and has been heavily used in the application of electrical signal and geophysical data processing. In this case, the user is attempting to reproduce the effect of an unknown filter h which produces one or more output signals y based on n known input signals x . For example, one can consider a case where two input signals x_1 and x_2 , such as sine functions, are fed into an unknown linear transfer function h that produces an output signal g that is further perturbed by random noise b

Least Mean Squares,

Fig. 3 Schematic of an approach to utilize an adaptive least mean squares filter to emulate an unknown transfer function h . Taking a stochastic gradient descent of the least mean squares of the raw error has the effect of modifying the weights in the adaptive filter until it replicates the effect of h .



$$y_i = g_i + b,$$

where y represents the combination of the unknown filter output with random noise. In order to simulate or mimic the effect of h on x , we can create an auxiliary system which includes an adaptive filter $\hat{h}$ which allows for the application of variable weights w to the input signals (Fig. 3) and produces out $\hat{y}$:

$$\hat{y}_i = w^T x_i = \hat{h}(x_i).$$

We can make adjustments to w for sample i and quantify our accuracy in recreating the effect of h by comparing the least mean square error ($J(w)$) between the output of the two filters:

$$J(w) = \frac{1}{2} \sum_{i=1}^m (y_i - w^T x_i)^2,$$

where $i=1, \dots, n$. As described in the Introduction section, the method works to find optimum estimates of w from input data by determining the gradient $\nabla J(w)$ and updating the weights by moving in steps following the steepest descent of the subset of the training dataset. Assuming convergence toward the global minimum, the application of the final updated weights to the input signals should reasonably approximate the effect of the unknown transfer filter h .

One significant advantage of the least mean squares filter is that it automatically adjusts the unknown filter weights incrementally as new data are added. This has allowed for their

application in the development of automated and adaptive filters such a noise cancellation and adaptive signal equalization (Widrow and Stearns 1985).

ADALINE and Neural Networks

Widrow and Hoff (1960) presented the least mean squares algorithm as part of an adaptive linear (hence, “ADALINE”) machine that automatically classifies input patterns, including those affected by random noise. ADALINE and the similar perceptron (Rosenblatt 1958, 1962) mark early significant advances in the development of neural networks. The basic ADALINE model consists of the following steps:

1. Activate input x_i , where $i=1, \dots, n$.
2. Determine y_i from $y_i = b + w^T x_i$, where b is a random bias term, using the activated input from step 1.
3. Apply the following activation function to determine the output based on y_i :

$$f(y_i) = \begin{cases} 1 & \text{if } y_i \geq 0 \\ -1 & \text{if } y_i < 0 \end{cases}$$

4. Adjust the weights and biases as follows. If the results from step 3 (+1 or -1) does not match the correct classification ($t \in \{-1, +1\}$), update the weight and bias via:

$$w^{t+1} = w^t - r \nabla J(w^t)$$

and

$$b^{t+1} = b^t - r(t - y_i).$$

Else, update the weight and bias via:

$$\mathbf{w}^{t+1} = \mathbf{w}^t$$

and

$$b^{t+1} = b^t.$$

5. Repeat steps 1–4 using the updated weights and biases, unless the largest change in weights during learning is smaller than the specified tolerance.

Because y_i is a linear function, even in cases where $t = f(y_i)$, there is still an error and updates to the bias and weight occur, which allows for more rapid convergence of ADALINE versus the perceptron. The method was eventually updated to create multiple layers of ADALINE neurons, an early form of deep neural networks (Ridgeway 1962).

Summary

Least mean squares methods, depending on the exact method of numerical determination, are flexible tools capable of solving a broad range of problems. The ability of least mean squares algorithm to adaptively update weights as data are sequentially processed produces a useful solution to many problems, making them ideal for noise reduction and other geophysical methods. Moreover, related algorithms such as ADALINE serve as important milestones in the development of modern machine learning methods.

Cross-References

- [Least Squares](#)
- [Machine Learning](#)
- [Optimization in Geosciences](#)
- [Regression](#)

Bibliography

- Ridgeway WC III (1962) An adaptive logic system with generalizing properties. PhD thesis, Stanford University, Stanford, 77 pp
- Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev* 65(6):386
- Rosenblatt F (1962) Principles of neurodynamics; Perceptrons and the theory of brain mechanisms. Spartan Books, Washington, DC, 616 pp

Widrow B, Hoff ME (1960) Adaptive switching circuits. Solid-state electronics laboratory Technical Report 1553-1. Stanford University, 34 p

Widrow B, Stearns SD (1985) Adaptive signal processing. Prentice-Halls, Englewood Cliffs, 224 pp

Least Median of Squares

Sara Taskinen and Klaus Nordhausen

Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland

Definition

The least median of squares (LMS) is a regression method introduced in Rousseeuw (1984) and further developed in Rousseeuw and Leroy (1987). The LMS estimator minimizes the median of the squared residuals. The method is shown to be highly robust having a breakdown point of 50% which is the highest possible. Because of this, the LMS regression method is fitted to the main bulk of the data while ignoring the rest of the data. Hence, atypical observations are easily identified. The method suffers from very low efficiency and is therefore recommended to be used primarily as an exploratory tool to identify regression outliers.

Introduction

Consider a data set consisting of n independent and identically distributed (iid) observations $(\mathbf{x}, y)$, $= 1, \dots, n$, where $\mathbf{x} = (x_1, \dots, x_p)'$ and y are the observed values of predictor variables and response variables, respectively. The data are assumed to follow a linear regression model

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i,$$

where the p -vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ contains the unknown regression coefficients to be estimated based on the data and the errors ε are iid and independent of $\mathbf{x}$.

The widely used least squares (LS) estimator for $\boldsymbol{\beta}$ was proposed in early 1800 by Gauss and Legendre (Stigler 1981). The LS estimate of $\boldsymbol{\beta}$ is the $\hat{\boldsymbol{\beta}}$ that minimizes

$$\sum_{i=1}^n r_i^2, \quad (1)$$

that is, the sum of squared residuals $r_i = y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}$. The popularity of the estimate arises mainly from its nice computational and theoretical properties. The LS estimator has a

closed form solution and can be computed explicitly from data using simple matrix algebra. The estimator has also a desired property of being regression, scale and affine equivariant meaning that it is known how the estimate changes under different transformations of the data (Rousseeuw and Leroy 1987). If the error distribution is Gaussian, then the LS estimate is optimal. It is however well known that the LS estimator is extremely sensitive to outlying observations and even a single outlier can have a huge effect on the estimate. Notice that with outlier we mean here regression outlier, that is, an observation, $(\mathbf{x}'_i, y)$, where the relationship between $\mathbf{x}$ and y differs from that of the majority of the data. For the identification of different types of regression outliers, see Rousseeuw and Leroy (1987), for example.

To improve the LS estimator, several robust regression estimators have been proposed in the literature. Many of the proposed estimators are obtained by replacing the squared residuals by another function of the residuals. The least absolute deviation (LAD) estimator or L_1 regression estimator is obtained by minimizing the sum of the absolute values of the residuals $\sum_{i=1}^n |r_i|$, and a more general family of estimators, that is, the M -estimators minimize $\sum_{i=1}^n \rho(r_i)$, where ρ is a symmetric function ($\rho(-x) = \rho(x)$) with a unique minimum at zero (Huber 1973). Such regression estimates are robust against outliers in response variables (y 's); however, they cannot tolerate outliers in the predictors ($\mathbf{x}_i$'s). Notice also that such outliers, which are also called as "leverage points", cannot always be identified using residuals based on the methods described above as the regression estimates may be strongly affected by the outliers yielding small residual values.

Least Median of Squares Estimator

As mentioned in the previous section, a single outlier can strongly affect the estimates that are obtained by minimizing the sum of the squared residuals (or other function of the residuals). A regression estimator that is very robust against outliers in responses as well as outliers in predictors is obtained by replacing the sum in Eq. (1) with the median (Rousseeuw 1984). The least median of squares (LMS) estimator thus minimizes the median of the squared residuals

$$\text{Median}_i r_i^2. \quad (2)$$

In the case of a single predictor the solution can be interpreted as the midline between the two closest parallel lines which contain half of the data between them. In the case of an intercept only model, i.e., when just the location of the y 's is to be computed, the estimate is known as the

Shorth estimate, the midpoint of the shortest interval to contain 50% of the data. If the amount of the points outside the parallel lines is to be reduced the method is known as least α -quantile estimator (Rousseeuw and Leroy 1987). It is shown in Rousseeuw (1984) that there always exists a solution for Eq. (2). However, the objective function may have many local minima (Steele and Steiger 1986). For large p the computation requires some subsampling and may be time-consuming (Souvaine and Steele 1987; Joss and Marazzi 1990). See also a cautionary note on the instability of the LMS solution in Hettmansperger and Sheather (1992).

The LMS estimate for the regression coefficient is one of the few robust estimates which does not need an initial scale estimate. The same property holds for the LS estimate. Based on the LMS estimate the scale of the residuals ε can be estimated as follows. First an initial scale s_0 is obtained as

$$s_0 = 1.4826 \left(1 + \frac{5}{n-p} \right) \sqrt{\text{Median}_i r_i^2},$$

where the first constant term is for achieving consistency in case the errors follow a Gaussian distribution, and the second constant term is a finite sample correction. The initial scale is then used to define the standardized residuals $r_{st} = r/s_0$, $= 1, \dots, n$. A general rule is that observations which have absolute standardized residuals smaller than 2.5 are considered as good data points. The actual scale based on LMS, σ_{LMS} , is then the standard deviation of the (unstandardized) residuals of the good data points.

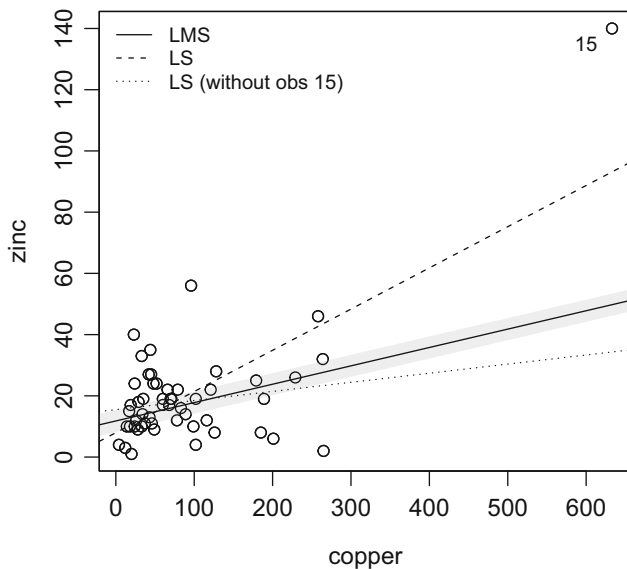
Often in robustness studies it is of interest to evaluate what is the amount of outliers the estimator can handle. To formalize this, we need a notion of breakdown point. In the next we shortly recall the definition of finite-sample breakdown point as introduced by Donoho and Huber (1983). Let a sample of n data points be $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$, where $\mathbf{z} = (\mathbf{x}', y)'$, and write $\hat{\boldsymbol{\beta}}(\mathbf{Z})$ for an regression estimate based on the data. Now consider all possible contaminated samples $\mathbf{Z}'$ that are obtained by replacing any m of the original data points by arbitrary values. Then denote the maximum bias caused by such contamination as

$$\text{bias}(m; \hat{\boldsymbol{\beta}}, \mathbf{Z}) = \sup_{\mathbf{Z}'} \|\hat{\boldsymbol{\beta}}(\mathbf{Z}') - \hat{\boldsymbol{\beta}}(\mathbf{Z})\|.$$

Infinite bias means that m outliers can have an arbitrarily large effect on $\hat{\boldsymbol{\beta}}$ and we say that the estimator "breaks down". Further, the finite-sample breakdown point is the minimum fraction of contamination that will make the estimator to break down. Mathematically, this can be expressed as

$$\varepsilon^*(\hat{\boldsymbol{\beta}}, \mathbf{Z}) = \min_{1 \leq m \leq n} \left\{ \frac{m}{n} : \text{bias}(m; \hat{\boldsymbol{\beta}}, \mathbf{Z}) = \infty \right\}.$$

Rousseeuw (1984) showed that if $p > 1$ and the observations are in general position, meaning that any p observations give a unique determination of β , then the finite sample breakdown point of the LMS estimator is $([n/2] - p + 2)/n$, where $[n/2]$ denotes the largest integer less than or equal to $n/2$. When considering the limit $n \rightarrow \infty$ (with p fixed), the asymptotic breakdown point of the LMS estimator is 50%,

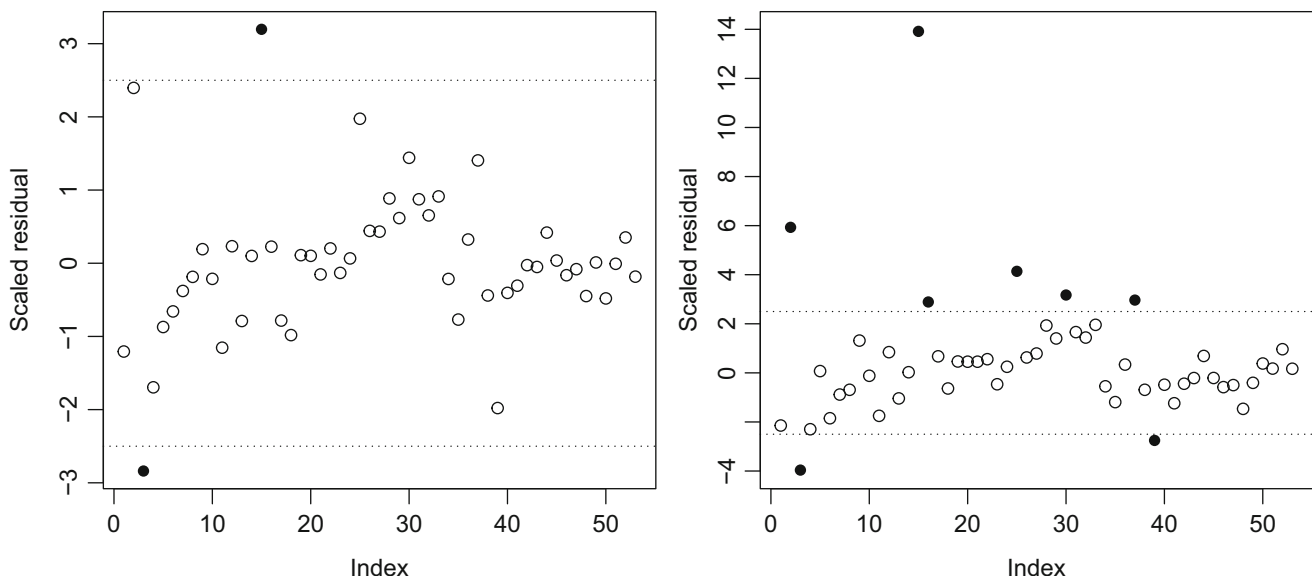


Least Median of Squares, Fig. 1 Zinc and copper contents of 53 samples of rocks and the regression lines based on least median of squares regression (LMS) and least squares regression (LS) applied to the original data and data with observation 15 as rejected. The grey area around the LMS regression line contains the “inner” 50% of the data points

that is, the best one can expect. The LMS thus extends the 50% breakdown property of the median to the regression setting. Another tool to study the robustness properties of an estimator is the influence function which measures the change of the estimator when an outlying observation is introduced. As the LMS estimator converges at the rate of $n^{-1/3}$ and does not have a normal limiting distribution, its influence function is not well-defined and the estimator suffers from low efficiency. Hence, Rousseeuw (1984) recommends that LMS should not be used for inferential purposes, but as an exploratory tool for diagnosing regression outliers. LMS is for example implemented in the R package MASS (Venables and Ripley 2002) as the function `lmsrob`.

Example

Smith et al. (1984) measured the chemical contents of 53 samples of rocks in Western Australia. The data are available as the mineral data set in R (R Core Team 2020) package `RobStatTM` accompanying the book by Maronna et al. (2018). In Fig. 1 we plot the chemical contents (in parts per million) of zinc vs. the chemical contents of copper. We notice that the observation 15 stands out as a clear outlier and the LS estimator is highly influence by this observation. If we reject observation 15, then the LS fit seems to model correctly the main bulk of the data. In Fig. 2 the scaled residuals based on the LS fit are plotted. As the observations with absolute value of standardized residual exceeding 2.5 are identified as outliers (Rousseeuw and Leroy 1987), the horizontal lines at 2.5



Least Median of Squares, Fig. 2 Scaled residuals based on LS regression (left) and LMS regression (right) applied to mineral data. An observation with an absolute value of standardized residual larger than 2.5 is identified as an outlier and marked using filled black dots

and -2.5 are also added in the figure. The plot indicates that all scaled residuals are very small and although the LS method reveals two outliers, none of them is classified extremely atypical.

The LMS estimator is not affected by observation 15 and the resulting fit is nearly equivalent to the LS fit obtained when the observation 15 is rejected. The LMS estimates for intercept and slope are 11.79 and 0.06, respectively (as compared to the LS estimates 7.96 and 0.13). The scaled residuals in Fig. 2 indicate that the method identifies altogether eight outliers and observation 15 clearly stands out in the residual plot.

To have an idea about the model fit, Rousseeuw and Leroy (1987) define the coefficient of determination in the case of LMS as

$$R_{LMS}^2 = 1 - \left(\frac{\text{Median}_i |r_i|}{\text{Mad}_i y_i} \right)^2,$$

where $\text{Mad } y$ denotes the median absolute deviation. For our example data $R_{LMS}^2 = 0.59$ which is much better than the regular coefficient of determination of the LS fit of $R_{LS}^2 = 0.46$.

Finally notice that in this simple example it was easy to identify an outlier just by plotting the data. However, outlier identification becomes more difficult when the number of predictors increase and cannot be performed just by plotting the data. In such a case one should proceed as guided in Rousseeuw and Leroy (1987), for example.

Summary and Conclusion

The paper introducing the LMS Rousseeuw and Leroy (1987) is considered as a seminal paper in the development of robust statistics and is, for example, appreciated in the Breakthroughs in Statistics series (Kotz and Johnson 1997). LMS demonstrates that high breakdown regression can be performed with outliers present in response and predictors. However, due to its low efficiency and the development of better robust regression methods such as MM-regression (Yohai 1987), LMS is nowadays mainly used for exploratory data analysis and as an initial estimate for more sophisticated methods.

Cross-References

- Iterative Weighted Least Squares
- Ordinary Least Squares
- Regression

Bibliography

- Donoho DL, Huber PJ (1983) The notion of breakdown point. In: Bickel PJ, Doksum KA, Hodges JL (eds) *A Festschrift for Erich L. Lehmann*. Wadsworth, Belmont, pp 157–184
- Hettmansperger TP, Sheather SJ (1992) A cautionary note on the method of least median squares. *Am Stat* 46(2):79–83. <https://doi.org/10.2307/2684169>
- Huber PJ (1973) Robust regression: asymptotics, conjectures and Monte Carlo. *Ann Stat* 1(5):799–821. <https://doi.org/10.1214/aos/1176342503>
- Joss J, Marazzi A (1990) Probabilistic algorithms for least median of squares regression. *Comput Stat Data Anal* 9(1):123–133. [https://doi.org/10.1016/0167-9473\(90\)90075-S](https://doi.org/10.1016/0167-9473(90)90075-S)
- Kotz S, Johnson NL (1997) *Breakthroughs in statistics*, vol 3. Springer, New York
- Maronna RA, Martin DR, Yohai VJ, Salibián-Barrera M (2018) *Robust statistics: theory and methods (with R)*, 2nd edn. Wiley, Hoboken
- R Core Team (2020) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- Rousseeuw PJ (1984) Least median of squares regression. *J Am Stat Assoc* 79(388):871–880. <https://doi.org/10.1080/01621459.1984.10477105>
- Rousseeuw PJ, Leroy A (1987) *Robust regression and outlier detection*. Wiley, Hoboken, pp 1–335
- Smith RE, Campbell NA, Litchfield R (1984) Multivariate statistical techniques applied to pisolithic laterite geochemistry at golden grove, Western Australia. *J Geochem Explor* 22(1):193–216. [https://doi.org/10.1016/0375-6742\(84\)90012-8](https://doi.org/10.1016/0375-6742(84)90012-8)
- Souvaine DL, Steele MJ (1987) Time- and space-efficient algorithms for least median of squares regression. *J Am Stat Assoc* 82(399):794–801. <https://doi.org/10.1080/01621459.1987.10478500>
- Steele JM, Steiger WL (1986) Algorithms and complexity for least median of squares regression. *Discret Appl Math* 14(1):93–100. [https://doi.org/10.1016/0166-218X\(86\)90009-0](https://doi.org/10.1016/0166-218X(86)90009-0)
- Stigler SM (1981) Gauss and the invention of least squares. *Ann Stat* 9(3):465–474. <https://doi.org/10.1214/aos/1176345451>
- Venables WN, Ripley BD (2002) *Modern applied statistics with S*, 4th edn. Springer, New York
- Yohai VJ (1987) High breakdown-point and high efficiency robust estimates for regression. *Ann Stat* 15(2):642–656. <https://doi.org/10.1214/aos/1176350366>

Least Squares

Mark A. Engle

Department of Geological Sciences, University of Texas at El Paso, El Paso, TX, USA

Definition

Least squares is a numerical method to estimate ideal solutions for model fit parameters in overdetermined systems by minimizing the sum of squares distance of the residuals between one or more variables for each data point and the corresponding model estimate(s). The approach only

considers uncertainty or error in the variables being used in the determination of residuals (e.g., response variables), but not in the other variables (e.g., predictor variables). The method is commonly used to approximate best-fitting solutions to both linear and nonlinear regression systems, but is also applied to more general applications including machine learning.

Introduction

Consider a set of n data points composed of one or more predictor variables (x_i) and one or more response variables (y_i), where $i = 1, \dots, n$. Myriad methods exist to determine optimal parameters for numerical models estimating values of the response variable(s) ($\hat{y}_i$) from the predictor value(s) for overfitted systems. One such method, least squares, is an efficient approach that utilizes the set of residuals or errors (e_i), defined as difference between the measured and predicted response values:

$$e_i = y_i - \hat{y}_i,$$

but can more generally refer to any variables or set of variables that the model is attempting to fit. Regardless of the exact mathematical form of the prediction model, the least squares method evaluates candidates for the m model parameters (β_k), where $k = 1, \dots, m$, by calculating the sum of squares residuals or errors (SS_e) for the data set:

$$SS_e = \sum_{i=1}^n e_i^2.$$

The set of m values for β_k that produces the lowest SS_e is considered the optimal solution. A generalized solution to the minimization of SSE for any numerical systems is to set a gradient with respect to β_k equal to zero:

$$\frac{\partial SS_e}{\partial \beta_k} = 0.$$

Linear Least Squares Regression

One of the oldest and most well utilized applications of least squares methods is linear least squares regression, which was developed by Legendre (1805) and Gauss (1809). Linear least squares is a commonly applied method to generate linear estimates of y_i in overdetermined mathematical systems, using m number of predictor variables. The method is pervasive throughout the earth sciences and used in a broad range of topics from evaluating potential controls on geologic processes to calibrating instrumentation and numerical models. Values of y_i are modeled from linear combinations of the

m predictor variables (x_{ik}) and model parameters β_k , where $k = 1, \dots, m$ with a generalized form of:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_m x_{im} + \varepsilon,$$

where ε represents model error or uncertainty. In the case of simple linear least squares (also called ordinary least squares), based on a single x_i , modeled values of $\hat{y}_i$ are calculated from estimates of model parameters $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

and the residuals are calculated as described in the “[Introduction](#)” section. As an example, let us say we want to provide rapid screening results in the field by developing a model to predict the density of water samples (typically a laboratory measurement) from their specific conductance measurements (a field measurement) for brackish groundwater samples from the Dockum aquifer in Texas, USA (data subset from Reyes et al. 2018). There exists an analytical solution for $\hat{\beta}_0$ and $\hat{\beta}_1$ in simple linear least squares regression that minimizes SS_e :

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where

$$SS_x = \sum_{i=1}^n (x_i - \bar{x})^2$$

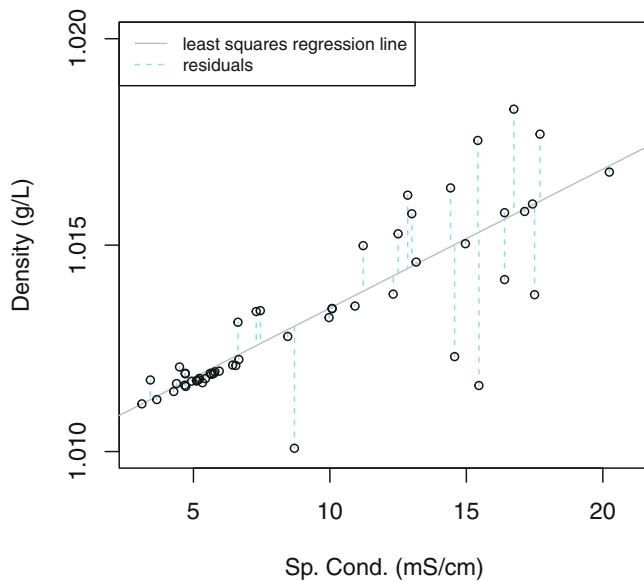
and

$$SS_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

SS_x is the sum of squared distances between the predictor values and its mean, which if divided by the degrees of freedom of x_i is equal to its variance. Similarly, SS_{xy} is the sum of squares for the x_i and y_i cross products. The sum of squares of y_i (SS_y) can be divided into the sum of squares accounted for by the model (SS_p) versus that not explained by the model, which is equal to SS_e or the sum of squares of the residuals:

$$SS_y = SS_p + SS_e,$$

and can be further explored using Analysis of Variance (ANOVA) methods. It is then relatively intuitive to develop



Least Squares, Fig. 1 Simple linear least squares regression model to predict density from specific conductivity data for groundwater samples from the Dockum Aquifer, Texas, USA. (Data subset from Reyes et al. 2018)

an estimator of fit quality (the coefficient of determination, or R^2) calculated as portion of SS_y explained by the model:

$$R^2 = \frac{SS_p}{SS_y}.$$

Relationships between x_i , y_i , $\hat{y}_i$, and e_i , including the least squares regression line for the above example are illustrated in Fig. 1. In this case, $SS_y = 2.05 \times 10^{-4}$, $SS_p = 1.42 \times 10^{-4}$, and $SS_e = 6.32 \times 10^{-5}$, producing an R^2 value of 0.69, meaning that nearly 70% of the sum of squares in the dependent variable is accounted for by the simple linear least squares regression model. Many texts suggest that in order for a linear regression model to satisfy its mathematical assumptions, the residuals must be independent, constant across the range of predictor variables (homoscedastic), and normally distributed. In reality, however, the requirement to meet assumptions varies depending on the purpose of the linear least squares regression model. For instance, Helsel et al. (2020) argue that if the intended use of the model is simply to predict y_i from x_i , then the only requirement is that the relationship between the predictor variable(s) and the response variable is linear and that the data used in the analysis are representative. Assumptions that the residuals be normally distributed are reserved for instances of hypothesis testing or calculating confidence intervals. Depending on the final purpose of the model predicting density from specific conductance, additional investigation about the nature of the residuals may or may not be required.

Least squares solutions are used to estimate model parameters for many types of linear regression models including

linear regression, linear multivariate regression (many predictor variables and one response variable), and linear multivariable regression (many predictor and response variables). One documented problem with least squares models utilizing multiple predictor and/or response variables is that if sets of variables are strongly correlated (i.e., multicollinearity), the model results can be misleading. Methods such as stepwise, lasso, and ridge regression have been developed to determine variables most important in model prediction, and subsequently remove the remaining variables, as one method to overcome multicollinearity problems. This approach has the advantage that a more parsimonious model is generated, requiring fewer model inputs, than one utilizing all of the available variables. An alternative approach, utilized by least partial squares and principal component regression methods, is to extract latent components from the predictor variables. The latent components, which should have minimal correlation to one another, are used as predictor variables in the regression model instead of the original predictor variables. These methods make more efficient use of the data, while reducing multicollinearity and the number of parameters in the final model.

Nonlinear Least Squares Regression

Linear regression methods are most useful when the relationship between the predictor variable(s) and the response variable(s): (1) follows a hyperplane and (2) can be reasonably recreated from linear combinations of the response variables. In some cases, the relationships between the original variables is not linear but can be approximated as such by transforming one or more variables and processing the transformed data with linear methods. However, in instances where variable transformation fails to produce linear relationships between x_i and y_i , or residuals vary in a nonrandom way around x_i , nonlinear least squares regression methods can be attempted by utilizing a nonlinear prediction function $f(x, \beta)$. In nonlinear least squares regression, there is no closed form analytical solution and optimal values of regression parameters are estimated iteratively, by adjusting the m model parameters at iteration t by a known quantity ($\Delta\beta$) for the next time step ($t + 1$):

$$\beta_k^{t+1} = \beta_k^t + \Delta\beta.$$

If the model contains m predictor variables, then the goal of the iterative approach is to find model parameters β_k that produce a value of zero for the following equation:

$$\frac{\partial S}{\partial \beta_k} = -2 \sum_{i=1}^m r_i \frac{\partial f(x_i, \beta)}{\partial \beta_k} = 0.$$

The equation is numerically solved through the application of iterative convergence methods such as the Gauss-Newton and Gauss-Siedel algorithms. This generalized solution approach allows for the application of least squares model-fitting approaches to be applied to a range of other applications and models. However, unlike linear systems, solutions to nonlinear least squares models are not unique as multiple local minima can exist. Thus, there is no guarantee that solutions are optimal.

Non-regression Least Squares Methods

Least squares methods require only the ability to measure residuals between response variable(s) and their modeled values and one or more model parameters to adjust and make efficient use of the data. This makes least squares attractive for a wide range of geologic problems. Some of the earliest applications of least squares methods were fitting models to estimate the size and shape of the earth and moon and of the orbits of the planets (Nieverfelt 2000), for instance. Least squares models have also made in-roads as an efficient solution for machine learning techniques, particularly variations of support vector machines. Support vector machines are non-probabilistic supervised classification models that define classifier functions (typically pairs of hyperplanes) that best separate each defined group of data, or projected data, in the dataset. For a given training data set $\{x_k, y_k\}$, for $k = 1, \dots, l$, where x_k are real data corresponding and y_k is the group classifier ($y_k \in \{-1, +1\}$ in the case of 2 groups), if there exists no hyperplane or set of hyperplanes that can perfectly separate the data, a slack variable (ζ_k) is added to the relationship such that:

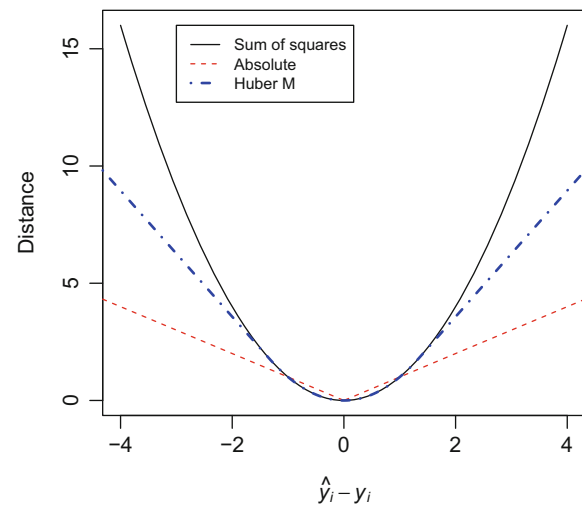
$$y_k = w^T \varphi(y_k) + b \geq 1 - \zeta_k$$

and $\zeta_k > 0$. w^T is a transformed classification vector, $\varphi(y_k)$ is nonlinear transformation of y_k into high level space created using the so-called kernel trick, and b is constant. A common method to find estimates of optimized model parameters ($\hat{w}^T$ and $\hat{b}$), which represent the hyperplane exhibiting the largest region between the groups of samples (i.e., margin) is to construct Lagrange multipliers, producing a quadratic optimization equation. Suykens and Vandewalle (1999) reformulated the problem into a least squares model producing, after utilization of Lagrange multipliers, a linear system to be solved. Least squares methods have made in-roads into other machine learning methods as well; Mao et al. (2017) applied a least squares approach to the discriminator function in generative adversarial networks to overcome the vanishing gradient problem using descent methods during learning.

Thus, despite the historic origin of the method, least squares still has utility in next generation analysis and machine learning tools.

Robust Least Squares Methods

One criticism of least squares methods is that calculating the square of the residuals, as opposed to other distance measurements, biases the results towards data points that are far away from the predicted values. Comparison of sum of squares versus absolute residual distances (as is used in L_1 linear least squares regression, which estimates the median value of y_i for a given x_i) shows that the former heavily weights points at distances when residuals are >1 while it gives relatively less weight when residuals are <1 (Fig. 2). As a result, several least squares-based robust and resistant regression techniques have been developed to reduce the impact of highly influential points. Earlier robust methods, such as Huber's M-estimator (Huber 1964), limit the distances above a specific residual threshold but do not consider errors in the predictor variables. Huber's method involves choosing a tuning parameter ranging from 0 (producing a form equivalent to L_1 regression) to ∞ (equivalent to least squares linear regression), which determines the efficiency of the method (Fig. 2). Both least squares regression and Huber's M-estimator both exhibit breakdown points of 0; even a single erroneous or problematic data point can impact the regression result. Newer sum of squares-based regression methods, such as least trimmed squares (Rousseeuw and Leroy 1987), were developed to



Least Squares, Fig. 2 Comparison of sum of squares, absolute, and Huber's M-estimator distance (using a tuning parameter of 1.345) based on residuals

adapt to outliers in both x_i and y_i and have breakdown points as high as 50%. Numerically, the objective function of least trimmed squares regression is nearly identical to linear least squares regression, except that: (1) the calculation is performed on a subset of the data (typically 50–75%) that exhibits the lowest minimum covariance determinant and (2) there exists no analytical solution. Comparison between least trimmed squares and least mean squares regression show that in instances with well-behaved data, the methods will produce nearly identical results. However, in cases where even a small proportion of outliers are present, as is common in many geological datasets, they are much more likely to have a significant impact on the optimized model parameters between the two methods, largely due to the difference in breakdown points. With improvements in computing power and optimization procedures and the similarity between models from traditional and robust methods, there is some logic in utilizing robust least squares methods as the default approach. Use of robust and resistant least squares methods, arguably, opens scientists and engineers to less potential criticism relative to an alternative approach of removing seemingly problematic points from the dataset. Work on methods to improve robust methods for least squares models and methods is continuing.

Summary

Least squares methods have been applied to a wide range of applications in the earth and planetary sciences, dating back more than two centuries. The well-studied behavior of least squares models, their efficient use of the data, and their minimal data requirements make them an attractive choice for model fitting in overdetermined systems. Though mostly known for their use in linear regression models, least squares methods have also been successfully applied in nonlinear model fitting and are finding revitalization in machine learning algorithms. Users of least squares models should be aware of the heavy weight the method places on data points with large residuals and that for some applications robust alternatives are available to prevent deleterious modeling results from erroneous or problematic data.

Cross-References

- [Iterative Weighted Least Squares](#)
- [Least Median of Squares](#)
- [Ordinary Least Squares](#)
- [Regression](#)
- [Support Vector Machines](#)

Bibliography

- Gauss CF (1809) *Theoria motus corporum coelestium*. Perthes, Hamburg
- Helsel DR, Hirsch RM, Ryberg KR, Archfield SA, Gilroy EJ (2020) Statistical methods in water resources, Chapter 4-A3. In: *Techniques of methods*. U.S. Geological Survey, Reston, 458 pp
- Huber PJ (1964) Robust estimation of a location parameter. *Ann Math Stat* 35:73–101
- Legendre A-M (1805) *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot, Paris
- Mao X, Li Q, Xie H, Lau RY, Wang Z, Smolley SP (2017) Least squares generative adversarial networks. In: *Proceedings of IEEE international conference on computer vision*, pp 2794–2802
- Nieverfelt Y (2000) A tutorial history of least squares with applications to astronomy and geodesy. *J Comput Appl Math* 121:37–72
- Reyes FR, Engle MA, Jin L, Jacobs MA, Konter JG (2018) Hydrogeochemical controls on brackish groundwater and its suitability for use in hydraulic fracturing: the Dockum Aquifer, Midland Basin, Texas. *Environ Geosci* 25:37–63
- Rousseeuw PJ, Leroy AM (1987) *Robust regression and outlier detection*. Wiley, 329 pp
- Suykens J, Vandewalle J (1999) Least squares support vector machine classifiers. *Neural Processing Lett* 9:293–300

LiDAR

Jaya Sreevalsan-Nair

Graphics-Visualization-Computing Lab, International
Institute of Information Technology Bangalore, Electronic
City, Bangalore, Karnataka, India

Definition

Light Detection and Ranging (LiDAR) is a remote sensing technique which uses electromagnetic radiation, i.e., light, to measure the distance of objects of interest from the sensor. The time taken for the reflected radiation to return to the sensor gives the distance from the sensor. The reflected radiation, in comparison to the emitted radiation, loses energy during its onward and return travel. The loss in energy may be attributed to the optical properties of the material of the objects it is incident upon, as well as, the medium through which it has traveled. After normalizing for the latter, the former is useful information for semantically analyzing the acquired data, and reconstructing the region that has been scanned.

LiDAR is an active remote sensing system, as it generates energy, which is emitted in the form of light from a laser sensor at a high rate. The LiDAR instrument has a transmitter to emit light pulses, and receiver to record the reflected pulses. The distance captured from the time taken by the reflected rays is corrected using the geolocation information of the sensor. This distance is the elevation in the case of airborne

sensors or aerial laser scanners, but is the distance to the sensor in the case of terrestrial or mobile sensors, mounted on static or dynamic platforms. The GPS (global positioning system) and the IMU (internal measurement unit) in the LiDAR instrument are used for obtaining accurate positional coordinates and orientation of the instrument to find absolute position coordinates of the objects of interest. Depending on the application and type of LiDAR system used, the objects of interest include objects in the scanned scene or topography, including buildings, vehicles, vegetation, and people. The advantages of LiDAR include high-density, high-accuracy discretized point observations, useful for generating digital elevation models (DEM), and diverse applications. LiDAR technology is versatile to be used for land management, hazard assessment (e.g., flooding), forest science, ocean sciences, etc.

The measurements acquired from the LiDAR system are stored and managed as a point cloud, which is a set of position coordinates (x, y, z) from where the light pulses have been reflected. The LiDAR sensor captures the reflected light energy in two different ways, namely, the discrete returns and the full waveform systems. The discrete LiDAR records *returns*, which are points in the peaks of the waveform. Such a system can record one to four returns. However, a full waveform LiDAR system records more information through a full distribution of reflected energy. Both systems record discrete points, which are stored in .las file format. GeoTIFF formats are also used for storing LiDAR data. This data generally stores position coordinates of the three-dimensional (3D) points in the point cloud, intensity (or remission), and optionally, semantic class labels. Semantic class labels refer to the class of objects the point belong to, e.g., buildings and road in urban data.

Overview

LiDAR technology is broadly categorized as airborne (aerial laser scanning (ALS)) and terrestrial (terrestrial laser scanning (TLS)) sensors, based on its data acquisition mode. Airborne LiDAR sensors are usually used for scanning larger areas, and are carried by low flying aircrafts or unmanned aerial vehicles. Terrestrial LiDAR is collected from land, from a fixed location or when mounted on land vehicles. Terrestrial LiDAR can scan 360° views. Airborne LiDAR is apt for scanning larger spatial scales, such as regions at city-level or for coastal projects, giving a bird's eye view, in comparison to that of terrestrial LiDAR. Terrestrial laser scanning covers regions where airborne sensors cannot access, such as moving traffic, building interiors, etc. Combining airborne and terrestrial LiDAR remote sensing can provide complete details of regions of interest.

Airborne LiDAR is subclassified as topographic and bathymetric, based on its applications. Topographic LiDAR

uses near-infrared light pulses to sense the land, and bathymetric LiDAR uses green light to penetrate through volume of water to sense the sea-floor or riverbed. The former has been originally used for studying topography or structures under dense tree cover or canopy, and in mountainous regions. The topographic LiDAR has also been used for computing accurate shorelines and flood maps. The bathymetric LiDAR has been widely used for studying underwater terrains, and varied hydrographic applications. The topographic and bathymetric maps essentially are *elevation* maps.

Terrestrial LiDAR, used for scanning built-up area, are of two types – static and mobile LiDAR. Terrestrial LiDAR measures *distance* maps, which measures distances from the sensor in its 360° view. TLS scanners enable scanning above ground using fast and dense sampling. They capture data with low oblique angle of signals, and the light signals which are reflected by obstacles also include shadows in 3D point clouds. To mitigate such effects, TLS scanning is done from multiple viewpoints (Baltensweiler et al. 2017). The irregular point distribution and shadowing by obstacles introduce challenges in separation of ground and non-ground points in TLS data, more than so in ALS data.

Applications

The LiDAR point clouds are used for generating digital elevation models (DEM), and TLS data is used for digital soil mapping, among other applications (Baltensweiler et al. 2017). The term “DEM” loosely include both digital surface model (DSM) and digital terrain model (DTM), of which the latter is specific to terrains (Chen et al. 2017). The spatial resolution of the DEM generated from ALS is within 0.1–1.5 m. The root mean square errors (RMSE) of the DEMs are dependent on the site/region types, and are of the order of 0.007–0.525 m and 0.286–0.306 m for TLS and ALS, respectively, considering at to uneven slope terrains.

Coastal mapping is done extensively using LiDAR technology. This includes the use of terrestrial LiDAR to provide dry-beach DEM, and use of airborne LiDAR bathymetry for shoreline mapping by gathering data for regional coastal geomorphology and identifying shallow-water depth measurement (Irish and Lillycrop 1999). ALS data has been used for qualitative and quantitative channel dynamics in fluvial geomorphology, applicable to river catchment. The applications of ALS data also include ground filtering, building detection and reconstruction, tree classification, road detection, and semantic segmentation and classification (Lohani and Ghosh 2017).

The use of Unmanned Aerial Vehicles (UAV), such as drones, has increased in various applications, including archeological remote sensing. Similarly, the modern usage of mobile LiDAR in sensing has been used widely in autonomous driving to support data analysis for blind-spots, lane

drifting, cruise control, etc. The interest in the area of vehicle LiDAR also include the development of varied computing platforms on the edge, and in the fog and the cloud. Overall the applications of LiDAR measurements are relevant based on the spatial resolution and accuracy values.

Point Cloud Analysis: With a Focus on ALS

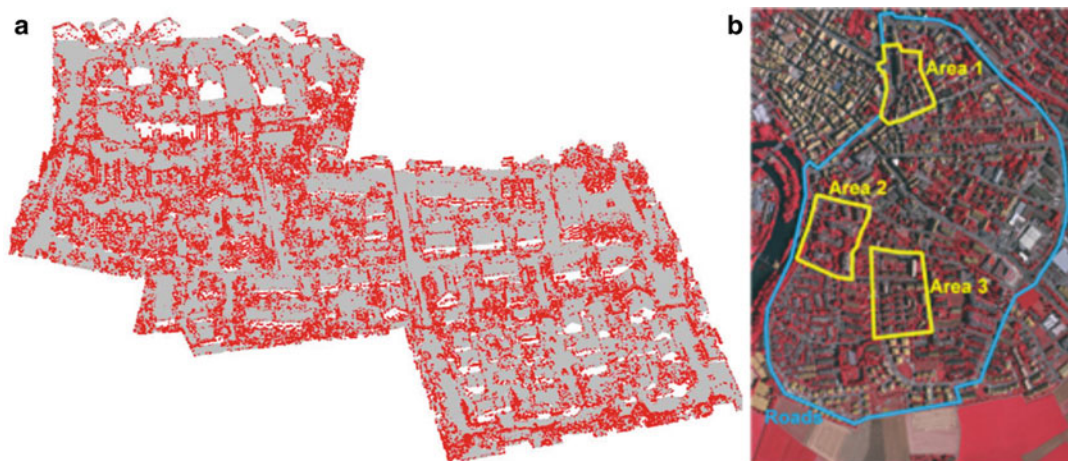
Airborne LiDAR or ALS point clouds include complete scans of top-view of sites, as shown in Fig. 1, which exploit geometric information for automated analysis. This geometric information is extracted using spatially local neighbors, as the LiDAR data preserves spatial locality. Such information is extracted using *local geometric descriptors*. These descriptors are defined as data that encodes the local geometric information, and are usually in the form of matrices or positive semidefinite second-order tensor fields (Sreevalsan-Nair and Kumari 2017). It is important for the matrices or second-order tensors to be positive semi-definite, as positive eigenvalues are required for further analysis. These local neighborhood are of varied shapes, depending on the application. Spherical, cylindrical, k-nearest, cubical neighborhoods are usually used.

Neighborhood search is a computationally intensive process, which can be made efficient using appropriate data structures. Octrees and kd-trees are used conventionally to partition the 3D bounding box containing the point cloud. The use of efficient hierarchical data structures reduces overall time in descriptor computation and subsequent feature extraction.

Semantic or object-based classification is a key data analytic process implemented on both ALS and TLS point

clouds. Supervised learning is largely used for ALS point clouds, where the point-wise feature vectors are hand-crafted. These features include two- and three-dimensional features, height-, and eigenvalue-based features. The feature extraction involves computation of features from local geometric descriptors using multiple spatial scales to gather more relevant information, of which an optimal scale is determined (Weinmann et al. 2015). The spatial scale considered here is the size of the local neighborhood where the optimal scale is where the entropy in the geometric classification of the point cloud is a global minimum. Depending on the relevance of the features, there are up-to 21 features, computed at the optimal scale included in the feature vector (Weinmann et al. 2015). Random forest tree classifier is best suited for the semantic classification of the LiDAR point cloud. Being a supervised learning, sufficiently large training data is required for training the model to improve classification accuracy. Detection and localization/instantiation of specific classes, such as, buildings, road/ground, etc., are well-researched topics (Sampath and Shan 2009; Chen et al. 2017).

The points can be geometrically classified as linear, planar, and volumetric features. They are alternatively called line-type, surface-type, and point/junction-type features. Combining semantic and geometric classification as a tuple of labels could be referred to as augmented semantic classification (Kumari and Sreevalsan-Nair 2015). The tuple of labels help in better graphical rendering of point clouds, where edges are delineated clearly. Larger sections of the point clouds tend to be surface-type features, as they include points on the exposed faces, interior to the object/instance boundaries. The normal to the surfaces can provide salient information. The normal information computed for multiple scales can be used to compute difference of normal (DoN) (Ioannou



LiDAR, Fig. 1 (a) Point rendering of Aerial Laser Scanned (ALS) point clouds of topographic data from Vaihingen site, Germany, using data provided by ISPRS as benchmark data. The red points are linear features, delineating structures, e.g., building roofs, and localizing objects, e.g., foliage, and gray are remaining points. (b) Corresponding orthoimage.

(Image source: (a) Author's own; (b) Dataset provider. Dataset Source: German Association of Photogrammetry, Remote Sensing and GeoInformation (DGPF) (Cramer 2010); <https://www2.isprs.org/commissions/comm2/wg4/benchmark/3d-semantic-labeling/>)

et al. 2012). DoN and other surface features have been effectively used for segmentation and classification.

Geometric reconstruction is another significant data analytic application implemented on point clouds. Building roof extraction is an application where points classified as building points, and localized as instances are used to reconstruct regions for graphical rendering, 3D printing, computer aided design (CAD), etc. Roof reconstruction involves finding contour lines or boundaries, and planar facets. The former can be done by using break-line extraction (Sampath and Shan 2009), Marching Triangles method on a triangulated irregular network (TIN) of points, and tensor-line extraction (Sreevalsan-Nair et al. 2018). The points can be clustered effectively to improve segmentation (Sampath and Shan 2009). Once the boundaries are extracted, one can fill in the planar facets using plane equations and geometric approximation methods of plane fitting.

Future Scope

Efficient and accurate DEM generation continues to be a problem of interest, specifically to improve the LiDAR point analysis. DTM generation can be improved by combining several related DTM generation methods (Chen et al. 2017), which is applicable to the larger set of DEMs. This involves combining different multi-scale aggregation algorithms, clustering, segmentation, classification, and geometric extraction methods. There are two approaches in combining methods. In the first approach, processes or components in the generation algorithms may be combined. In the second approach, the different outputs (DEMs) may be merged to give an improved DEM.

Another way of improving DEMs is the multimodal method, which involves inclusion of different data sources, in addition to LiDAR data. Thus, this method addresses the deficiencies in information capture and processing in LiDAR. Spectral features from aerial images, high resolution satellite imagery, high resolution photogrammatic point clouds are some of the data sources that have been successfully combined with LiDAR to extend its applications beyond DEM or DTM generation (Chen et al. 2017). Apart from the widely used discrete return LiDAR, the full-waveform LiDAR data has demonstrated strong potential for generating DTMs.

Some of the challenges faced in mobile LiDAR in data processing in the presence of high data uncertainty are handled by the use of machine learning and neural network models. However, this work requires in-depth research, as in ALS point clouds, to improve on the different classes of methodology for point cloud processing. For DTM generation using ALS, there are six different classes of methods, namely, surface-based corrections, morphological filters, TIN refinement, segmentation in conjunction with classification,

statistical analysis, and multi-scale aggregation methods. Use of the vehicle/mobile LiDAR for autonomous driving application is the latest entrant to be added to potential application of LiDAR point clouds. Geometric modeling is not directly extensible from ALS to mobile LiDAR owing to the incompleteness in scans of objects, especially of moving ones, in 360° views. Apart from bringing in variety in processing steps, efficiency is also a key requirement for edge, fog, and cloud computing that is posed to be used in autonomous driving, which is a potential use-case of interconnected network of sensors and *things*, referring to the internet of things (IoT).

In summary, LiDAR imagery in the form of 3D point clouds is an effective remote sensing method which is useful for varied applications in geoscience and GIS (geographical information systems), owing to its accuracy and sampling.

Bibliography

- Baltensweiler A, Walther L, Ginzler C, Sutter F, Purves RS, Hanewinkel M (2017) Terrestrial laser scanning improves digital elevation models and topsoil pH modelling in regions with complex topography and dense vegetation. *Environ Model Softw* 95:13–21
- Chen Z, Gao B, Devereux B (2017) State-of-the-art: DTM generation using airborne LiDAR data. *Sensors* 17(1):150
- Cramer M (2010) The DGPF-Test on Digital Airborne Camera Evaluation – Overview and Test Design. *Photogrammetrie – Fernerkundung – Geoinformation (PFG)* 2:73–82. https://www.dgpf.de/pfg/2010/pfg2010_2_Cramer.pdf
- Ioannou Y, Taati B, Harrap R, Greenspan M (2012) Difference of normals as a multi-scale operator in unorganized point clouds. In: 2012 second international conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), IEEE, pp 501–508
- Irish JL, Lillycrop WJ (1999) Scanning laser mapping of the coastal zone: the SHOALS system. *ISPRS J Photogramm Remote Sens* 54(2–3):123–129
- Kumari B, Sreevalsan-Nair J (2015) An interactive visual analytic tool for semantic classification of 3D urban LiDAR point cloud. In: Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems, ACM, p 73
- Lohani B, Ghosh S (2017) Airborne LiDAR technology: a review of data collection and processing systems. *Proc Natl Acad Sci India Sect A Phys Sci* 87(4):567–579
- Sampath A, Shan J (2009) Segmentation and reconstruction of polyhedral building roofs from aerial LiDAR point clouds. *IEEE Trans Geosci Remote Sens* 48(3):1554–1567
- Sreevalsan-Nair J, Kumari B (2017) Local geometric descriptors for multi-scale probabilistic point classification of airborne LiDAR point clouds. In: Modeling, analysis, and visualization of anisotropy. Springer, New York, pp 175–200
- Sreevalsan-Nair J, Jindal A, Kumari B (2018) Contour extraction in buildings in airborne LiDAR point clouds using multiscale local geometric descriptors and visual analytics. *IEEE J Sel Top Appl Earth Obs Remote Sens* 11(7):2320–2335
- Weinmann M, Jutzi B, Hinz S, Mallet C (2015) Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers. *ISPRS J Photogramm Remote Sens* 105:286–304

Linear Unmixing

Katherine L. Silversides

Australian Centre for Field Robotics, Rio Tinto Centre for Mining Automation, The University of Sydney, Sydney, NSW, Australia

Definition

Linear unmixing is a method of calculating the composition of a mixed sample (Weltje 1997). A geological example is when a rock sample is known to be a physical mix of certain minerals. These minerals are the endmembers. If the chemical composition of the endmembers and the sample is known, linear unmixing can be used to estimate the amount of each mineral in the sample. If the endmember minerals or their compositions are unknown, instead bilinear unmixing can be used to estimate both the endmember and sample compositions (Full 2018; Weltje 1997).

Introduction

A linear mixture occurs when endmembers are physically mixed or a sample has a composition that can be mathematically considered to be a combination of distinct endmembers with fixed compositions (Weltje 1997). An example of a linear mixture is when red and blue grains are mixed and the result is a mix of individual red and blue grains, not purple grains. When the composition of a sample and the endmembers is known, linear unmixing can be used to calculate the percentage of endmembers present in the sample. If the composition of the endmembers is unknown, the problem becomes bilinear (or explicit) unmixing. In this case, the composition of the endmembers is also estimated at the same time (Tolosana-Delgado et al. 2011; Weltje 1997).

Geological specimens or rocks can be considered to be comprised of linear mixes of endmembers, where the endmembers are a set of minerals with known chemical compositions (Weltje 1997). All of the samples are then comprised of a mixture of these endmembers. If the number of endmembers and their compositions is known, then the bulk composition of each rock can be calculated or estimated using linear unmixing (Weltje 1997). Linear mixtures are used in situations such as identifying minerals from bulk geochemistry (Renner et al. 2014; Silversides and Melkumyan 2017; Tolosana-Delgado et al. 2011; Weltje 1997) and hyperspectral data analysis (Bioucas-Dias et al. 2012; Chase and Holyer 1990; Dobigeon et al. 2009; Liu et al. 2018; van der Meer 1999).

Theory

Linear mixing can be considered as a matrix problem. The first matrix, X , contains D characteristics observed in N samples. Each sample is assumed to each be a mixture of P endmembers, where the characteristics of each composite sample are a weighted average of the endmember characteristics. The second matrix, T , is a $P \times D$ matrix that describes the characteristics of each endmember. These can be used to solve

$$X = Y \cdot T,$$

where $Y = (y_{ij})$ is a matrix containing the weights of the j th endmember in the i th sample. Solving for Y therefore provides the composition of each sample. In linear unmixing, the endmember characteristics, and therefore T , are known. In the bilinear problem, the endmember characteristics are unknown, and more complex solutions are required to solve for Y and T simultaneously (Tolosana-Delgado et al. 2011; Full 2018).

Mathematical Example

A system containing three endmembers, which are measured using four characteristics, can be described using a 3×4 matrix. For example,

$$T = \begin{bmatrix} 4 & 1 & 2 \\ 4 & 1 & 6 \\ 1 & 4 & 2 \\ 1 & 4 & 0 \end{bmatrix}$$

If there are two composite samples, they can be represented by a 4×2 matrix, for example,

$$Y = \begin{bmatrix} 1 & 1 & 4 & 4 \\ 2 & 5 & 3 & 2 \end{bmatrix}$$

These can then be used to find the weights using the equation above. Using these examples, the weights are

$$X = \begin{bmatrix} 22 & 28 & 24 \\ 33 & 27 & 40 \end{bmatrix}$$

This means that the first sample contains the endmembers in the ratio 22:28:24 and the second sample contains the endmembers in the ratio 33:27:40.

Applications

Linear unmixing can be applied to any sequence of data that contains mixtures of materials where each mixture produces observations that are a linear combination of the endmembers. Common applications include hyperspectral analysis and determining mineralogy from bulk chemistry. These are briefly discussed below.

Hyperspectral cameras or sensors can be used in geoscience applications to determine the proportion of materials such as minerals. These sensors measure electromagnetic energy across numerous wavelengths. In the geosciences, bands covering the visible, near-infrared, and shortwave infrared spectral bands are frequently used. The signal or spectrum recorded by a hyperspectral sensor is a result of the light reflected by the objects located in its field of view. Spectroscopic analysis can then be used to determine the composition of these objects. In mineral scanning, it is common for each pixel in a sensor to receive a mixed spectrum due to the mineral size being finer than the sensor resolution. If certain conditions are met, the mineral mixing is considered linear, and linear unmixing can be used to measure the abundance of each mineral present. These conditions can be summed up by each photon only interacting with a single mineral, so that the signal received is a sum of photons from individual minerals. These minerals are the endmembers, and their known spectral signatures can be used as the characteristics for linear unmixing of the hyperspectral data to find the fractional abundances of the minerals (Bioucas-Dias et al. 2012; Dobigeon et al. 2009; van der Meer 1999). Linear unmixing can also be applied to hyperspectral data in many other diverse applications, for example, mapping coral bathymetry (Liu et al. 2018) or identifying sea ice type and concentration (Chase and Holyer 1990).

The chemical composition of a sample alone cannot provide all of the information that is desired for mining and processing. For example, factors such as hardness, crystallinity, and element substitution can affect the processing and upgradability of ore. While there are methods that can directly measure the minerals, such as X-ray diffraction (XRD), they are expensive and time-consuming. Therefore, they are typically not routinely performed during a mining operation. However, when the minerals present in a deposit are known, along with their typical chemical composition, linear unmixing can be used to calculate the mineral composition of a particular sample (Silversides and Melkumyan 2017; Tolosana-Delgado et al. 2011). In this situation, the characteristics are the measured chemical species, the endmembers are the pure minerals, and the samples are the geochemical assays of the mixed composition rocks. Examples of using linear unmixing to determine mineralogy include mineral

composition in production holes in banded iron formation (BIF) hosted iron ore (Silversides and Melkumyan 2017), determining sediment mixtures in moraines in retreating glaciers (Tolosana-Delgado et al. 2011) and determining the mineral composition of deep-sea manganese nodules (Renner et al. 2014). Note that this application uses compositional data. Further information on compositional data and methods of processing is presented in other chapters.

Summary

Linear unmixing involves calculating the composition of a mixed sample using the characteristics of the sample and the characteristics of the endmembers. If the characteristics of the endmembers are unknown, then bilinear unmixing is used to find these characteristics simultaneously with the proportions of the endmembers present in the mixture. Linear unmixing has a wide range of geoscientific applications including mineral identification from hyperspectral sensing and determining mineralogy from bulk chemistry.

Cross-References

- [Compositional Data](#)
- [Hyperspectral Remote Sensing](#)

Bibliography

- Bioucas-Dias JM, Plaza A, Dobigeon N, Parente M, Du Q, Gader P, Chanussot J (2012) Hyperspectral unmixing overview: geometrical, statistical, and sparse regression-based approaches. *IEEE J Sel Top Appl Earth Obs Remote Sens* 5(2):354–379
- Chase JR, Holyer RJ (1990) Estimation of sea ice type and concentration by linear unmixing of Geosat altimeter waveforms. *J Geophys Res* 95(C10):18015–18025
- Dobigeon N, Moussaoui S, Coulon M, Tournet J, Hero AO (2009) Joint Bayesian endmember extraction and linear unmixing for hyperspectral imagery. *IEEE Trans Signal Process* 57(11):4355–4368
- Full WE (2018) Linear Unmixing in the geologic sciences: more than a half-century of progress. In: Daya SB, Cheng Q, Agterberg F (eds) *Handbook of mathematical geosciences*. Springer, Cham
- Liu Y, Deng R, Li J, Qin Y, Xiong L, Chen Q, Liu X (2018) Multispectral bathymetry via linear unmixing of the benthic reflectance. *IEEE J Sel Top Appl Earth Obs Remote Sens* 11(11):4349–4363
- Renner RM, Nath BN, Glasby GP (2014) An appraisal of an iterative construction of the endmembers controlling the composition of deep-sea manganese nodules from the Central Indian Ocean Basin. *J Earth Syst Sci* 123:1399–1411
- Silversides KL, Melkumyan A (2017) Mineralogy identification through linear unmixing of blast hole geochemistry. *Appl Earth Sci* 126(4):188–194

- Tolosana-Delgado R, von Eynatten H, Karius V (2011) Constructing modal mineralogy from geochemical composition: a geometric-Bayesian approach. *Comput Geosci* 37:677–691
- van der Meer F (1999) Image classification through spectral unmixing. In: Stein A, van der Meer F, Gorte B (eds) *Spatial statistics for remote sensing*. Kluwer, Dordrecht, pp 185–193
- Weltje GJ (1997) End-member modeling of compositional data: numerical-statistical algorithms for solving the explicit mixing problem. *Math Geol* 29(4):503–549

Linearity

Christien Thiart

Department of Statistical Sciences, University of Cape Town, Cape Town, South Africa

Definition

Linearity possesses the properties of additivity and homogeneity and furthermore, from a statistical point of view, it means linear in the parameters/coefficients or weights. For clarity consider the following simple relationship between Y and X $Y = \beta_0 + \beta_1 f(X)$, where β_0 and β_1 are known as the parameters/coefficients and $f(X)$ is any function of X that does not depend on any of the parameters. If $f(X) = X$ then this relationship would be a straight line crossing the Y axis at β_0 and the slope of the line is determined by β_1 , e.g., one-unit change in X will cause a change of β_1 in Y .

Introduction

Linearity is everywhere: conversion from one scale to another; an assumption of linear models, linear regression, logistic regression, and even in interpolation methods for spatial data. To expand our simple definition, consider the following scenarios:

Conversion Formula

Linearity can be as simple as a formula for conversion from one scale to another, e.g., to convert temperature from degrees Celsius (C) to degrees Fahrenheit (F):

$$F = \frac{9}{5}C + 32$$

This can easily be visualized by an XY-plot, with the Celsius temperature on the X axis and the Fahrenheit temperature on the Y axis. A straight line will show the relationship between the different measurements of temperature. It would be an exact line, with the intercept of the Y axis (F) at 32, the

slope of the line $\frac{9}{5}$ and the correlation coefficient between Fahrenheit and Celsius equal to one.

Linear Regression

An approach to find the best linear relationship between a response (dependent variable) and a set of independent/explanatory variables is known as linear regression. Consider the following linear regression model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y}$ is a $n \times 1$ response vector, $\boldsymbol{\beta}$ is an $r \times 1$ vector of unknown coefficients, $\mathbf{X}$ is a $n \times r$ matrix of nonrandom explanatory variables whose rank is r ($n > r$), and $\boldsymbol{\varepsilon}$ is $n \times 1$ vector of random error, with $\boldsymbol{\varepsilon} \sim (\boldsymbol{\tau}, \boldsymbol{\Sigma}_\varepsilon)$, $E(\boldsymbol{\varepsilon}) = \boldsymbol{\tau}$ and $\text{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}_\varepsilon$.

In the case of uncorrelated errors the expectation and variance of $\boldsymbol{\varepsilon}$ are $E(\boldsymbol{\varepsilon}) = \boldsymbol{\tau} = 0$ and $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$, where $\mathbf{I}$ is an identity matrix. Ordinary least square (OLS) produces best linear unbiased estimates (BLUE). Sometimes we also make the convenient assumption of normality, $\boldsymbol{\varepsilon} \sim N(\boldsymbol{\tau}, \sigma^2 \mathbf{I})$, allowing us to investigate various hypotheses for the $\boldsymbol{\beta}$'s and goodness-of-fit tests. Furthermore, under normality the conditional expectation of Y , given that $X = x$, is linear in x :

$$E(\mathbf{Y}|\mathbf{X}=x) = \mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(x - \mu_X)$$

where the parameters μ_Y and μ_X are the population means of Y and X , ρ is the correlation coefficient between X and Y , and σ_Y and σ_X the standard deviations. The conditional expectation is also called the regression curve of Y on x . A special case is the simple linear model where, for example, we only have one X variable:

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X}$$

This equation is linear both in the parameters (the β 's) and in the variable X . If we consider a polynomial function of the X variables, say of order s

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X} + \beta_2 \mathbf{X}^2 + \dots + \beta_s \mathbf{X}^s$$

then this equation is linear in the parameters (the β 's) but not linear with respect to the variable X . Even if at first glance an equation/line does not look linear/straight, we can transform it to obtain linearity. Consider

$$\mathbf{Y} = e^{\beta_0 + \beta_1 \mathbf{X}}$$

then by taking the natural log on both sides we obtain

$$\log \mathbf{Y} = \beta_0 + \beta_1 \mathbf{X}$$

which is now linear. When presented with a nonlinear function, it is useful to transform the variables in order to obtain linearity. Transformations to achieve linearity are outside the scope of this article, and help is easily available in any multivariate regression book (e.g., Mosteller and Tukey 1977).

Visualizing Linearity

In the case of bivariate variables it is easy to visualize the association between them with a scatterplot (XY-plot). One variable (usually the independent variable) defines the X axis and the other (dependent) variable defines the Y axis. The points on this graph represent the relationship between the two points. If the points lie on a straight line (as in our temperature example), we have perfect correlation between X and Y . In this case the correlation coefficient $r_{XY} = 1$, indicating a perfect positive relationship.

The correlation coefficient lies between -1 and $+1$. If the correlation coefficient is near 0 , no linear relationship between the two variables exists. Values of the correlation coefficient near ± 1 are an indication of strong linear relationships. But be aware the correlation coefficient is not a “test” or an indication of linearity. It is good practice to always inspect the residuals after fitting a model. The residuals are defined as the difference between the observed value and the predicted value ($\hat{\mathbf{Y}}$)

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

In the residual (some practitioners prefer the studentized residuals) by predicted scatterplot ($\hat{\mathbf{Y}}$), we see that the residuals are randomly scattered around zero, with no obvious nonrandom pattern, the so-called “null plot.” Nonlinearity, outliers, and influential observations will be easily identified for further investigation. Outliers are observations that are inconsistent with the rest of the observations, and influential points are single points far removed from the others. Both outliers and influential observation can affect the regression results (the regression coefficients $[\beta's]$ and the correlation coefficient).

Linear Dependence

Linearity can also be linked to the collinearity problem. Collinearity happens easily in observational studies, and the

$\mathbf{X}$ matrix representing the independent variables in the linear model can be rank-deficient. The rank of a matrix is defined as the number of linearly independent columns/rows in the matrix. No linear dependencies among the columns indicate that the matrix is of full rank or nonsingular. If the matrix is singular (some linear dependencies) we have collinearity. Collinearity leads to instability of the regression coefficients, sometimes resulting in nonsense coefficients and coefficients with very large standard deviations. The whole process is unstable and adding small random errors in $\mathbf{Y}$ causes large shifts in the coefficients. The regression results are unstable (Rawlings et al. 1998).

Linearity in Interpolation Methods for Spatial Data

Suppose that observations are made on an attribute Z at n spatial locations $\mathbf{s} = (\mathbf{s}_1, \dots, \mathbf{s}_n)$ in a designated geographic region D . Let $Z(\mathbf{s}_g)$ denote the observed response at a generic location $\mathbf{s}_g$, where $\mathbf{s}_g \in D \subset \mathbb{R}^d$, and let $\mathbf{Z}(\mathbf{s})$ denote the $n \times 1$ vector of responses at the n locations. The prediction $Z(\mathbf{s}_0)$ at location $\mathbf{s}_0$ in D is usually a (linear) weighted sum of the observations observed around it:

$$\hat{Z}(\mathbf{s}_0) = \sum_{i=1}^{N(\mathbf{s}_0)} \lambda_i \cdot Z(\mathbf{s}_i) = \boldsymbol{\lambda}^T \mathbf{Z}(\mathbf{s}),$$

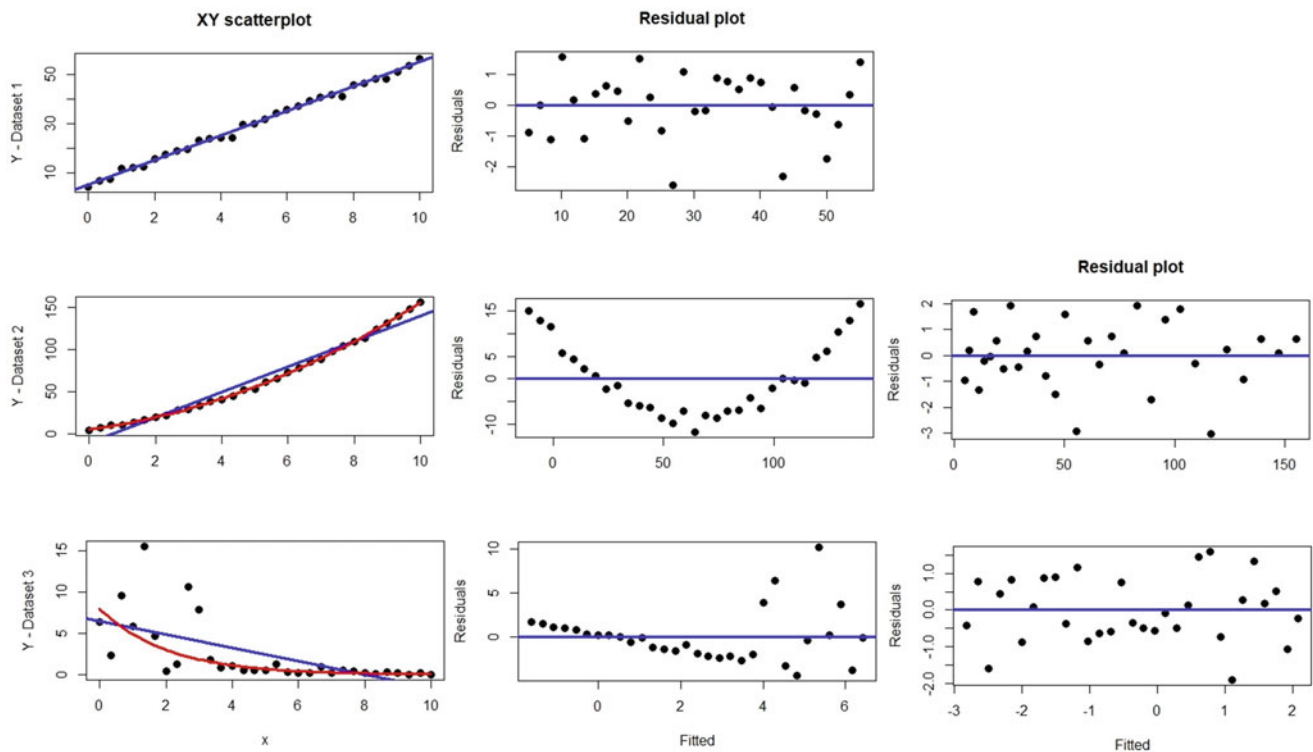
where λ_i is the weight associated with the i th observation $Z(\mathbf{s}_i)$ with $\sum_{i=1}^{N(\mathbf{s}_0)} \lambda_i = 1$, $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_{N(\mathbf{s}_0)}]$, and $N(\mathbf{s}_0)$ is the number of points in the search neighborhood around the prediction location $\mathbf{s}_0$.

For inverse distance interpolation the weights will be a function of the distance between points and for kriging the weights are obtained by minimizing the mean square prediction error.

Example

To illustrate the concept of linearity and the importance of visualization, we consider three small datasets. Dataset 1 is generated using a simple linear regression model: $Y = 5 + 5X + \varepsilon$, where $\varepsilon \sim N(0, 1.44)$ is a random error term, X is a sequence of numbers evenly spaced from 0 to 10 , and $n = 31$. Dataset 2 is generated by adding a quadratic term: $Y = 5 + 5X + X^2 + \varepsilon$ and dataset 3 is generated using an exponential model: $Y = 5 * \exp[-\frac{1}{2}X + \varepsilon]$.

Dataset 1: We fit a simple linear regression model to give $Y = 5.18 + 4.99X$, $R^2 = 0.99$ (blue line, Fig. 1, top left). For the fitted values versus residuals plot we observed a roughly



Linearity, Fig. 1 Analysis of simulated datasets. Row one is dataset 1, generated from a simple linear regression model; row two is dataset 2, generated from a quadratic model; and row three is dataset 3, generated from an exponential model. First column: scatter plot of observed points; blue line is the fitted simple linear regression model, red line is a fitted

model improving on the simple regression model (if applicable). The middle column presents fitted values versus residuals for the regression models and the last column the fitted values versus residuals for the improved model (if applicable)

centered random pattern around 0 with one possible outlier (Fig. 1, top middle). The residual plot and the $R^2 = 0.99$ are an indication of a good fit.

Dataset 2: We first fit a simple linear regression model to give $Y = -10.9 + 15X$, $R^2 = 0.97$ (blue line, Fig. 1, row 2 left). The corresponding fitted values versus residual plot for this model (Fig. 1, row 2 middle) shows a systematic pattern; the residuals are positive for small X values and large X values and negative for X values in between. When we add a quadratic term, the resulting fitted model $Y = 5 + 5.1X + X^2$, $R^2 = 0.99$ (red line, Fig. 1, row 2, left) produces a residual versus fitted plot that shows a centered random pattern around zero (Fig. 1, row 2 right). Note that the R^2 value of 97% for the simple regression model is very high and misleading, thereby illustrating the concept that one cannot just interpret the output of a model without inspecting the residuals.

Dataset 3: We first fit a simple linear regression model to give $Y = 6.4 - 0.8X$, $R^2 = 0.41$ (blue line, Fig. 1, bottom left). But if one looks at fitted values versus residuals (Fig. 1, bottom middle), we observe a nonrandom pattern. If we fit a model by taking the log of the Y 's (red line, Fig. 1, bottom left), then $Y = \exp(2.1 - 0.49X)$, $R^2 = 0.74$, and the resulting plot of fitted values versus residuals (Fig. 1, bottom right) is now a random pattern around zero.

Summary

Linearity is everywhere, either exactly or approximately; it can be a simple conversion formula but can also be an important assumption of various models, e.g., linear models, linear and logistic regressions. In the case of spatial data, predictions at unknown locations are linear weighted sums of the locations observed around the unknown locations. Visualization is a useful tool to judge the linearity between bivariate variables; if we have linear dependence between exploratory variables it can lead to collinearity, resulting in instability of the fitted model. Even when a function is nonlinear, for many purposes it can be useful as a start to approximate it by a linear function.

Cross-References

- [Interpolation](#)
- [Inverse Distance Weight](#)
- [Kriging](#)
- [Nonlinearity](#)
- [Ordinary Least Squares](#)
- [Regression](#)

Bibliography

- Mosteller F, Tukey JW (1977) Data analysis and regression: a second course in statistics. Addison-Wesley, Reading
- Rawlings JO, Pantula SG, Dickey DA (1998) Applied regression analysis, a research tool, 2nd edn. Springer, Cham

Local Singularity Analysis

Wenlei Wang¹, Shuyun Xie² and Zhijun Chen²

¹Institute of Geomechanics, Chinese Academy of Geological Sciences, Beijing, China

²China University of Geosciences (Wuhan), Wuhan, China

Definition

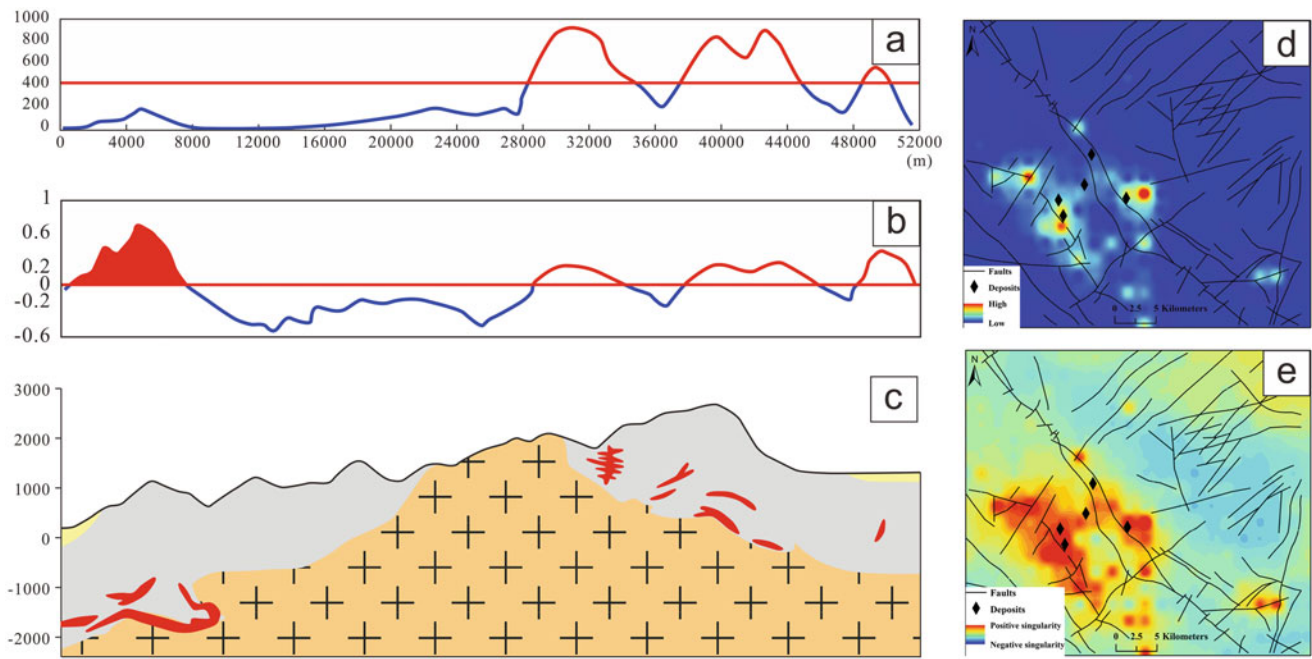
Local singularity analysis as a fractal-based concept was initially proposed to investigate anomalous distributions of ore elements (Cheng 1999). Formally defined in the *Dictionary of Mathematical Geosciences* (Howarth 2017), the concept of singularity analysis is “In geochemical applications, the singularity also can be directly estimated from the slope of a straight line fitted to a log transformed element concentration value as a function of a log transformed size measure. Areas with low singularities can be used to delineate anomalies characterized by relatively strong enrichment of the elements considered. Local singularity maps usually differ significantly from maps with geochemical anomalies delineated by other statistical methods.” In geologic applications, local singularity analysis is an effective method of quantitatively and qualitatively characterizing irregular energy release and/or material accumulation within relatively short spatial-temporal intervals generated by nonlinear geoprocesses (Cheng 2007). In a local comparison of element concentrations (or other physical quantities) between samples and their vicinities, spatially varied physicochemical signatures can be quantified using a singularity index (i.e., a measure of the intensity of variations or changes). As an example of the investigation of local signatures (e.g., ore element concentrations) across space, spatial variations of geoanomalies associated with mineralization can be evaluated. After more than 10 years of development, local singularity analysis has been extended to characterize many nonlinear geoprocesses and extreme geoevents as part of geological, geophysical, and litho-geochemical data analyses.

Introduction

In the context of earth science, geoanomalies having geologic signatures distinctive from those of their surroundings are a

major focus in the study of the nature of hazards, the environment, and energy and mineral resources for the survival and development of human society. Geoanomaly analysis has become a fundamental technical route in many subdisciplines of earth science. In past centuries, both linear and nonlinear methodologies have played important roles in the locating and characterizing of geoanomalies. Statistical methods including univariate and multivariate analyses (e.g., the use of a Q-Q plot or bi-plot cluster) are broadly adopted for their effectiveness in solving problems in the statistical frequency domain rather than the spatial domain. With the development of applications in geology, a branch of statistics termed geostatistics has been proposed to predict ore grades in the mining industry and is maturely applied in various branches of geology and geography. Most observational datasets in spatial and spatial-temporal forms can be effectively interpreted in the spatial domain. In the early stage of its application in geological exploration, geostatistics was based on the hypothesis that mineralization-related geovariates (i.e., geochemical data) follow a normal or log-normal distribution (Ahrens 1953). The simplest and classical method of identifying a geoanomaly from the background is the use of an algorithm based on the mean (μ) $\pm$ standard deviation(s) (e.g., σ , 2σ , or 3σ) (Fig. 1). Many mineral deposits have been discovered from geoanomalies identified by geostatistical methods, especially geochemical anomalies. Besides the adoption of the above type of algorithm, other approaches including the use of multielement indices, principal component analysis, independent component analysis, multidimensional scaling, and random forests have been adopted successfully in Australia, Canada, and many other countries (Grunsky and Caritat 2020). However, with the advance of mineral exploration around the world, most easily discoverable mineral resources buried at a shallow depth have been exploited. Exploration fields now tend to be in covered areas, the deep earth, and other untraditional spaces. Geoanomalies now tend to be weak and difficult to identify from observational datasets adopting traditional geostatistical methods (Fig. 1).

Fractal methods that consider both the frequency distribution and spatial self-similarity have become popular in recent decades. Against this background, local singularity analysis, a typical fractal method, has been proposed to identify weak anomalies that have not been discovered owing to the strong variance of the background and/or deep burial (Fig. 1). Local singularity analysis has been shown to be efficient in enhancing weak geoanomalies and effective in discovering causative bodies (e.g., ore bodies) of geoanomalies in covered areas and deeply buried spaces. As introduced by Cheng (2007), local variations of ore elements have been investigated by local singularity analysis and evaluated using a singularity index. Weak geochemical anomalies produced by deeply buried ore bodies have been appropriately enhanced. Meanwhile, strong anomalies identifiable using traditional methods are



Local Singularity Analysis, Fig. 1 Schematic diagrams of the traditional method (a) and local singularity analysis (b) for the identification of a geoanomaly in a mineral district (c). In comparison with the

identification of a geoanomaly using the traditional method (d), local singularity analysis (e) is more effective in characterizing a weak geoanomaly

preserved without the loss of indicative patterns (Fig. 1). Local singularity analysis has since flourished in mineral exploration and other subdisciplines of earth science. Its application has been extended to investigate geological and geophysical data. The concept of the fault singularity has been proposed and used to characterize the development of fault systems indicative of mineralization-favoring spaces produced by faulting activity (Wang et al. 2012). In application to gravity and magnetic data, geophysical anomalies associated with mineralization have been enhanced, demonstrating that local singularity analysis allows improved and simplified high-pass filtering with the advantage of scale independence (Wang et al. 2013a). Recently, local singularity analysis has been applied to investigate earthquakes, magmatic flare-ups, continental crust evolution, and other extreme geological events (Cheng 2016). The progressive development of the theory, application and algorithms of local singularity analysis has had profound implications for the simulation, prediction, and interpretation of nonlinear geoprocesses.

Singularity Index

As local singularity analysis was originally defined in the context of hydrothermal mineralization (Cheng 1999), the mass $\mu(V)$ and density $C(V)$ of ore materials each have a power-law relationship with V :

$$\mu(V) = c(V)^{\alpha/E}, \quad (1)$$

$$C(V) = c(V)^{\alpha/E-1}, \quad (2)$$

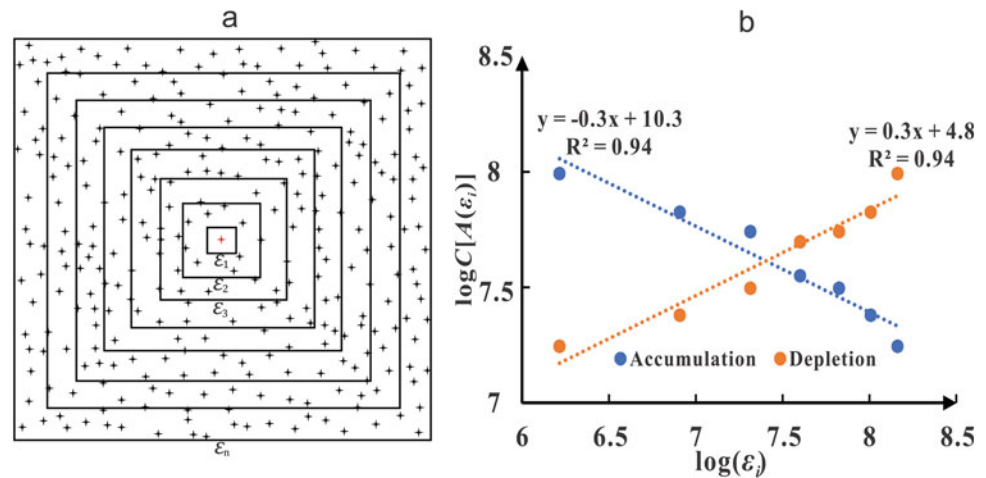
where the constant c determines the magnitude of functions; E is the Euclidian dimension; and α is termed the singularity index. The scaling exponent α is a kernel parameter of local singularity analysis that preserves the shape of functions and changing behaviors of the ore materials at different scales of spatial-temporal intervals. Different physicochemical behaviors can be characterized on the basis of changes in the singularity index α . For $\alpha = E$, the ore materials follow a monofractal distribution that indicates that the mass $\mu(V)$ and density $C(V)$ are independent of the volume V . For $\alpha > E$, there is a negative singularity that corresponds to the depletion of ore materials. This implies that ore materials follow a multifractal distribution. For $\alpha < E$, there is a positive singularity that indicates that ore materials follow a multifractal distribution, implying enrichment as introduced in the definition of local singularity analysis.

Methods of Estimating the Singularity Index

Methods of estimating the singularity index have progressively advanced over two decades. The first proposed estimation method is the square-window-based method (Cheng 2007). In its application to geochemical data in two dimensions, the general estimation steps (Fig. 2) of the method are: (1) centering on a location i of the study area and predefining a

Local Singularity Analysis,

Fig. 2 Schematic diagram of the square-window-based method of estimating the singularity index (after Wang et al. 2018)



set of square windows with size $A(\varepsilon_i \times \varepsilon_i)$ and density $C[A(\varepsilon_i)]$; (2) plotting $C[A(\varepsilon_i)]$ and $A(\varepsilon_i \times \varepsilon_i)$ on a log-log plot (Eq. 3); (3) adopting the least-squares method to estimate the slope (k) of the linear relation between $\log C[A(\varepsilon_i)]$ and $\log \varepsilon_i$; (4) calculating the singularity index $\alpha_i = k + E$ or $k + 2$; and (5) iterating these steps at each location to estimate the spatial distribution of the singularity index of the study area. Equation 3 is written as

$$\log C[A(\varepsilon_i)] = c + (\alpha - E) \log(\varepsilon_i). \quad (3)$$

Selected Case Studies

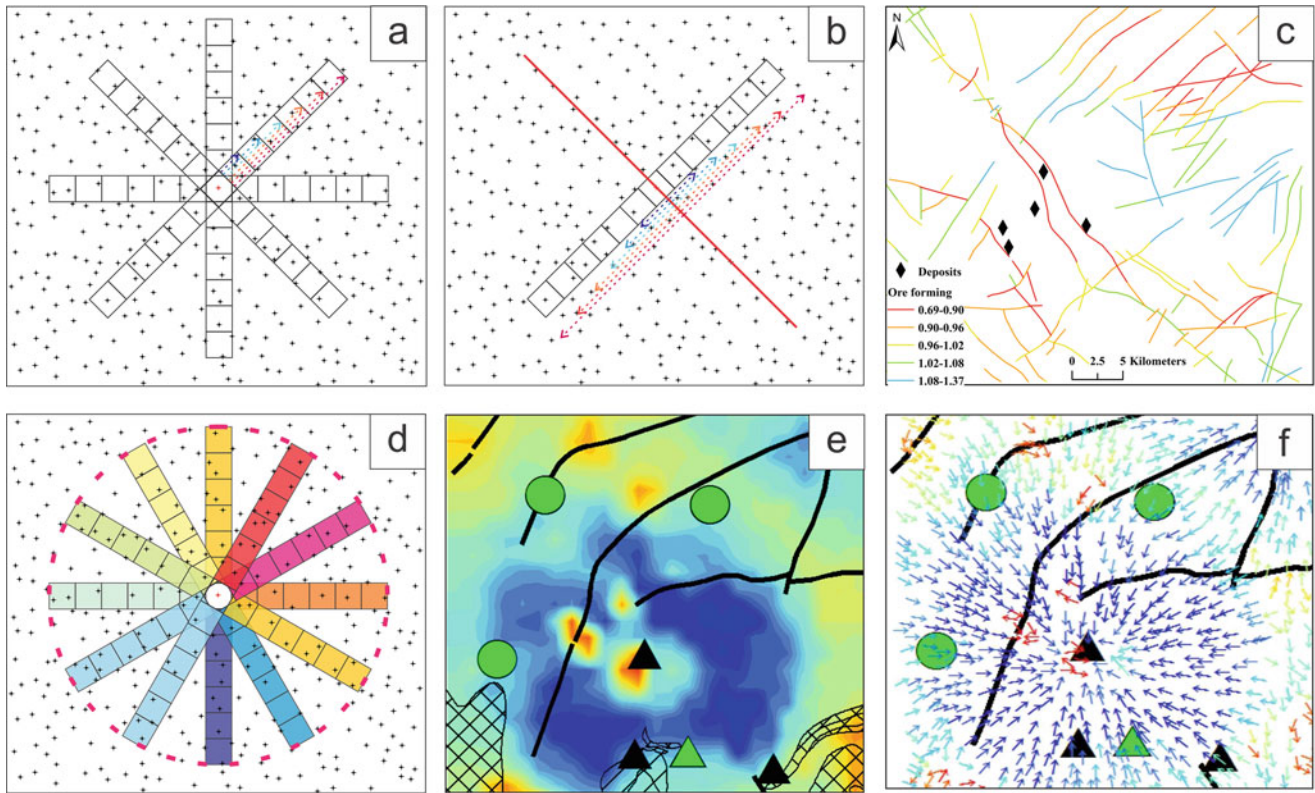
Furthermore, algorithms of local singularity analysis have been developed to characterize geoanomalies more precisely and appropriately. As an example, geochemical signatures inherited from multiple geoprocesses are often those of heterogeneity and anisotropy. In addition to identifying the locations of geoanomalies indicative of mineralization, the characterization of the anisotropic distributions of the related mineralization is necessary. The directional window-based method was first proposed by Li et al. (2005). In this method, a series of directional (rectangular) windows are predefined to obtain $C[A(\varepsilon_i)]$ and ε_i (Fig. 3a). Other steps of estimating the singularity index α are similar to those in the square-window-based method. The variation in window size only follows changes in ε_i , and the estimation thus becomes a one-dimensional problem and $\alpha = k - 1$. Adopting this method, the spatial variation of geochemical behaviors along the initial (i.e., northward) direction is consequently characterized. Moreover, other directional singularity indices, such as the eastward-trending window-based index (Wang et al. 2018), can be estimated.

Wang et al. (2013b) proposed a geologically constrained local singularity analysis that is termed the fault-trace-

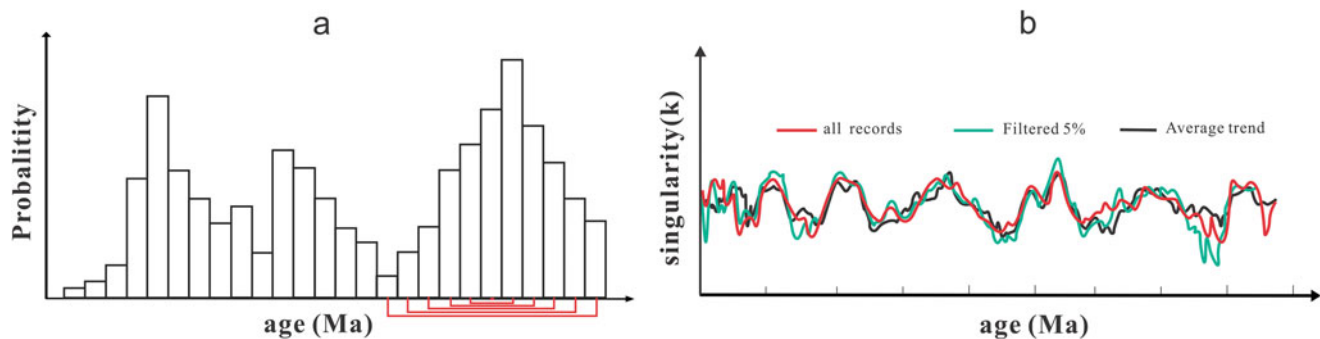
oriented singularity mapping technique and quantitatively describes interrelations between fault structures and ore-bearing fluids. In contrast with former window-based methods, fault traces are first divided equally into segments (Fig. 3b). Centering on one segment, a series of rectangular windows along the direction vertical to the segment is defined. Then, in steps similar to those discussed above, the fault-oriented singularity index is estimated to describe interactions between fault activity and fluid flow. A new fault property (Fig. 3c) is obtained by assigning the index to the fault segment: $\alpha > 1$ indicates a negative fault, or a gradual depletion of metals approaching the fault space, whereas $\alpha < 1$ indicates a positive fault, or a continuous enrichment of metals approaching the fault space.

Wang et al. (2018) further developed the directional window method and proposed the concept of the anisotropic singularity. Following the directional window method, the anisotropic singularity estimation focuses more on the intensity of anisotropic variations between samples and their vicinities. Centering on one sample location, series of rectangular windows trending along all directions are predefined (Fig. 3d). Using the estimation method discussed above, the direction with the maximum slope, or $|\alpha - 1|_{\max}$, is determined and used to obtain the singularity index for the current location, such that the variation (accumulation or depletion) characterized by $|\alpha - 1|_{\max}$ is greatest along the selected direction (Fig. 3e). In addition, as shown in Fig. 3f, directional patterns of element accumulation (with arrows directed from low to high) and depletion (with arrows directed from high to low) depict the anisotropy of geochemical behaviors appropriately.

Cheng (2017) applied the local singularity analysis to global databases of igneous and detrital zircon U–Pb ages. Figure 4 shows the U–Pb ages on a histogram. The episodic growth of the continental crust and the development of



Local Singularity Analysis, Fig. 3 Schematic diagrams of directional window-based (a), fault-trace-oriented (b, c), and anisotropic singularity index estimation methods (d, e, f) (after Wang et al. 2018)



Local Singularity Analysis, Fig. 4 Schematic diagram of the method of estimating the singularity index for histogram data in one dimension (a) and practical application in analyzing the global zircon U–Pb age series (Cheng 2017)

supercontinents are characterized by the age peaks. In analyzing the one-dimensional data, the estimation method (Fig. 4) was modified to have the steps of (1) choosing one bin of the histogram as the initial one-dimensional window; (2) while centering on this bin, defining a series of one-dimensional windows with size ε_i ; (3) plotting average numbers of recorded ages of bins within each window $C[A(\varepsilon_i)]$ and ε_i on a log-log graph; and (4) estimating the slope k using the least-squares method and using k to calculate the

singularity index for the initial bin, $\alpha = k + 1$. The case study well extends the application fields of local singularity analysis to lithogeochemical data, focusing on the shapes of age peaks depicted by the singularity index regardless of the bin size of the histogram (i.e., scale independency) rather than the amplitudes and periodicity of age series. The singularity indices of global data show that all age peaks have an episodic pattern with a duration of 600–800 Myr (Fig. 4) (Cheng 2017).

Conclusions

After more than 20 years of development, local singularity analysis has become a broadly used geoanomaly identification method, especially in the exploration of weak geoanomalies obscured by covering materials, owing to its advantages in quantifying spatial-temporal variations of physicochemical signatures generated by nonlinear geoprocesses. Initially applied in geochemical data analysis, spatially varied ore element concentrations have been characterized using the singularity index to represent accumulation and depletion in the comparison of samples within local vicinities. In the context of two-dimensional data analysis, geological and geophysical data have been investigated to delineate mineralization and other geoanomalies related to nonlinear geoprocesses, quantitatively and qualitatively. Algorithms have been further developed to satisfy geological constraints and the anisotropic nature of geoprocesses. In recent practice, local singularity analysis was adopted to simulate and predict extreme geological events that extend the application fields in earth sciences. In more widespread and profound application, local singularity analysis is expected to be adopted for new situations of geological singular events.

Cross-References

- [Fractal Geometry in Geosciences](#)
- [Nonlinearity](#)

Bibliography

- Ahrens LH (1953) A fundamental law of geochemistry. *Nature* 172: 1148–1148
- Cheng Q (1999) Multifractality and spatial statistics. *Comput Geosci* 25: 949–961
- Cheng Q (2007) Mapping singularities with stream sediment geochemical data for prediction of undiscovered mineral deposits in Gejiu, Yunnan Province, China. *Ore Geol Rev* 32:314–324
- Cheng Q (2016) Fractal density and singularity analysis of heat flow over ocean ridges. *Sci Rep* 6:1–10
- Cheng Q (2017) Singularity analysis of global zircon U-Pb age series and implication of continental crust evolution. *Gondwana Res* 51: 51–63
- Grunsky E, Caritat P (2020) State-of-the-art analysis of geochemical data for mineral exploration. *Geochemistry* 20(2):217–232
- Howarth RJ (2017) *Dictionary of mathematical geosciences*. Springer, Cham
- Li Q, Liu S, Liang G (2005) Anisotropic singularity and application for mineral potential mapping in GIS environments. *Prog Geophys* 20: 1015–1020. (in Chinese with English abstract)
- Wang W, Zhao J, Cheng Q, Liu J (2012) Tectonic-geochemical exploration modeling for characterizing geo-anomalies in southeastern Yunnan district, China. *J Geochem Explor* 122:71–80
- Wang W, Zhao J, Cheng Q (2013a) Application of singularity index mapping technique to gravity/magnetic data analysis in southeastern Yunnan mineral district, China. *J Appl Geophys* 92:39–49
- Wang W, Zhao J, Cheng Q (2013b) Fault trace-oriented singularity mapping technique to characterize anisotropic geochemical signatures in the Gejiu mineral district, China. *J Geochem Explor* 134: 27–37
- Wang W, Cheng Q, Zhang S, Zhao J (2018) Anisotropic singularity: a novel way to characterize controlling effects of geological processes on mineralization. *J Geochem Explor* 189:32–41

Locally Weighted Scatterplot Smoother

Klaus Nordhausen and Sara Taskinen

Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland

Definition

The locally weighted scatterplot smoother (LOWESS) uses local regression models to obtain a smooth estimator for the conditional expected value of response variable. The locality is defined based on a nearest neighbour-type approach assigning weights to the observations based on their distances from the point of interest as well as on their outlyingness which are then used in an iterative weighted least squares approach.

Introduction

A common problem in statistics is to identify the relationship between two variables, *i.e.*, between an explaining variable x and a response variable y based on a sample $(x_i, y_i), i, \dots, n$. Traditionally, the conditional expectation of y given x is modelled as a function of x , and we assume that the data are generated by

$$y_i = g(x_i) + \epsilon_i,$$

where $g(\cdot)$ is a smooth function and ϵ_i is some random fluctuation with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$. When visualizing a scatterplot of x_i against y_i , one thus wants to see a smooth relationship in the conditional mean of y as a function of x .

Assuming $g(x) = \beta_0 + \beta_1 x$, where β_0 and β_1 are constants, a straight line is used to model the conditional mean of y , which is known as linear regression. However, in many practical problems, it seems an oversimplification to approximate $g(\cdot)$ with a linear function. The use of a polynomial of order d , *i.e.*,

$g(x) = \beta_0 + \beta_1 x + \dots + \beta_d x^d$, may also not be reasonable and may overfit the data and cause problems at the boundaries of x .

In many cases, it may be reasonable to assume that locally, in a neighbourhood of a point x , a linear regression provides a good approximation for the conditional mean of y . For such a local regression, observations close to x are naturally more relevant than observations further away, thus leading to the following weighted regression problem

$$\operatorname{argmin}_{\beta_0, \beta_1} \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right) (y_i - \beta_0 - \beta_1 x_i)^2, \quad (1)$$

where K is a non-negative weight function called kernel, and the parameter h which defines locality is called the bandwidth. The bandwidth can be chosen, for example, using cross validation. This modelling approach is also known as local linear regression and is easily extended to the case of several predictors and to local polynomial regression.

Local linear regression has a long tradition and is discussed in detail in Loader (1999), for example. The main advantages of local linear regression are, according to Cleveland & Loader (1996), that (i) it is easy to understand and interpret, (ii) it adapts well to problems at the boundaries and regions of high curvature, and (iii) it does not require smoothness or regularity conditions. Many computational tools for local linear regression exist. In R, the method is available, for example, in the package *locfit* (Loader 2020).

Despite its flexibility, local linear regression as defined in (1) has some shortcomings. One disadvantage is that the method uses the quadratic loss function, which is well known to be sensitive to outliers. It is therefore natural to replace the quadratic loss function with a robust loss function and fit locally some robust regression model instead of a least squares regression. Local robust polynomial regression of order d can then be defined as

$$\operatorname{argmin}_{\beta_0, \beta_1, \dots, \beta_d} \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right) \rho(y_i - \beta_0 - \beta_1 x_i - \dots - \beta_d x_i^d), \quad (2)$$

where ρ is a loss function. The data analyst must now choose the kernel function K , the loss function ρ , the order of the polynomial d , and the bandwidth h .

Locally Weighted Scatterplot Smoother

In this context, Cleveland (1979) suggested the locally weighted scatterplot smoother (LOWESS) which uses as a kernel the tricube weight function

$$K(x) = \left(1 - |d|^3\right)^3,$$

when $|d| < 1$, and 0 otherwise, where d is the distance from the neighbourhood sample points to the value x . Cleveland (1979) argues that such kernel is beneficial as it assigns points far away a zero weight and is therefore better than kernels based on symmetric densities which give usually all points some positive weight. Depending on the scale of x , it may then however be advisable to scale the observations. For the neighbourhood structure, LOWESS suggests an adaptive nearest neighbour method. Let $0 < f \leq 1$ be a tuning parameter and denote then $r = \lceil fn \rceil$ as the number of local data points to be used. At point x , those r points, for which $|x - x_i|$ is the smallest, then form the neighbourhood. As the next step, initial fits $\hat{y}_i$ at each data point are obtained using the local polynomial least square regression in order to obtain the residuals $r_i = y_i - \hat{y}_i$. These residuals are then robustly scaled using six times the median absolute deviation (mad), i.e., the median of $\{|r_1|, \dots, |r_n|\}$. To downweight large residuals, biweight kernel $B(t) = (1 - t^2)^2 I_{[-1,1]}(t)$ is then used to assign robust weights $\delta_i = B(r_i / (6 \text{ mad}))$. Based on this initial step, the procedure is then iterated M times so that the kernel $K(x)$ is multiplied with robust weights δ_i . Thus, this corresponds to local iterative weighted least squares (IWLS) where the weights depend on the distance from the point of interest and whether the point is considered as outlying.

The parameter values for LOWESS that need to be chosen are f , d , and M . Cleveland (1979) suggested to choose d from 0, 1, and 2 with a tendency to $d = 1$ as higher order polynomials tend to overfit the data and the method becomes numerically unstable. It is also argued that the number of iterations can be small, say $M = 2$ or $M = 3$, and no convergence criterion is needed. Parameter f has most impact on the results, as, when the value is increased, the smoothness of the fit increases. Thus, it is a trade-off between variability and recognizing the pattern. Cleveland (1979) suggests to start with $f = 0.5$ and limit the value to the range of $[0.2, 0.8]$. For automatic procedures, Cleveland (1979) suggests to first choose the initial f which minimizes $\sum_{i=1}^n (y_i - \hat{y}_{i,f})^2$ and then to choose the final f as the minimizer of $\sum_{i=1}^n \delta_i (y_i - \hat{y}_{i,f})^2$. Based on the final robust weights, one can then compute the fitted values $\hat{y}_1^*, \dots, \hat{y}_m^*$ for a grid of, not necessarily observed values, $x_1^*, \dots, x_m^*$ covering the range of the observed x 's. Using linear interpolation, these points can then be used to visualize the smooth function. For the theoretical properties of LOWESS under the assumption of iid errors following a Gaussian distribution, we refer to Cleveland (1979). The procedure in an algorithmic form is summarized in Algorithm 1.

Algorithm 1: LOWESS algorithm

Input: n data points (y_i, x_i) , $i = 1, \dots, n$;
 1 Set $f \in (0, 1]$ to specify the neighborhood;
 2 Set integer $d \geq 0$ as the degree of the local polynomial;

- 3 Set integer $M \geq 1$ for the number of iterations;
- 4 Compute the size of the nearest neighbourhood $r = \lceil fn \rceil$;
- 5 **for** $i \leftarrow 1$ **to** n **do**
- 6 For x_i , identify the r points forming the neighbourhood and denote them $x_{r_l}^i$, where $r_l \in \{1, \dots, i-1, i+1, \dots, n\}$ and $l = 1, \dots, r$;
- 7 Use these r points to fit a local polynomial regression model of order d using the kernel $K(x_{r_l}^i) = \left(1 - |x_{r_l}^i - x_i|\right)^3$;
- 8 Obtain the initial fit $\hat{y}_i$;
- 9 **for** $iter \leftarrow 1$ **to** M **do**
- 10 Calculate the residuals $r_i = y_i - \hat{y}_i$ and the median of absolute deviations (mad) of the residuals;
- 11 Compute the robust weights $\delta_i = B(r_i/(6 \text{ mad}))$;
- 12 **for** $j \leftarrow 1$ **to** n **do**
- 13 Use the neighbourhood points of x_j , $x_{r_l}^j$ and their corresponding robust weights $\delta_{r_l}^j$ to fit a local polynomial regression model of order d using the kernel $K(x_{r_l}^j) = \delta_{r_l}^j \left(1 - |x_{r_l}^j - x_j|\right)^3$;
- 14 Obtain the robust fit $\hat{y}_j$;
- 15 Predict y for any point of interest x ;

In R, LOWESS with $d = 1$ can be applied using, for example, the function `lowess`.

Extensions of LOWESS

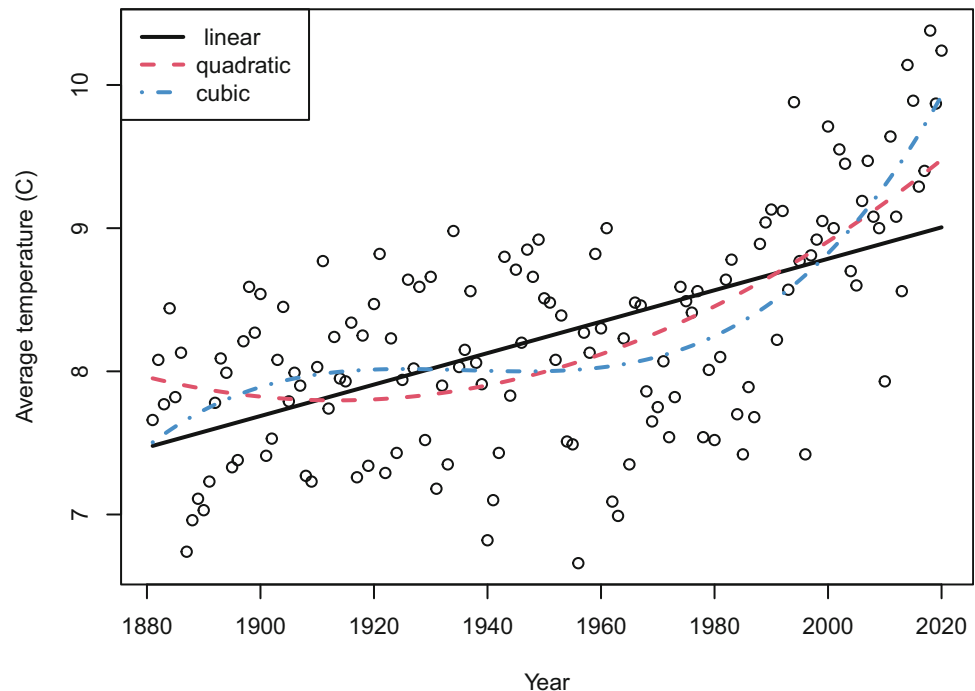
Cleveland (1979) pointed out that LOWESS can be used with any kernel function, and also variants exist to speed up the computations for large data sets and for accommodating prior weights for the observations.

The main extension of LOWESS is however suggested in Cleveland & Devlin (1988) as locally estimated scatterplot smoothing (LOESS). LOESS generalizes LOWESS to the case of several predictors by replacing the simple regression with a multiple regression. Implementations such as LOESS in R let the user choose the degree of the polynomial d as well as the need for robust updates.

Example

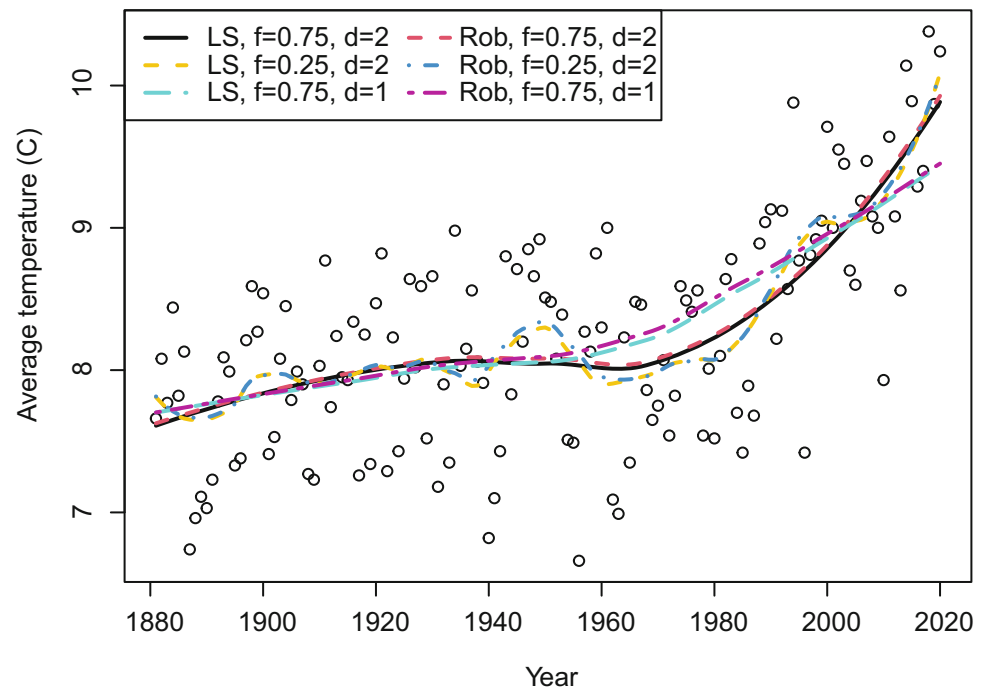
To demonstrate LOWESS, we consider average yearly temperatures (in centigrades) of the German state Baden-Wuerttemberg from 1880 to 2020. The data can be obtained from the CDC-portal of the Deutscher Wetterdienst. Figure 1 shows the data together with three global polynomial least squares fits. All fits indicate an increase in average temperature but seem not to fit the data very well for early temperature records. We then apply LOWESS via the R function `loess` and visualize the results in Fig. 2 for different degrees of the polynomial (d) and different parameters defining the

Locally Weighted Scatterplot Smoother, Fig. 1 Yearly average temperatures ($^{\circ}\text{C}$) for Baden-Wuerttemberg in Germany together with three global fits



Locally Weighted Scatterplot

Smoother, Fig. 2 Yearly average temperatures ($^{\circ}\text{C}$) for Baden-Wuerttemberg in Germany together with locally smoothed fits



neighbourhood (f). Both, the robust variant with the IWLS steps as well as the pure local LS fits, are illustrated. The figure clearly shows that the differences between robust fits and LS fits are quite small indicating the lack of outlying observations. A quadratic polynomial seems however preferable, especially with a large value of f , and the LOWESS fit shows that after a long period of slow rise in average temperature, the rise has accelerated in the last 50 years. A more thorough analysis of temperature data using LOWESS is available in Wanishsakpong & Notodiputro (2018), for example.

Summary

LOWESS and its extension LOESS are powerful and simple regression methods ideal for situations where an initial model assumption is not available, as the methods do not use any special model specification. LOWESS exploits locality which means that it requires dense observations to fully achieve its strengths. While LOWESS was originally designed for models with signal predictors, LOESS allows more predictors meaning that the method creates smooth surfaces which then need even denser data and a careful choice for the notion of distance.

Cross-References

- [Smoothing Filter](#)
- [Splines](#)

References

- Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc* 74(368):829–836. <https://doi.org/10.1080/01621459.1979.10481038>
- Cleveland WS, Devlin SJ (1988) Locally weighted regression: an approach to regression analysis by local fitting. *J Am Stat Assoc* 83(403): 596–610. <https://doi.org/10.1080/01621459.1988.10478639>
- Cleveland WS, Loader C (1996) Smoothing by local regression: principles and methods. In: Härdle W, Schimek MG (eds) *Statistical theory and computational aspects of smoothing*. Physica-Verlag HD, Heidelberg, pp 10–49. https://doi.org/10.1007/978-3-642-48425-4_2
- Loader C (1999) *Local regression and likelihood*. Springer, New York
- Loader C (2020) *locfit: local regression, likelihood and density estimation*. R package version 1.5–9.4. <https://CRAN.R-project.org/package=locfit>
- Wanishsakpong W, Notodiputro KA (2018) Locally weighted scatter-plot smoothing for analysing temperature changes and patterns in Australia. *Meteorol Appl* 25(3):357–364. <https://doi.org/10.1002/met.1702>

Logistic Attractors

Balakrishnan Ashok
Centre for Complex Systems and Soft Matter Physics,
International Institute of Information Technology Bangalore,
Electronics City, Bangalore, Karnataka, India

Definition

The logistic map is a common starting point for modeling growth phenomena in various fields due to not only its simple

form but due to the extensive and rich dynamics it displays. Systems in nature typically display nonlinear behavior, and the logistic map is a quadratic nonlinear equation that can be extensively analyzed, even as it shows both regular and chaotic behavior, depending upon the value of the control parameter. A logistic attractor, as the name implies, is an attractor associated with the logistic map and can correspond to either a fixed point or a limit cycle or even chaotic oscillations.

Logistic Map and Logistic Function

To understand how logistic attractors appear in nature, an overview of the associated logistic map or difference equation is required, as also the associated logistic differential equation and logistic function, which are discussed briefly in this section. The logistic map is the quadratic map defined as

$$x_{n+1} = rx_n(1 - x_n), \quad (1)$$

with $0 \leq x_n \leq 1$, $0 \leq r \leq 4$.

This was first discussed at length by May (1976) in the context of modeling population growth from one generation x_n , to the next, x_{n+1} , with a growth rate r serving as the control parameter. The logistic map is the discrete analog to the logistic differential equation

$$\frac{d}{dt}x(t) = rx(t)(1 - x(t)), \quad (2)$$

which has the logistic function

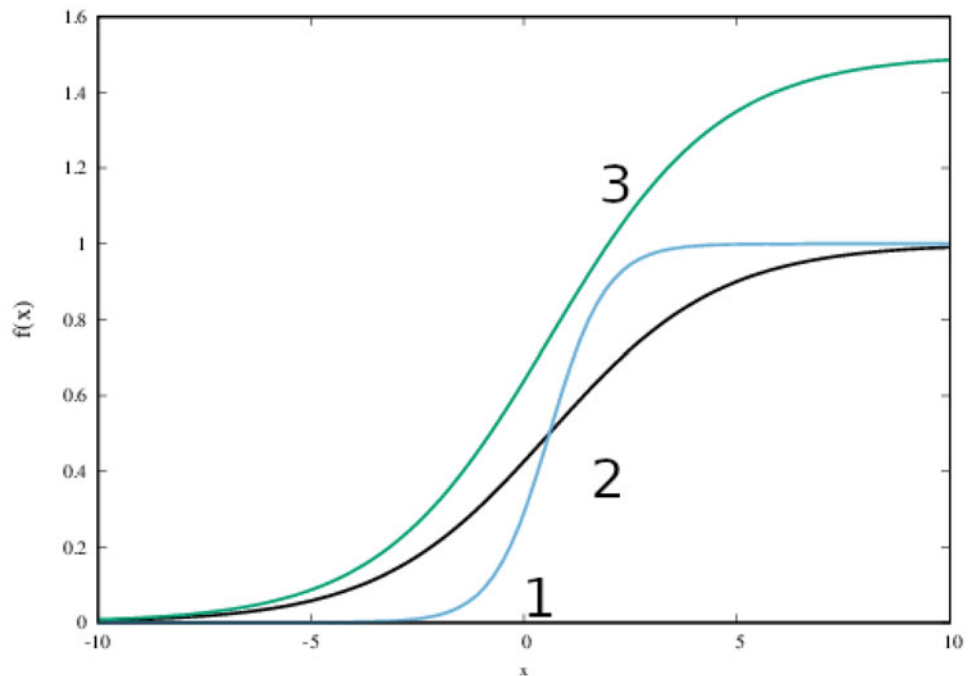
$$x(t) = m/(1 + \exp(-r(t - t_0))) \quad (3)$$

as its solution, with r being termed the logistic growth rate, and m being the maximum value of $x(t)$. Figure 1 shows examples of logistic functions with various values of growth rate and maximum value. The logistic differential equation was introduced by Pierre Francois Verhulst as an improvement over the Malthusian exponential growth model for population growth (Verhulst 1838). The assumptions made in formulating the logistic function are that the population in each new generation is proportional to the then-current one for smaller populations, and there is a maximum population level, M , reaching that results in population extinction in the subsequent generation. This gives

$$N_{n+1} = rN_n(1 - N_n/M), \quad (4)$$

If the population N_n in generation n is taken to represent the fraction of the maximum value M , and normalizing M so that $M = 1$ and $0 \leq N_n \leq 1$, with r as the growth rate, this has the same form as Eq. (1), which is also known as the logistic difference equation or the discrete logistic equation.

Logistic Attractors, Fig. 1 Plot of the logistic function $f(x)$ versus x for different values of the growth rate r and maximum value m and midpoint $x_0 = 0.6$: curve (1) $m = 1.0$, $r = 0.5$; curve (2) $m = 1.5$, $r = 0.5$; curve (3) $m = 1.0$, $r = 1.5$



Stability and Attractors

The logistic map has several attractors, which vary depending upon the value of the control parameter. These can be fixed points, limit cycles, or chaotic attractors. Fixed points, x^* , for the logistic map exist when

$$f(x^*) \equiv x^* = rx^*(1 - x^*), \quad (5)$$

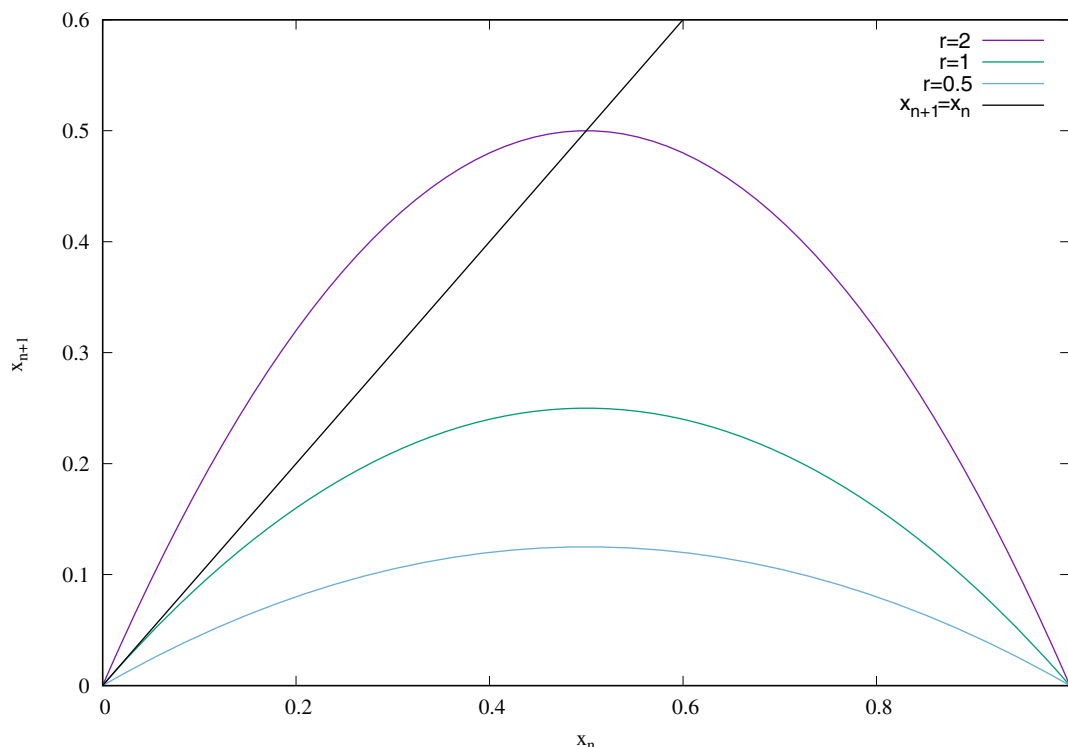
which would imply that $x^* = 0$ or $x^* = 1 - 1/r$. Regardless of the value of the control parameter, the logistic map always has a fixed point at the origin $x^* = 0$. $x^* = 1 - 1/r$ is a fixed point for $r \geq 1$. The stability of these fixed points can be obtained from $f'(x^*) = r - 2rx^*$, with stable fixed points resulting from $|f'(x^*)| < 1$ and unstable fixed points from $|f'(x^*)| > 1$. At $x^* = 0$, $f'(0) = r$, which would be stable for $r < 1$ and unstable for $r > 1$.

At the fixed point $x^* = 1 - 1/r$, $f'(x^*) = 2 - r$, stable fixed points would require $1 < r < 3$, and $r > 3$, resulting in all extant fixed points being unstable. Evolution of the stability of the system as a function of the control parameter can be observed from a return map of x_{n+1} versus x_n , as shown in Fig. 2. The intersection of the diagonal with the parabola would be at the fixed point since $x_{n+1} = x_n$ there. Three typical curves are shown for values of r less than, equal to, and greater than 1. As can be seen, for $r < 1$, the origin can be

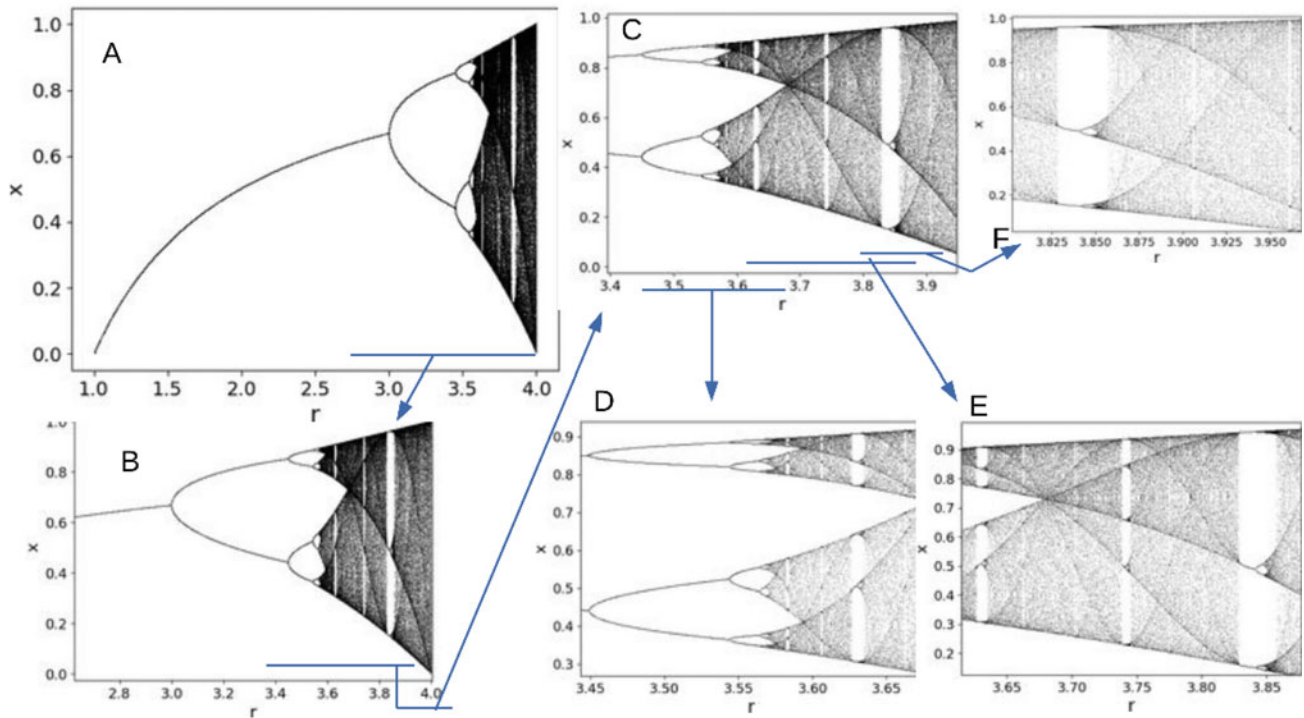
the only fixed point since the parabola falls below the diagonal. It will be noted that increasing r makes the parabola steeper so that the diagonal is tangent to it when $r = 1$. For $r > 1$, the second fixed point is located at the intersection of the diagonal with the parabola, with the other fixed point continuing to be the origin, which, however, loses stability. Hence, at $r = 1$, x^* bifurcating from the origin occurs through a transcritical bifurcation.

For $1 < r < 3$, the logistic map has an attractor at $x^* = 1 - 1/r$, which is a fixed point for the system. The system shows its first bifurcation to a period-2 cycle at $r = r_1 = 3$, where a flip bifurcation or period-doubling bifurcation occurs. This attractor exists for $3 \leq r \leq 3.449$, after which a second period-doubling bifurcation occurs at $r = r_2 = 3.449$ with the appearance of another attractor, a period-4 cycle, followed by further period-doubling attractors at $r_3 = 3.54409$, $r_4 = 3.5644$, and so on till $r = r_\infty = 3.569946$, at which point chaotic behavior sets in. The presence of these attractors can also be easily seen from the orbit diagram for the logistic map, shown in Fig. 3. As can be seen, the first period-doubling bifurcation occurs at $r = 3$, followed thereafter by further period-doubling bifurcations as mentioned above.

After the onset of chaotic behavior, there are again windows of periodic behavior, including the period-3 window that appears at $r \approx 3.83$. The appearance of a period-3 cycle in a system is an indicator of the presence of chaotic behavior –



Logistic Attractors, Fig. 2 Plot of x_{n+1} versus x_n for the logistic map, showing the location of the fixed points for three different r values



Logistic Attractors, Fig. 3 Orbit diagram for the logistic map showing period-doubling. Sections of panel (A) are shown magnified in (B–F). A period-3 cycle can be seen in panel F after $r \approx 3.83$

“period 3 implies chaos” as was proved by Li and Yorke in their famous paper with the same title (Li and Yorke 1975). Self-similarity can also be seen in the orbit diagram of the logistic map, with the branches seeming to repeat endlessly at finer and finer scales, as can be seen in Fig. 3. The appearance of n -cycles and attractors can also be observed by constructing a cobweb diagram for the map, which is a graphic method of locating fixed points and cycles. Cobweb diagrams illustrating the stability of the attractors of the logistic map are shown in Fig. 4 for six distinct r values, $r = 1.5, 2.0, 2.8, 3.2, 3.5$, and 3.9 . For $r = 1.5$ and $r = 2.0$, and $r = 2.8$, the fixed point at the origin $x^* = 0.0$ is a repelling one while the fixed point at $x^* = 1 - 1/r$ is attracting. That is, $x^* = 0.33, 0.5$, and 0.64 are attracting for $r = 1.5, 2$, and 2.8 , respectively. At $r = 3.2$, both the fixed points (at $x^* = 0$ and $x^* = 0.6875$) are repelling fixed points, as seen in part (d). The attractor at $r = 3.5$ (e) is a period-4 cycle while for $r = 3.9$ shown in (e), a multitude of cycles are possible, and a chaotic attractor is evident. The behavior of the logistic equation and the logistic map is extensively discussed in the dynamical systems theory literature, including, for example, in Strogatz (1994). In the context of geophysical applications, one can also refer to the work of Turcotte (Turcotte 1997) for an introduction to the logistic map and dynamical systems.

In Geophysics

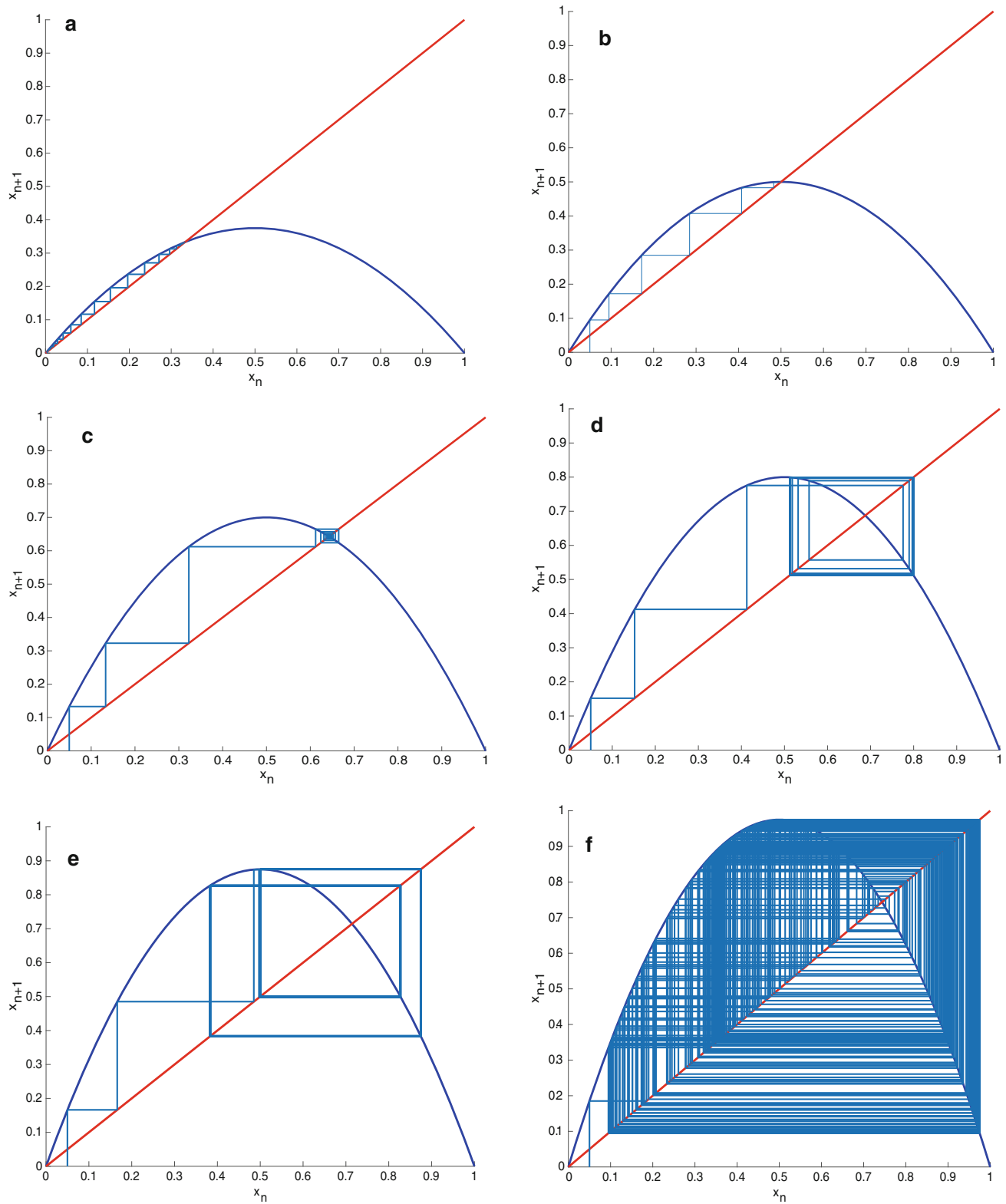
The logistic equation can be extended to more generalized forms, such as

$$\frac{d}{dt}x(t) = px(t)^q(1 - (x(t)/k)^c)t^n, \quad (6)$$

with p, q, k, c, n being constants, allowing for the modeling of various systems showing still further rich dynamics. In a recent work (Maslov and Anokhin 2012), Maslov and Anokhin used such a generalized logistic equation to derive the empirical Gutenberg–Richter equation relating the magnitude m of an earthquake to its frequency f (Gutenberg and Richter 1942), similar to that obtained by Ishimoto and Iida, earlier (Ishimoto and Iida 1939). The Gutenberg–Richter equation

$$\log(F) = a - bm \quad (7)$$

depicts the logarithmic relation between earthquake frequency and magnitude, a and b being constants. Considering one special case of the Eq. (6):



Logistic Attractors, Fig. 4 Cobweb diagram for the logistic map for $r =$ (a) 1.5, (b) 2.0, (c) 2.8, (d) 3.2, (e) 3.5, and (f) 3.9

$$\frac{d}{dt}x(t) = rx(t)(M - x(t))t^\alpha, \quad (8)$$

Maslov and Anokhin show that a generalized Gutenberg–Richter formula

$$\ln F(m) \approx \ln A - \frac{s}{\alpha + 1} m^{\alpha+1} \quad (9)$$

can be obtained by appropriate identification and redefinition of variables with those in the generalized logistic equation. The usual Gutenberg–Richter formula can be obtained by taking $\alpha = 0$, $a = \ln A$ and $b = s$.

A generalized logistic equation can also be used to model natural phenomena, including earthquakes and landslides, and the kinetics of such events that involve hierarchical aggregation, as has been proposed by Maslov and Chebotarev (Maslov and Chebotarev 2017), where the function $x(t)$ in Eq. (6) is replaced by $N(A)$, with A corresponding to the basic size of an element in a structure, and may correspond to, in the context of the problem being modeled, the magnitude of an earthquake or even area or mass. The function N is interpreted as a measure of the number of elements that are of size smaller than A . Equation 6 then serves to model the rate of growth of a particular phenomenon. The distributions of fore- and aftershocks of the 2014 Napa Valley earthquake, as well as the 2004 Sumatra earthquake, were modeled using the solution of the generalized logistic equation in Maslov and Chebotarev (2017) and showed good agreement, both qualitatively and quantitatively, with observed data points. Other researchers have also used generalized logistic functions to model fracture propagation and growth and include self-similarity in the fracture process (Lyakhovsky 2001).

Summary

Logistic attractors are the fixed points or limit cycles of the logistic difference equation. Despite the very simple form of this unimodal map, its nonlinear behavior gives rise to both simply periodic behavior, as well as period-doubling behavior with the number of stable solutions doubling, eventually showing chaotic behavior, after appropriately changing the value of the control parameter. The logistic map or its counterpart in continuous systems, the logistic equation, can be used to model varied growth phenomena and is useful in modeling geological phenomena such as earthquakes, fluid flows, as well as fracture propagation.

Bibliography

- Gutenberg B, Richter CF (1942) Earthquake magnitude, intensity, energy and acceleration. *Bull Seismol Soc Am* 32:163–191
- Ishimoto M, Iida K (1939) Observations of earthquakes registered with the microseismograph constructed recently. *Bull Earthq Res Inst Univ Tokyo* 17:443–478
- Li TY, Yorke JA (1975) Period three implies chaos. *Am Math Mon* 82: 985–992
- Lyakhovsky V (2001) Scaling of fracture length and distributed damage. *Geophys J Int* 144:114–122
- Maslov L, Anokhin V (2012) Derivation of the Gutenberg–Richter empirical formula from the solution of the generalized logistic equation. *Nat Sci* 4:648–651
- Maslov LA, Chebotarev VI (2017) Modeling statistics and kinetics of the natural aggregation structures and processes with the solution of generalized logistic equation. *Physica A* 468:691–697
- May R (1976) Simple mathematical models with very complicated dynamics. *Nature* 261:459–467
- Strogatz S (1994) *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry and engineering*. Perseus Books
- Turcotte DL (1997) *Fractals and chaos in geology and geophysics*, 2nd edn. Cambridge University Press
- Verhulst P-F (1838) Notice sur la loi que la population suit dans son accroissement. *Correspondance mathematique et physique* 10:113–121

Logistic Regression

D. Arun Kumar

Department of Electronics and Communication Engineering and Center for Research and Innovation, KSRM College of Engineering, Kadapa, Andhra Pradesh, India

Definition

Logistic regression (LR) generates the probability of outcome for the given input variable. The most commonly used LR is with binary outcome such as 1 or 0, true or false, and yes or no. In multinomial LR model the number of discrete outcomes are more than two. Logistic regression is used to predict the class label of an unknown sample/pattern to a category. LR is used in the data classification applications.

Introduction

Advancement in the sensor technology produced a large amount of spatial data. The spatial data is processed to obtain the information related to the ground area of interest (Bishop 1995). There are various methods of spatial data analysis techniques to extract the information from the data of

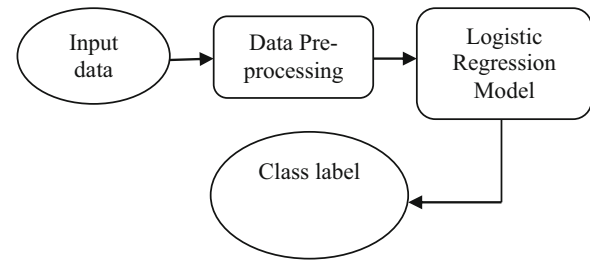
different dimensions. The spatial data analysis techniques are broadly classified as data classification, data regression, and data clustering. Data classification is assigning class label to the data sample (Duda et al. 2000). Data clustering is grouping the data samples in to individual groups without class labels. Data regression analysis is predicting the numeric values of output variable for the given input variables. The classification methods like Logistic regression, K-nearest neighbor classifier, minimum distance to mean classifier, etc., are used for spatial data analysis. The regression methods like linear regression, robust regression, probit regression, ridge regression, etc., are used to compute the values of output variable. The methods like K-means clustering, density-based clustering, and hierarchal clustering are used to group the data samples without class labels.

There exist various types of regression models to predict the unknown values of output variable. The models such as linear regression and logistic regression are frequently used to predict the values of a output variable. The linear regression model is used to predict continuous values of output variable such as predicting the slope value of a terrain from digital elevation model (DEM) image. The predicted continuous values of output variable are not sufficient to identify the number of classes present in the dataset. With this motivation logistic regression is suggested for data classification.

LR was first introduced in nineteenth century to predict the population growth (Bacaër 2011). The LR model was later updated in order to expand the model to classify the population of various countries (Wilson and Lorenz 2015). LR model is used to predict the discrete values of output variable. The discrete values are also called as class labels. The logistic regression model is mathematically represented in the form of nonlinear equation with activation function to provide the discrete values. There are various types of logistic regression models for data classification such as binary logistic regression, multinomial logistic regression, ordinal logistic regression model, etc. In the present study, a binary logistic regression model is explained with an example.

Logistic Regression Analysis

With traditional linear regression model, the output variable is generated as continuous values (Duda et al. 2000). The continuous values produced by linear regression have the limitation in producing class discrimination among the samples in the dataset. This limitation of regression model is solved by producing the probability of input sample/pattern to a given class. The functional block diagram of logistic regression model is provided in Fig. 1. The data samples are collected to prepare the dataset and furthermore, a logistic regression model is trained with the training samples. Let a sample/Pattern (P) in the dataset is associated with n number of



Logistic Regression, Fig. 1 Logistic Regression model for data analysis. The model considers input data, preprocessing of input data, training the logistic regression model, classifying the data, and assigning the class label to the data

feature values. The patterns are represented as points in n -dimensional feature space with each axis named as feature axis. The linear regression model for the training dataset is mathematically given in Eq. 1.

$$z = a_0 + a_1x_1 + a_2x_2 + \dots + a_nx_n, \quad (1)$$

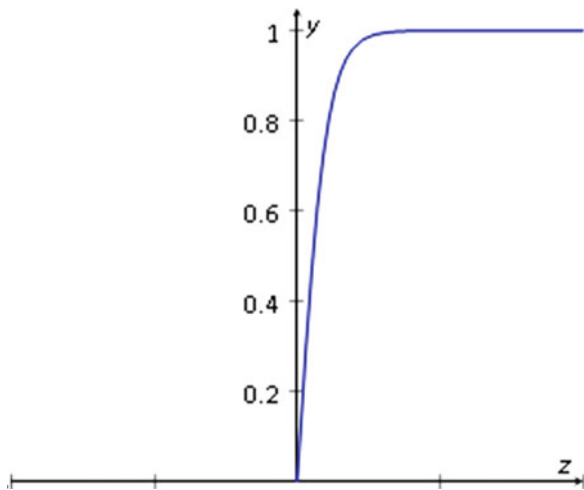
where $a_0, a_1, \dots, a_n$ are coefficients or weight parameters, $x_1, x_2, \dots, x_n$ are feature values for n features in a sample. During the training stage, the coefficients $a_0, a_1, \dots, a_n$ are updated by passing each input sample/pattern based on back propagating the output error (Theodoridis and Koutroumbas 1999). Linear regression model predicts continuous numeric values to the output variable data analysis (Cramer 2002). The model considers input data, preprocessing of input data, training the logistic regression model, classifying the data, assigning the class label to the input sample.

The limitation of linear regression is to predict discrete output value and represent the corresponding class label (Lever et al. 2016). The logistic regression model uses sigmoid activation function to generate the probabilities of input samples to the respective classes in the dataset. The sigmoid function generates output value in the range (0,1) for all the input real numbers (Hilbe 2009). Hence, if the input to the function is either a very large negative number or a very large positive number, the output is always between 0 and 1.

The predicted probabilities are used by the model to label sample to the respective class. The output of logistic regression model is computed as according to Eq. 1.

$$y = \frac{1}{(1 + e^{-z})} \quad (2)$$

The output value z is passed through sigmoid activation function and final output is obtained according to Eq. 2. The value of y is estimated in range [0,1] (according to Eq. 2) and the value indicate the probability of input sample belonging to a class (Ji and Telgarsky 2018). The plot for sigmoid activation function is given in Fig. 2.



Logistic Regression, Fig. 2 Pictorial representation of sigmoid activation function. The x-axis represents input variable (z) and y-axis represents output variable (y)

During the training phase, each sample/pattern is passed as an input and the corresponding output is computed. The output value of logistic regression model is compared with the target value and the output error is computed to update the coefficients. Updating the coefficients of logistic regression model is also termed as learning process. The learning process is implemented by decreasing the cost function near to zero by training the logistic regression model to multiple number of iterations. The cost of logistic regression model is given as,

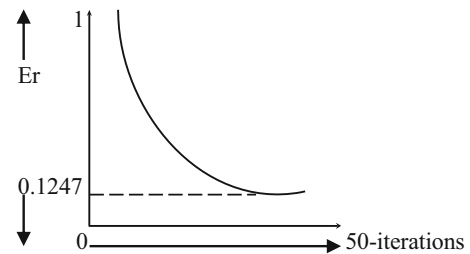
$$E_r = \frac{1}{2} \sum_{i=1}^c y_i - d_i \quad (3)$$

where i is the number of output variables, y_i is the computed output value for i th class and d_i is desired value for i th class. Pictorial representation of cost function for multiple number of iterations is given in Fig. 3.

Furthermore, the logistic regression model is tested with test data, and corresponding class label is assigned to the test pattern by applying max operator. The labeled patterns are evaluated based on various evaluation metrics. The detailed list of evaluation metrics is given in the following section.

Evaluation of Logistic Regression Model

The classified samples using logistic regression model in output image are evaluated using different performance metrics such as overall accuracy (OA), precision (P), recall (R), kappa coefficient (KC), and dispersion score (DS). These performance metrics are derived from confusion matrix (CM). CM is a table of actual class labels assigned by logistic regression model to the patterns of different classes. OA is defined as the sum of diagonal elements of CM divide by total number of samples. Precision is ratio of relevant classified samples to the



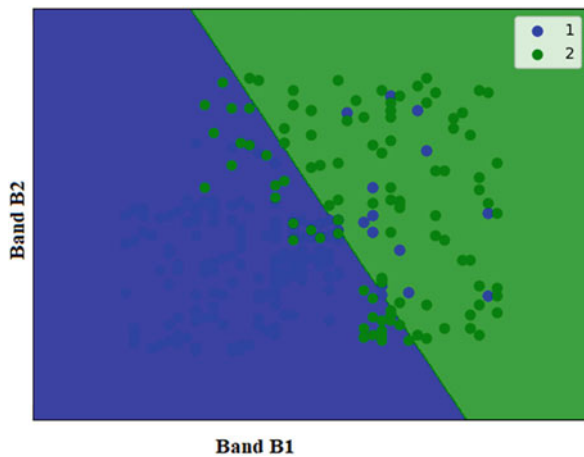
Logistic Regression, Fig. 3 Cost value of logistic regression reduced from maximum value down to zero over 50 number of iterations

retrieved classified samples. Recall (R) is the ratio of retrieved classified samples to relevant classified samples. Kappa coefficient (KC) is the measure of overall performance of the model based on the individual classwise agreement. Dispersion score (DS) measures the distribution of classified samples among the classes present in dataset and also, specify the overlapping characteristics of various class boundaries. A classifier with less DS to a particular class has a better ability to classify the samples to the corresponding class.

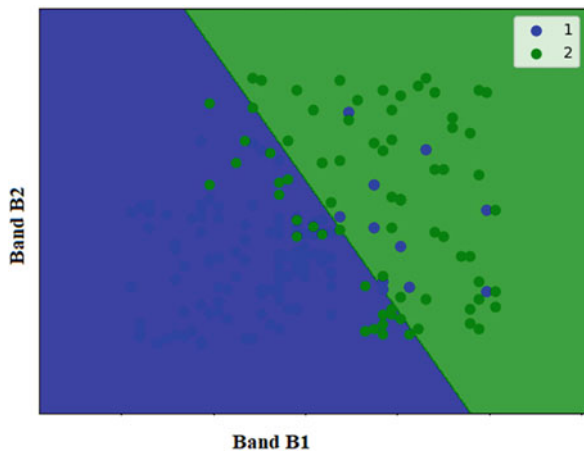
Example – Logistic Regression

In the present study, Sentinel Multispectral Imager (MSI) dataset with two Land use/Land cover classes like Forest and Water are considered to fit the logistic regression model. The dataset consists of 600 pixel vectors such as 300 pixel vectors for each class. Each pixel in MSI Sentinel image is characterized by 12 features (12 spectral bands such as Band 1, Band 2, Band 3, Band 4, ..., Band 12). The pixel is represented as a point in 12D feature space. The dataset is divided in to training and testing sets with 60% of the pixels selected in to training set and remaining 40% pixels are used for test set. The logistic regression model is trained with train dataset and tested with test dataset. The application of logistic regression model on training dataset is given in Fig. 4. In Fig. 4, B1 and B2 are feature axis and 1, 2 are land use/land cover classes where class 1 is Forest and class 2 is Water. In Fig. 4, few samples of class 1 are assigned to class 2 and few samples of class 2 are assigned to class 1. The logistic regression model is tested with test dataset and the classified pixels in B1-B2 feature space is given in Fig. 5. In Fig. 5, the model classified the test samples with 98% OA. The performance of logistic regression model is similar to OA as tested with other metrics. A few test samples of test data are misclassified to other classes as shown in Fig. 5.

Logistic regression performs better in discriminating the samples of various classes in comparison with similar type of models such as supervised kernel based support vector machines (SVM) and ensemble methods. The LR model has the limitation of handling the data with complex relationship between independent variable and dependent variable. The performance of LR is less when the decision boundary among



Logistic Regression, Fig. 4 Logistic regression model applied on training set. X-axis represents the feature axis (Band – B1) and Y-axis represents the feature axis (Band – B2). The training data consists of class 1 – Forest (1) and class 2 – Water (2)



Logistic Regression, Fig. 5 Logistic regression model applied on test set. X-axis represents the feature axis (Band – B1) and Y-axis represents the feature axis (Band – B2). The test data consists of class 1 – Forest (1) and class 2 – Water (2)

the classes in the dataset is nonlinear. Furthermore, the combination of LR and dimensionality reduction technique would reduce the complexity between the independent variable and dependent variable and in turn would produce better classification results.

Conclusion

In the present study, logistic regression model for data classification is discussed. Linear regression model is used to predict the continuous values for output variable. The limitation of linear regression model is its inability to produce discrete classes to the pixels. Logistic regression model produces discrete classes to the pixels unlike continuous values produced by linear regression model. The advantage of

logistic regression model is due to the usage of sigmoid activation function. In the present study, logistic regression model is trained and tested with Sentinel MSI dataset. The logistic regression model produced 98% overall accuracy as tested with Sentinel MSI dataset. The logistic regression model can also be extended to multiclass data classification problems in higher dimensional feature space.

Cross-References

- [Geographically Weighted Regression](#)
- [Machine Learning](#)
- [Regression](#)
- [Spatiotemporal Weighted Regression](#)

Bibliography

- Bacaër N (2011) Verhulst and the logistic equation (1838). In: A short history of mathematical population dynamics. Springer, London, pp 35–39
- Bishop CM (1995) Neural networks for pattern recognition. Oxford University Press, New York
- Cramer JS (2002) The origins of logistic regression. Tinbergen Institute Working Paper
- Duda RO, Hart PE, Stork DG (2000) Pattern classification, 2nd edn. Wiley Interscience Publications, New York
- Hilbe JM (2009) Logistic regression models. Chapman and Hall/CRC, Boca Raton
- Ji Z, Telgarsky M (2018) Risk and parameter convergence of logistic regression. In: Foundations of Machine Learning Reunion, University of California Press, Berkeley
- Lever J, Krzywinski M, Altman N (2016) Logistic regression. Nat Methods 13(8):603–604
- Theodoridis S, Koutroumbas K (1999) Pattern recognition and neural networks. In: Advanced course on artificial intelligence. Springer, pp 169–195
- Wilson JR, Lorenz KA (2015) Short history of the logistic regression model. In: Modeling binary correlated responses using SAS, SPSS and R. ICSA book series in statistics, vol 9. Springer, Cham. https://doi.org/10.1007/978-3-319-23805-0_2

Logistic Regression, Weights of Evidence, and the Modeling Assumption of Conditional Independence

Helmut Schaeben

Technische Universität Bergakademie Freiberg, Freiberg, Germany

Definition

Logistic regression is a generalization of linear regression in case the target variable is binary or categorical. Then the error

term is not normally distributed, and a least squares approach does not apply. The coefficients are estimated in terms of maximum likelihood leading to a system of nonlinear equations. A major application of logistic regression is the estimation of conditional probabilities updating prior unconditional probabilities. Assuming binary predictor variables that are jointly conditionally independent given the target variable the logistic regression model largely simplifies such that its coefficients get independent of one another and can be estimated by mere counting. They are recognized as Turing's Bayes factors in favor of $T = 1$ as against $T = 0$ provided by $B = 1$ (Good 2011), the Bayes log-factors, or eventually the weights of evidence (Good 1950, 1968). Mineral prospectivity modeling, where the data and the probabilities refer to given locations, applies among other methods logistic regression and weights of evidence even though they cannot account for spatially induced dependencies in the data and the probabilities to be estimated.

Introduction

The very definition of probability and its estimation have kept mankind, dilettantes, and professionals busy for centuries. Logistic regression and weights of evidence are two methods to estimate the conditional probability of an event given additional information conditioning the targeted event. The practical application of the latter is hampered by the requirement that the random variables representing the conditions are jointly conditionally independent given the random variable representing the target event. In prospectivity and exploration, geosciences are expected to predict the probability of a specified mineralization at a given location if conditions in favor or to its disadvantage are provided at this location. Neglecting the spatial reference of both the target and the predictor variables, logistic regression and, if joint conditional independence holds, weights of evidence apply. Weights of evidence has been geologists' most favorite method because of its apparent simplicity. However, the simplicity is deceiving as the modeling assumption of joint conditional independence is involved. The choice of the predictor variables to be sampled is usually guided by geological knowledge of the processes triggering a specified mineralization. Thus a flavor of causality modeling is added to the obvious regression problem to estimate a probability. In fact, conditional independence is a probabilistic approach to causality.

Mathematical Basics

Stochastic Independence

Let $(\Omega, \mathcal{A}, P)$ be a probability space. The events $B_1, \dots, B_m \in \mathcal{A}$ are jointly independent, if

$$P\left(\bigcap_{\ell=1}^k B_{i_\ell}\right) = \prod_{\ell=1}^k P(B_{i_\ell}) \quad (1)$$

for each k -tupel $(i_1, \dots, i_k)$ with $1 \leq i_1 < \dots < i_k \leq m$, $k = 2, \dots, m$. The random variables $B_1, \dots, B_m$ defined on Ω mapping into the sets S_ℓ of the measurable spaces $(S_\ell, \mathcal{S}_\ell)$, $\ell = 1, \dots, m$, are jointly independent, if the events $B_\ell = B_\ell^{-1}(E_\ell)$ for any $E_\ell \in \mathcal{S}_\ell$, $\ell = 1, \dots, m$, are jointly independent. If the random variables possess the probability measures P_{B_ℓ} with $P_{B_\ell}(E_\ell) = P(B_\ell^{-1}(E_\ell))$, $\ell = 1, \dots, m$, then their joint independence is equivalent to

$$P_{B_1, \dots, B_m} = P_{\otimes_{\ell=1}^m B_\ell} = \bigotimes_{\ell=1}^m P_{B_\ell}, \quad (2)$$

that is, their joint probability is equal to the product of their individual probabilities.

Stochastic Conditional Independence

The random events $B_1, \dots, B_m \in \mathcal{A}$ are jointly conditionally independent of the event $A \in \mathcal{A}$ with $P(A) \neq 0$, if

$$P\left(\bigcap_{\ell=1}^m B_\ell | A\right) = \prod_{\ell=1}^m P(B_\ell | A), \quad (3)$$

that is, if the joint conditional probability of $B_1, \dots, B_m$ given A agrees with the product of the individual conditional probabilities of B_ℓ , $\ell = 1, \dots, m$, given A . Eq. (3) is equivalent to

$$P\left(B_k | \bigcap_{\ell=1, \ell \neq k}^m B_\ell \cap A\right) = P(B_k | A). \quad (4)$$

The random variables $B_1, \dots, B_m$ are jointly conditionally independent of the random variable T defined on Ω , if

$$P_{B_1, \dots, B_m | T} = P_{\otimes_{\ell=1}^m B_\ell | T} = \bigotimes_{\ell=1}^m P_{B_\ell | T}, \quad (5)$$

that is, if their joint conditional probability given T agrees with the product of their individual conditional probabilities given T .

- Independence does not imply conditional independence and vice versa.
- Conditionally independent random variables may be (significantly) correlated or not.
- Conditionally independent random variables are conditionally uncorrelated.

Conditional independence is a probabilistic approach to causality while correlation is not. A comprehensive account of conditional independence is given in Dawid (1979).

Odds, logit Transform, and Logistic Function

In stochastics, an odds (in favor) O is defined as the ratio of the probabilities of an event and its complement. The logit transform is defined as the natural logarithm of an odds. Thus, for a random indicator variable T where $T = 1$ indicates presence and $T = 0$ indicates absence of some event, the logit transform of the probability P_T is defined as

$$\text{logit } P_T = \ln O_T = \ln \frac{P_T(1)}{P_T(0)} = \ln \frac{P_T(1)}{1 - P_T(1)} \quad (6)$$

provided that $P_T(0) \neq 0$. Odds and logit transform are depicted in Fig. 1.

The logistic function is defined as

$$A(z) = \frac{1}{1 + \exp(-z)}, \quad z \in \mathbb{R}. \quad (7)$$

Logit transform and logistic function are mutually inverse. The graph of the latter is shown in Figs. 1 and 2.

Estimating Probabilities

Odds, logit transform, and logistic function are instrumental in resolving the problem of estimating probabilities.

Let T denote a random indicator variable often referred to as target variable T coded such that the value $T = 1$ indicates the occurrence of an appealing event or the presence of an interesting phenomenon, for example, a certain mineralization. Let B_ℓ , $\ell = 1, \dots, m$, $m \geq 1$, denote a collection of categorical or real-valued continuous random predictor variables providing favorable or unfavorable conditions for the event $T = 1$. What are reasonable ways to estimate $P_{T|B_0 \otimes B_\ell}(1|1, \dots, b_\ell, \dots)$,

that is, the conditional probability of the event $T = 1$ given predictors B_ℓ , $\ell = 1, \dots, m$, and $B_0 \equiv 1$?

Logistic Regression

Logistic regression in its basic linear form statistically models, that is, estimates the logit transform $\text{logit } P_{T|\otimes_{\ell=0}^m B_\ell}$ of the conditional probability of a random indicator target variable T to be equal 1 given categorical or real-valued continuous predictor variables B_ℓ , $\ell = 1, \dots, m$, $m \geq 1$, by a linear combination of the predictor variables plus a constant, the intercept $B_0 \equiv 1$, that is, with $\otimes_{\ell=0}^m B_\ell = (B_0, B_1, \dots, B_m)^T$

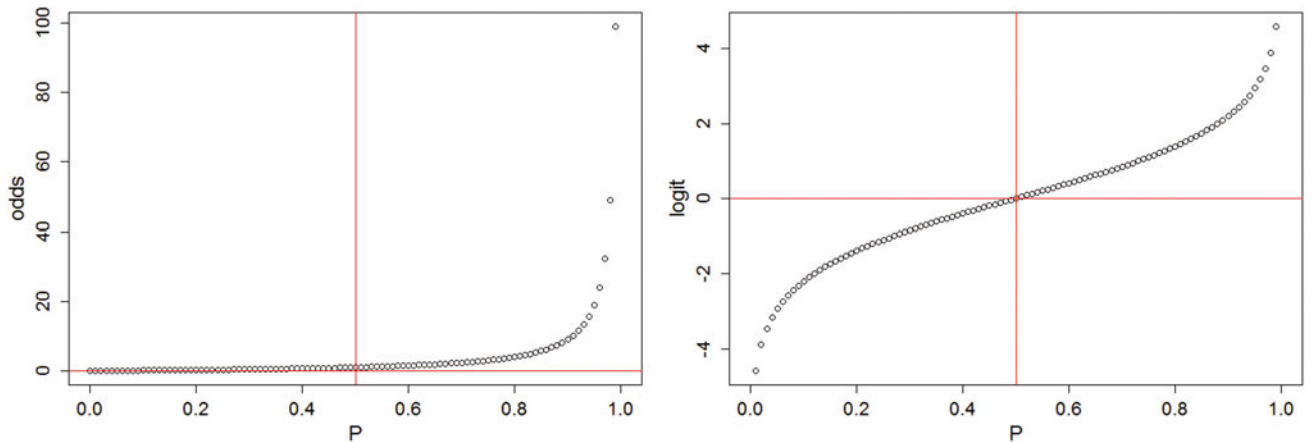
$$\text{logit } \widehat{P}_{T|\otimes B_\ell} = \beta_0 + \sum_{\ell} \beta_{\ell} B_{\ell}, \quad (8)$$

where the $\hat{\cdot}$ denotes the estimator. In practical applications the coefficients β_{ℓ} of the linear combination are estimated by $\hat{\beta}_{\ell}$ from data

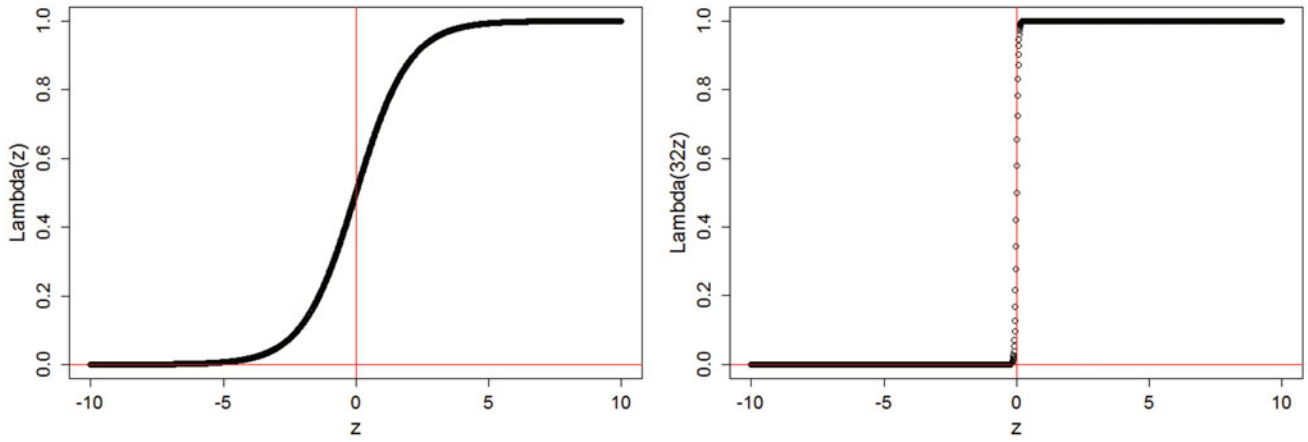
$$\text{logit } P_{T|\otimes B_\ell}(\widehat{1}|1, \dots, b_\ell, \dots) = \hat{\beta}_0 + \sum_{\ell} \hat{\beta}_{\ell} b_{\ell} \quad (9)$$

applying maximum likelihood estimation, and usually determined numerically by solving the corresponding system of nonlinear equations by application of iteratively reweighted least squares. Significance tests exist to check the reliability of the model.

Since logit transform and logistic function are inverse, the conditional probability of $T = 1$ given predictors $B_0, B_1, \dots, B_m$ is given by



Logistic Regression, Weights of Evidence, and the Modeling Assumption of Conditional Independence, Fig. 1 Graphs of odds O associated to probabilities P (left), and logit transform of probabilities P . (From Schaeben (2014b))



Logistic Regression, Weights of Evidence, and the Modeling Assumption of Conditional Independence, Fig. 2 Graphs of the sigmoidal logistic function $\Lambda(z)$ (left) and $\Lambda(32z)$ (right), indicating its fast convergence to the Heaviside function. (From Schaeben (2014b))

$$P_{T|\otimes_{\ell} B_{\ell}}(\widehat{1}|1, \dots) = \Lambda\left(\beta_0 + \sum_{\ell} \beta_{\ell} B_{\ell}\right) \quad (10)$$

Linear logistic regression is optimum, that is, exact, as the estimator agrees with the true conditional probability, if the predictors B_{ℓ} are indicator or categorical random variables and jointly conditionally independent given the target variable T (Schaeben 2014a).

A linear logistic regression model can be generalized into nonlinear logistic regression models when interaction terms, that is, products of predictors, are included

$$\begin{aligned} \text{logit} \widehat{P_{T|\otimes_{\ell} B_{\ell}}} &= \beta_0 + \sum_{\ell} \beta_{\ell} B_{\ell} \\ &+ \sum_{\ell_i, \dots, \ell_j} \beta_{\ell_i, \dots, \ell_j} B_{\ell_i} \dots B_{\ell_j}. \end{aligned} \quad (11)$$

Then

$$\widehat{P_{T|\otimes_{\ell} B_{\ell}}} = \Lambda\left(\beta_0 + \sum_{\ell} \beta_{\ell} B_{\ell} + \sum_{\ell_i, \dots, \ell_j} \beta_{\ell_i, \dots, \ell_j} B_{\ell_i} \dots B_{\ell_j}\right). \quad (12)$$

For indicator or categorical random predictor variables, proper interaction terms compensate any lack of conditional independence exactly. Then logistic regression with interaction terms is optimum, that is, the estimator agrees with the true conditional probability (Schaeben 2014a).

Given $m \geq 2$ predictor variables $B_{\ell} \neq 1$, $\ell = 1, \dots, m$, there is a total of $\sum_{\ell=2}^m \binom{m}{\ell} = 2^m - (m+1)$ possible interaction terms. To compensate a violation of conditional independence the interaction term $B_{\ell_1} \otimes \dots \otimes B_{\ell_k}$, $k \leq m$, is actually required, if $B_{\ell_1}, \dots, B_{\ell_k}$ are not jointly conditionally independent given T . Thus, recognizing which variables or interaction terms cause violations of conditional independence becomes

essential because the total number 2^m of all possible terms would have to be reasonably smaller than the sample size n to be a practically feasible model.

Logistic regression seems to have been systematically developed first by Berkson (1944), a standard reference for modern applied logistic regression is Hosmer et al. (2013), and a recent publication that focused on its application to prospectivity modeling is Kost (2020). Logistic regression is a method of statistical learning, the focus of which is inference of interpretable models. Often, the major progress of understanding and knowledge takes place by a trial-and-error approach to select the most powerful predictor variables or products thereof advancing possible inference most efficiently.

Weights of Evidence

Without any additional mathematical/statistical modeling assumption Bayes' theorem for the random indicator target variable T and several random indicator variables B_{ℓ} , $\ell = 1, \dots, m$, $m \geq 2$, gives

$$\begin{aligned} \text{logit} P_{T|\otimes_{\ell=1}^m B_{\ell}}(1|\dots) &= \text{logit} P_{T|\otimes_{\ell=1}^{m-1} B_{\ell}}(1|\dots) \\ &+ \ln F_m \end{aligned} \quad (13)$$

$$= \text{logit} P_T(1) + \sum_{\ell=1}^m \ln F_{\ell} \quad (14)$$

with

$$F_{\ell} = \frac{P_{B_{\ell}|\otimes_{i=0}^{\ell-1} B_i \otimes T}(\cdot|1, \dots, 1)}{P_{B_{\ell}|\otimes_{i=0}^{\ell-1} B_i \otimes T}(\cdot|1, \dots, 0)}, \ell = 1, \dots, m. \quad (15)$$

Assuming joint conditional independence of $B_1, \dots, B_m$ given T simplifies the factors F_{ℓ} to

$$F_\ell = \frac{P_{B_\ell|T}(\cdot|1)}{P_{B_\ell|T}(\cdot|0)}, \ell = 1, \dots, m. \quad (16)$$

Explicitly distinguishing the two possible outcomes for each B_ℓ and taking logarithms results in the weights of evidence

$$W_\ell(i) = \ln \frac{P_{B_\ell|T}(i|1)}{P_{B_\ell|T}(i|0)}, i = 0, 1, \quad (17)$$

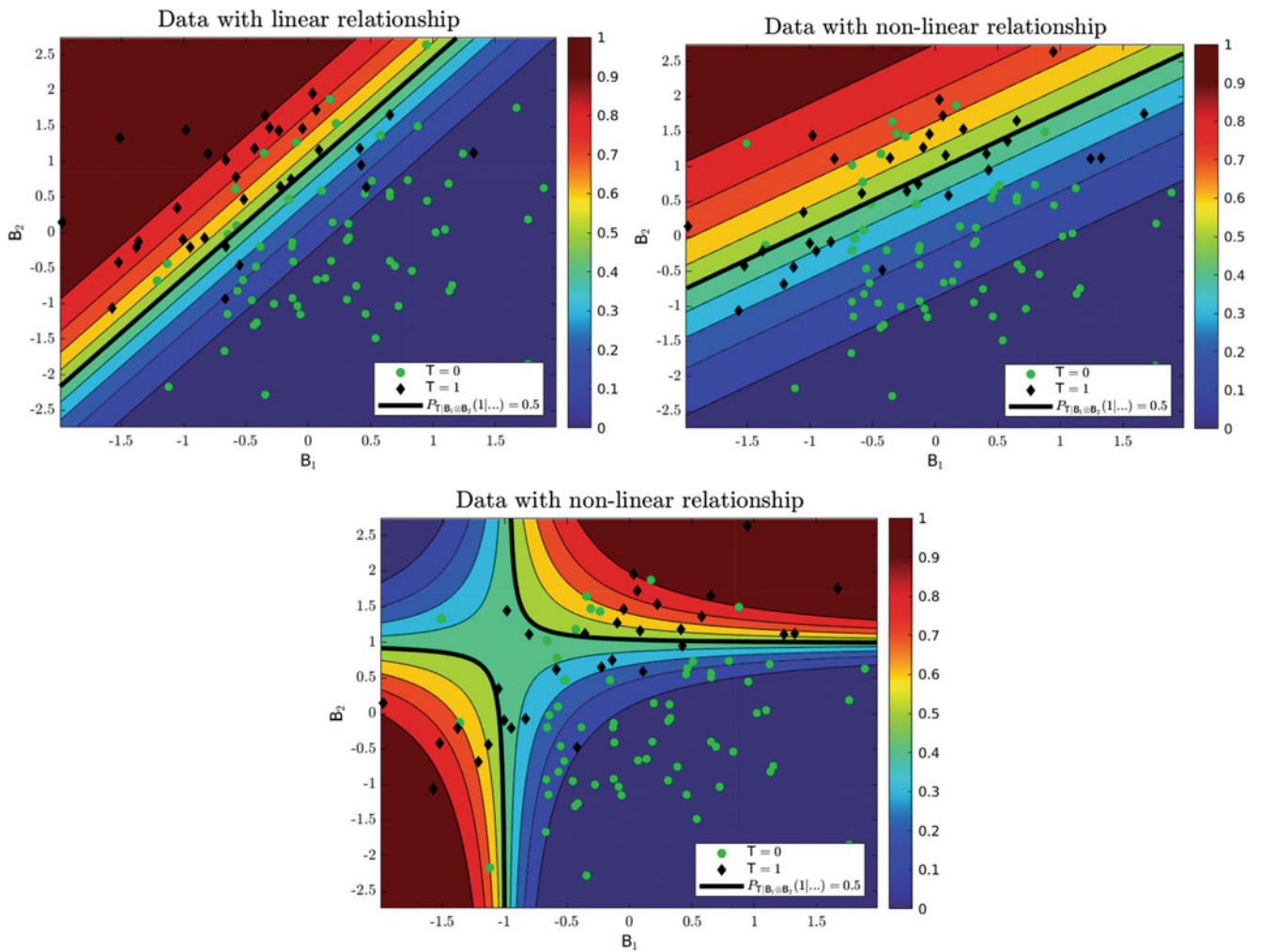
and their contrasts

$$C_\ell = W_\ell(1) - W_\ell(0), \ell = 1, \dots, m, \quad (18)$$

and eventually in the log-linear form of Bayes' theorem referred to as the method of "weights of evidence" in geology

$$\begin{aligned} \text{logit } P_{T|\otimes_\ell B_\ell}(1|\dots) &= C_0 + \sum_\ell C_\ell B_\ell \\ &= C_0 + \sum_{\ell: B_\ell \neq 0} C_\ell B_\ell \end{aligned} \quad (19)$$

$$P_{T|\otimes_\ell B_\ell}(1|\dots) = \mathcal{A} \left(C_0 + \sum_{\ell: B_\ell \neq 0} C_\ell B_\ell \right) \quad (20)$$



Logistic Regression, Weights of Evidence, and the Modeling Assumption of Conditional Independence, Fig. 3 Classification induced by fitted logistic regression models. Realizations of B_1 , B_2 are generated according to a bivariate normal distribution with mean 0 and standard deviation 1. They are plotted in the plane spanned by B_1 and B_2 . Realizations of the indicator random variable T are generated according to the binomial distribution $B(1, p)$ with (i) p as provided by the linear model, that is, $p = \mathcal{A}(-2 - 2B_1 + 2B_2)$, and (ii) p as provided by the nonlinear model including the interaction term, that is,

$p = \mathcal{A}(-2 - 2B_1 + 2B_2 + 2B_1B_2)$, respectively. Linear logistic regression analysis applied to simulated data of model (i) yields the two classes separated by the straight line shown at the top left; linear logistic regression analysis applied to simulated data of model (ii) yields the two classes separated by the straight line shown at the top right; non-linear logistic regression analysis including the interaction term B_1B_2 applied to simulated data of model (ii) yields the two classes separated by the curve shown at the bottom center.

with $C_0 = \logit P_T(1) + \sum_{\ell} W_{\ell}(0)$. Thus, weights of evidence is first of all a straightforward application of Bayes' theorem. In practical applications the weights, that is, the involved conditional probabilities, are estimated by relative frequencies, that is, by counting.

Weights of evidence were introduced by (Good 1950, 1968), and with respect to potential modeling or prospectivity modeling into geology by Bonham-Carter et al. (1989) and Agterberg et al. (1990).

Relationship of Linear Logistic Regression and Weights of Evidence

Comparing Eq. (19) and Eq. (8) reveals with $\beta_0 = C_0$ and $\beta_{\ell} = C_{\ell}$, $\ell = 1, \dots, m$, immediately that weights of evidence is the special case of logistic regression for jointly conditionally independent random indicator predictor variables (Schaeben 2014b).

Additional light is shed on logistic regression and weights of evidence, respectively, from a classification point of view. Thresholding of estimated conditional probabilities may be applied to design a $\{0,1\}$ classification, usually with respect to the estimator of the conditional median of the target variable

$$\hat{T}_{0.5} = \begin{cases} 0 & \text{if } \hat{P}_{T|\otimes_{\ell} B_{\ell}}(1|\dots) \leq 0.5, \\ 1 & \text{otherwise.} \end{cases} \quad (21)$$

For logistic regression the locus for which $\hat{P}_{T|\otimes_{\ell} B_{\ell}}(1|\dots) = 0.5$ is given by setting

$$\hat{\beta}_0 + \sum_{\ell} \hat{\beta}_{\ell} B_{\ell} + \sum_{\ell_i, \dots, \ell_j} \hat{\beta}_{\ell_i, \dots, \ell_j} B_{\ell_i} \otimes \dots \otimes B_{\ell_j} = 0. \quad (22)$$

As Fig. 3 illustrates, if all interaction terms vanish, that is, in the case of linear logistic regression, Eq. (22) represents a hyperplane in the space spanned by the contributing B_{ℓ} ; in the general nonlinear case Eq. (22) represents a curved surface separating the two classes.

The method of weights of evidence cannot include interaction terms, because along the derivation from Bayes' general theorem for several random indicator variables, Eq. (14), to weights of evidence, Eq. (17), they are made to vanish by the very assumption of joint conditional independence. Therefore the two classes of the classification induced by weights of evidence are always separated by a hyperplane.

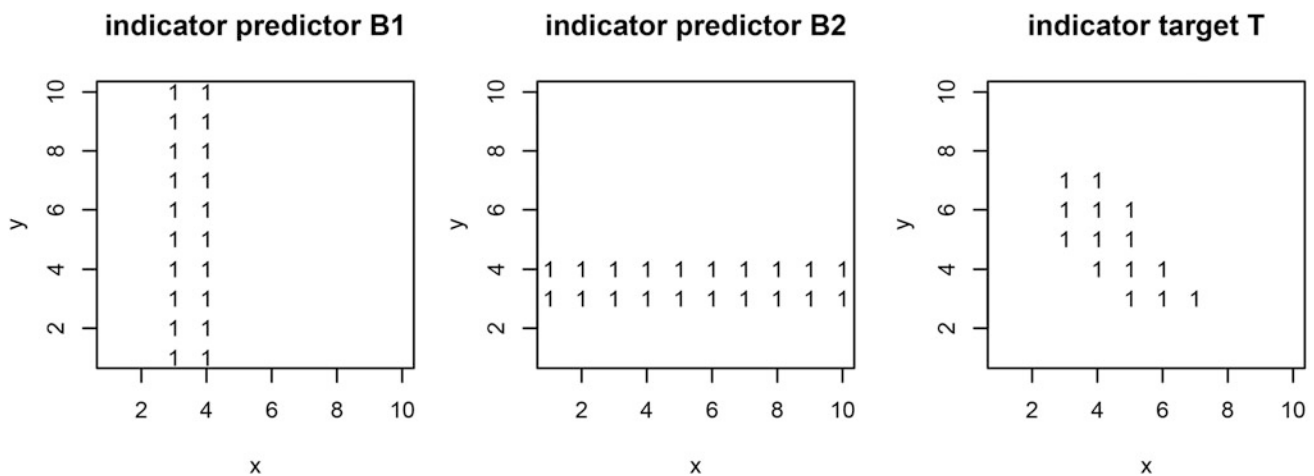
Another way of comparing methods of estimating probabilities or conditional probabilities is by precision-recall curves (PR) rather than receiver operating characteristic (ROC) curves (Davis and Goadrich 2006) because PR curves are more appropriate if the target event $T = 1$ is a rare event as often experienced in practical applications.

Practical Example with Fabricated Data

One hundred triplets of realizations (b_{1i}, b_{2i}, t_i) , $i = 1, \dots, 100$, of random indicator variables (B_1, B_2, T) are assigned to 100 pixels arranged in a (10×10) -array of pixels of digital map images. A specific assignment is depicted in Fig. 4 and referred to as dataset RANKIT.

The assignment could be different without changing the ordinary statistics; however, spatial statistics as captured by a variogram would be different (Schaeben 2014b). Obviously, logistic regression and weights of evidence are nonspatial methods, as they are independent of any specific spatial assignment.

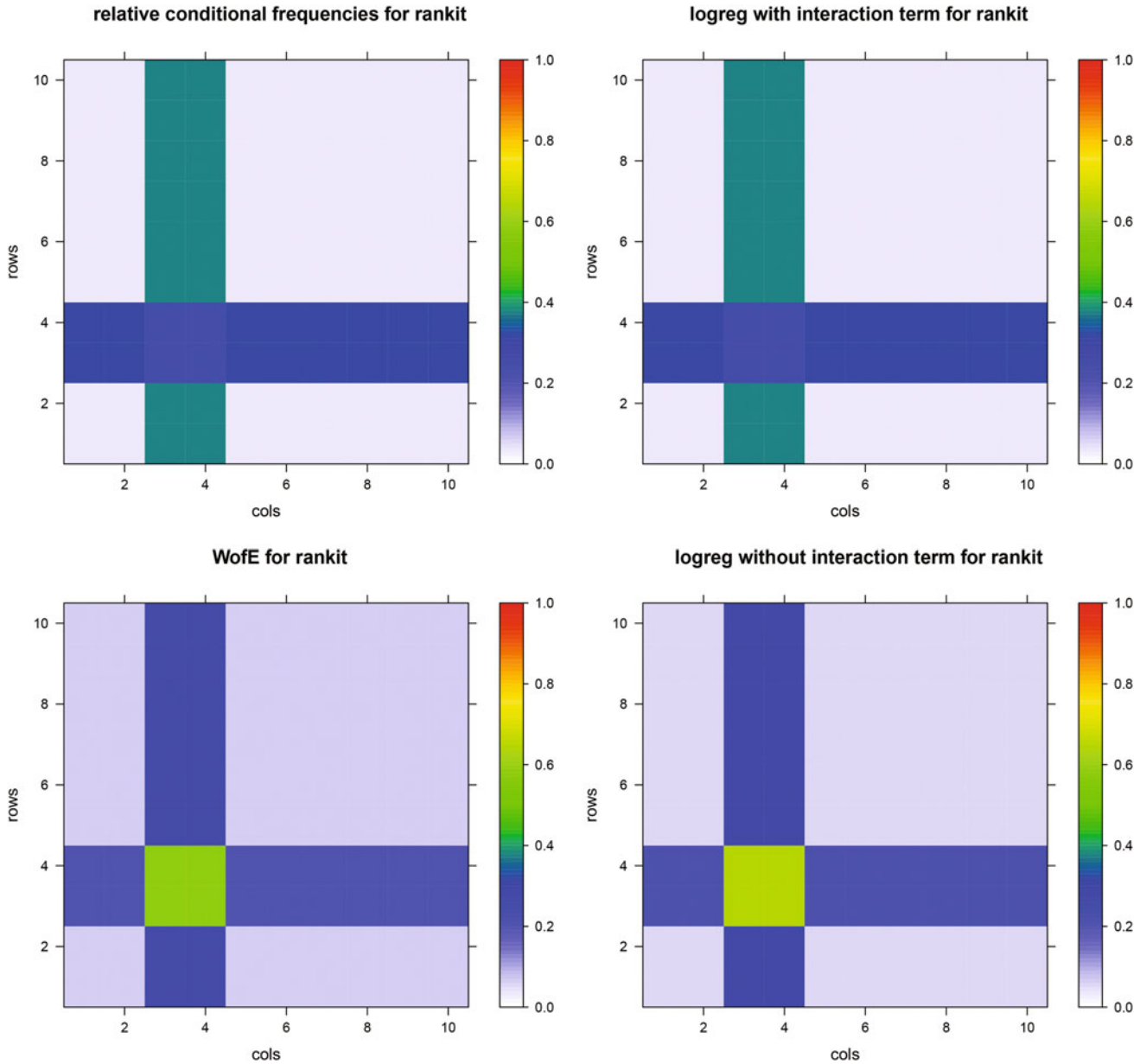
Testing the null hypothesis of joint conditional independence with the standard log-likelihood ratio test yields



Logistic Regression, Weights of Evidence, and the Modeling Assumption of Conditional Independence, Fig. 4 Spatial distribution of two indicator predictor variables B_1 , B_2 , and the indicator target variable T of the dataset RANKIT. (From Schaeben (2014b))

Logistic Regression, Weights of Evidence, and the Modeling Assumption of Conditional Independence, Table 1 Comparison of predicted conditional probabilities for various methods applied to the training dataset RANKIT

$\hat{P}_{T B_1 \otimes B_2}(1 \cdot)$	Elementary counting	WofE	Linear LogReg	Nonlinear LogReg
$B_1 = 1, B_2 = 1$	0.25000	0.58636	0.64055	0.25000
$B_1 = 1, B_2 = 0$	0.37500	0.26875	0.27736	0.37500
$B_1 = 0, B_2 = 1$	0.31250	0.20156	0.21486	0.31250
$B_1 = 0, B_2 = 0$	0.03125	0.06142	0.05565	0.03125



Logistic Regression, Weights of Evidence, and the Modeling Assumption of Conditional Independence, Fig. 5 Spatial distribution of predicted $\hat{P}(T=1|B_1B_2)$ for the training dataset RANKIT according to elementary estimation (top left), logistic regression with

interaction term (top right), weights-of-evidence (bottom left), and logistic regression without interaction term (bottom right). (From Schaben (2014b))

$p = 0.06443468$ and suggests the decision to reject the null hypothesis (Schaeben 2014b).

The fitted models of weights of evidence, linear logistic regression without interaction term, and nonlinear logistic regression with interaction term read explicitly as follows:

$$\begin{aligned} \text{WofE} : \hat{P}_{T|B_1 \otimes B_2}(1|..) \\ = A(-2.726 + 1.725 B_1 + 1.349 B_2) \end{aligned} \quad (23)$$

$$\begin{aligned} \text{linLogReg} : \hat{P}_{T|B_1 \otimes B_2}(1|..) \\ = A(-2.831 + 1.874 B_1 + 1.535 B_2) \end{aligned} \quad (24)$$

$$\begin{aligned} \text{non - linLogReg} : \hat{P}_{T|B_1 \otimes B_2}(1|..) \\ = A(-3.434 + 2.923 B_1 + 2.646 B_2 - 3.233 B_1 B_2) \end{aligned} \quad (25)$$

Since the mathematical modeling assumption of conditional independence is violated, only logistic regression with interaction terms yields a proper model and predicts the conditional probabilities almost exactly.

The results of weights of evidence, logistic regression with or without interaction terms, applied to the fabricated dataset RANKIT are summarized in Table 1, and Fig. 5 depicts the results of dataset RANKIT.

The example illustrates that application of weights of evidence even though the predictor variables are not jointly conditionally independent of the target variable corrupts its estimated conditional probabilities and assigns their largest values to false locations, that is, false pixels of the digital map image. Logistic regression is the canonical generalization of weights of evidence as it exactly compensates lack of joint conditional independence of random indicator predictor variables by including interaction terms corresponding to the actual violations of joint conditional independence. For random indicator predictor variables logistic regression is optimum. Thus there is no need for ad hoc recipes to fix weights of evidence corrupted by lack of joint conditional independence. Eventually logistic regression is not restricted to random indicator predictor variables, the predictors may be real-valued continuous random variables, rendering their transform to indicator predictors by thresholding as required by weights of evidence dispensable.

Conclusions

Logistic regression provides a functional model to estimate conditional probabilities. The regression coefficients are determined according to the maximum likelihood method

applying the truncated iteratively re-weighted least squares algorithm. If the functional form is restricted to linear dependencies only, and if the predictor variables are assumed to be binary and jointly conditionally independent given the binary target variable, then logistic regression largely simplifies to the formula of Bayes' theorem for several variables referred to as weights of evidence; its numerics is largely simplified to counting as means to estimate probabilities from empirical frequencies of events.

References

- Agterberg FP, Bonham-Carter GF, Wright DF (1990) Statistical pattern integration for mineral exploration. In: Gaal G (ed) Computer applications in resource exploration. Pergamon, Oxford, pp 1–22
- Berkson J (1944) Application of the logistic function to bio-assay. *J Am Stat Assoc* 39:357
- Bonham-Carter GF, Agterberg FP, Wright DF (1989) Weights of evidence modeling: a new approach to mapping mineral potential. In: Bonham-Carter GF, Agterberg FP (eds) Statistical applications in the earth sciences, paper 89-9. Geological Survey of Canada, pp 171–183
- Davis J, Goadrich M (2006) The relationship between precision-recall and roc curves. In: Cohen WW, Moore A (eds) Proceedings of the 23rd international conference on machine learning (ICML'06), pp 233–240
- Dawid AP (1979) Conditional independence in statistical theory. *J R Stat Soc B* 41:1
- Good IJ (1950) Probability and the weighting of evidence. Griffin, London
- Good IJ (1968) The estimation of probabilities: an essay on modern Bayesian methods. Research monograph no 30. The MIT Press, Cambridge
- Good IJ (2011) A list of properties of Bayes-Turing factors unclassified by NSA in 2011. <https://www.nsa.gov/news-features/declassified-documents/tech-journals/assets/files/list-of-properties.pdf>
- Hosmer DW, Lemeshow S, Sturdivant RX (2013) Applied logistic regression, 3rd edn. Wiley, Hoboken
- Kost S (2020) PhD thesis, TU Bergakademie Freiberg. <https://nbn-resolving.de/urn:nbn:de:bsz:105-qucosa2-728751>
- Schaeben H (2014a) A mathematical view of weights-of-evidence, conditional independence, and logistic regression in terms of markov random fields. *Math Geosci* 46:691
- Schaeben H (2014b) Potential modeling: conditional independence matters. *Int J Geomath* 5:99

Log-Likelihood Ratio Test

Alejandro C. Frery
School of Mathematics and Statistics, Victoria University of Wellington, Wellington, New Zealand

Definition

The log-likelihood ratio test contrasts two parametric models, of which one is a subset or restriction of the other.

Consider the random sample $\mathbf{X} = (X_1, X_2, \dots, X_n)$ from the model characterized by the distribution $\mathcal{D}(\theta)$, with $\theta \in \Theta \subset \mathbb{R}^p$. One may be interested in checking the null hypothesis that the model for the data belongs to a subset $\mathcal{H}_0 : \theta \in \Theta_0 \subset \Theta$ of all the possible models, versus the alternative $\mathcal{H}_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$.

An approach consists in comparing the likelihoods of the sample under $\mathcal{H}_0$ and under the unrestricted model. The comparison can be made by computing the ratio of the likelihoods, or the logarithm of such ratio.

Adapting the definition given by Casella and Berger (2002), we will say that a likelihoods ratio test for contrasting $\mathcal{H}_0 : \theta \in \Theta_0$ against $\mathcal{H}_1 : \theta \in \Theta \setminus \Theta_0$ is any test that rejects $\mathcal{H}_0$ if

$$\frac{\sup_{\theta \in \Theta} L(\theta; \mathbf{X})}{\sup_{\theta \in \Theta_0} L(\theta; \mathbf{X})} > c', \quad (1)$$

where $L(\theta; \mathbf{X})$ is the likelihood of the parameter θ for the sample $\mathbf{X}$ and c' is a constant that depends on the desired significance.

Assume the maximum likelihood estimator for the problem exists. The numerator in Eq. (1) is the likelihood over all possible values of the parameter, which is maximized by the maximum likelihood estimator $\hat{\theta}$. The denominator in Eq. (1) is the maximum likelihood estimator restricted to Θ_0 ; denote it $\hat{\theta}_0$. Inequality (Eq. 1) becomes

$$\frac{L(\hat{\theta}; \mathbf{X})}{L(\hat{\theta}_0; \mathbf{X})} > c', \quad (2)$$

Since both numerator and denominator are positive,

$$\begin{aligned} \log \frac{L(\hat{\theta}; \mathbf{X})}{L(\hat{\theta}_0; \mathbf{X})} &= \log L(\hat{\theta}; \mathbf{X}) - \log L(\hat{\theta}_0; \mathbf{X}) \\ &= \ell(\hat{\theta}; \mathbf{X}) - \ell(\hat{\theta}_0; \mathbf{X}) > \log c' = c \end{aligned}$$

is the same test, in which $\ell(\theta; \mathbf{X})$ is the reduced log-likelihood of θ for $\mathbf{X}$.

We will see two examples before completing the theory of likelihood ratio tests.

Testing Under the Multinomial Distribution

Consider the Example 12.6.6 from Lehman and Romano (2005). The random sample $\mathbf{X} = (X_1, X_2, \dots, X_n)$ follows a multinomial distribution, i.e., each random variable may assume one of $k + 1$ values with probability $\Pr(X = j) = p_j$,

$j \in \{1, 2, \dots, k + 1\}$. The parameter space is $\Theta = \{\mathbf{p} = (p_1, p_2, \dots, p_k) : p_j \geq 0 \text{ and } \sum_{j=1}^k p_j \leq 1\}$. We want to test the null hypothesis that we know the k probabilities that index the distribution, i.e., $\mathcal{H}_0 : \{p_i = \pi_i, i \in \{1, 2, \dots, k\}\}$, since $p_{k+1} = 1 - \sum_{j=1}^k p_j$. The parameter space in the null hypothesis is the point $\Theta_0 = \boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_k)$, which is known. Denote Y_i the number of observations in $\mathbf{X}$ that are equal to i . The likelihood of the sample is

$$\begin{aligned} L(\mathbf{p}; \mathbf{X}) &= \frac{n!}{Y_1! Y_2! \dots Y_{k+1}!} p_1^{Y_1} p_2^{Y_2} \dots p_{k+1}^{Y_{k+1}} \\ &= \frac{n!}{Y_1! Y_2! \dots Y_k! (n - \sum_{j=1}^k Y_j)!} p_1^{Y_1} p_2^{Y_2} \dots \\ &\quad p_k^{Y_k} \left(1 - \sum_{j=1}^k p_j\right)^{n - \sum_{j=1}^k Y_j}, \end{aligned}$$

and the maximum likelihood estimator in Θ is the vector of proportions

$$\begin{aligned} \hat{\mathbf{p}} &= (\hat{p}_1, \hat{p}_2, \dots, \hat{p}_k, \widehat{p_{k+1}}) \\ &= \left(\frac{Y_1}{n}, \frac{Y_2}{n}, \dots, \frac{Y_k}{n}, 1 - \frac{\sum_{j=1}^k Y_j}{n}\right). \end{aligned}$$

The likelihoods ratio test statistic is

$$\frac{L(\hat{\mathbf{p}}; \mathbf{X})}{L(\boldsymbol{\pi}; \mathbf{X})} = \frac{\hat{p}_1^{Y_1} \hat{p}_2^{Y_2} \dots \widehat{p_{k+1}}^{Y_{k+1}}}{\pi_1^{Y_1} \pi_2^{Y_2} \dots \pi_{k+1}^{Y_{k+1}}} = \left(\frac{\hat{p}_1}{\pi_1}\right)^{Y_1} \left(\frac{\hat{p}_2}{\pi_2}\right)^{Y_2} \dots \left(\frac{\widehat{p_{k+1}}}{\pi_{k+1}}\right)^{Y_{k+1}}.$$

Both numerator and denominator are positive, and the latter, being constrained, cannot be larger than the former. Taking the logarithm, we obtain the log-likelihoods ratio:

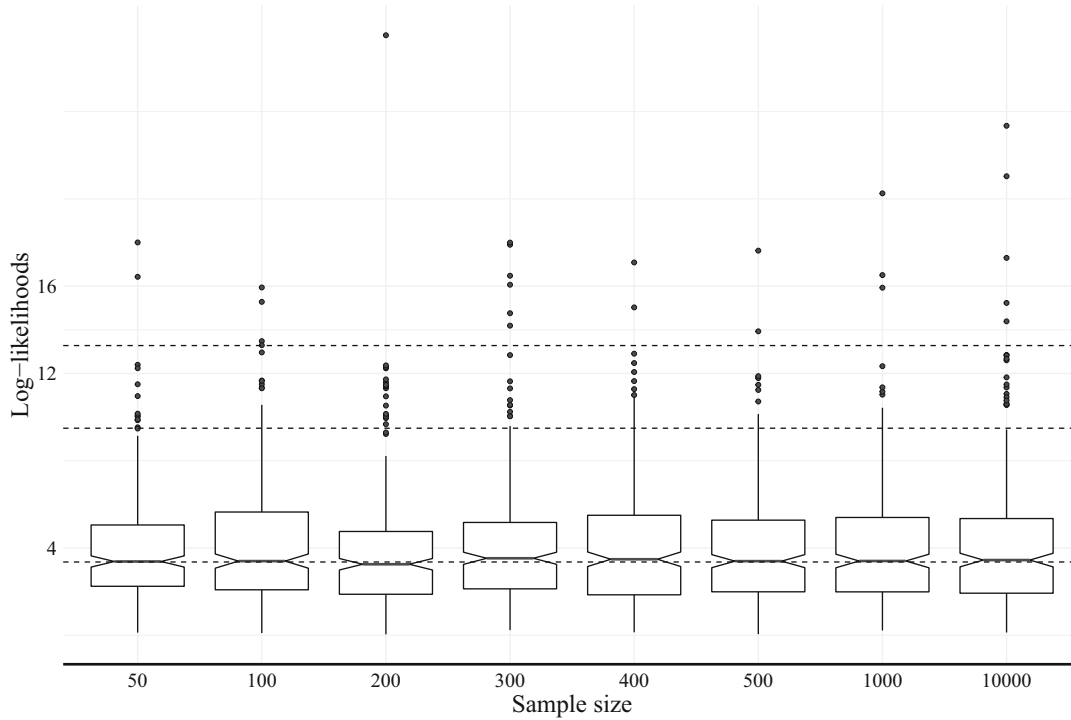
$$\ell(\hat{\mathbf{p}}; \mathbf{X}) - \ell(\boldsymbol{\pi}; \mathbf{X}) = \sum_{j=1}^{k+1} Y_j \log \frac{\hat{p}_j}{\pi_j} = n \sum_{j=1}^{k+1} \hat{p}_j \log \frac{\hat{p}_j}{\pi_j}. \quad (3)$$

Finally, the test statistic

$$R_n(\mathbf{X}) = 2[\ell(\hat{\mathbf{p}}; \mathbf{X}) - \ell(\boldsymbol{\pi}; \mathbf{X})]$$

becomes useful by knowing that, under $\mathcal{H}_0$, it is asymptotically a χ_k^2 -distributed random variable. Notice that Eq. (3) is closely related to the Kullback-Leibler divergence between $\hat{\mathbf{p}}$ and $\boldsymbol{\pi}$.

We now illustrate this example with simulations. Fixing $\boldsymbol{\pi} = (0.3, 0.2, 0.1, 0.25, 0.15)$, we obtain samples of size $n \in \{50, 100, 200, 300, 400, 500, 10^3, 10^4\}$ from this distribution, then we compute $\hat{\mathbf{p}}$ and obtain R_n . Since R_n is asymptotically distributed as χ_4^2 , the asymptotic values of the



Log-Likelihood Ratio Test, Fig. 1 R_n samples boxplots under $\mathcal{H}_0$ for $n \in \{50, 100, 200, 300, 400, 500, 10^3, 10^4\}$, along with the mean (Eq. 4), median, and 95% and 99% quantiles (dashed lines) of a χ_4^2 -distributed random variable

mean, (approximate) median, and quantiles of order 95% and 99% are, respectively, 4, 3.36, 9.49, and 13.28. Figure 1 shows the boxplots of these test statistics after 300 independent replications for each situation, along with these values. Notice that there is good agreement between the sample values and asymptotic quantities.

Testing Under the Normal Distribution

Consider the situation in which $\mathbf{X}$ is a collection of normal random variables with unknown mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. The parameter space is $(\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+$. The densities are of the form $f(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp \{-(-x - \mu)^2/(2\sigma^2)\}$, so the reduced likelihood and log-likelihood are, respectively,

$$L(\mu, \sigma^2; \mathbf{X}) = (\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \right\}, \quad \text{and} \quad (4)$$

$$\ell(\mu, \sigma^2; \mathbf{X}) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2. \quad (5)$$

The maximum likelihood estimator of (μ, σ^2) is, thus, the pair $(\hat{\mu}, \hat{\sigma}^2)$ given by

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i, \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu})^2. \quad (6)$$

The estimator given in Eq. (6) maximizes both Eqs. (4) and (5) in the unrestricted parameter space Θ . Assume now we are interested in testing the null hypothesis $\mathcal{H}_0: \mu = 0$. The parameter space under $\mathcal{H}_0$ is $\Theta_0 = \mathbb{R}_+$, and we now need to estimate σ^2 under the restriction $\mu = 0$, which amounts to computing

$$\hat{\sigma}_0^2 = \frac{1}{n} \sum_{i=1}^n X_i^2.$$

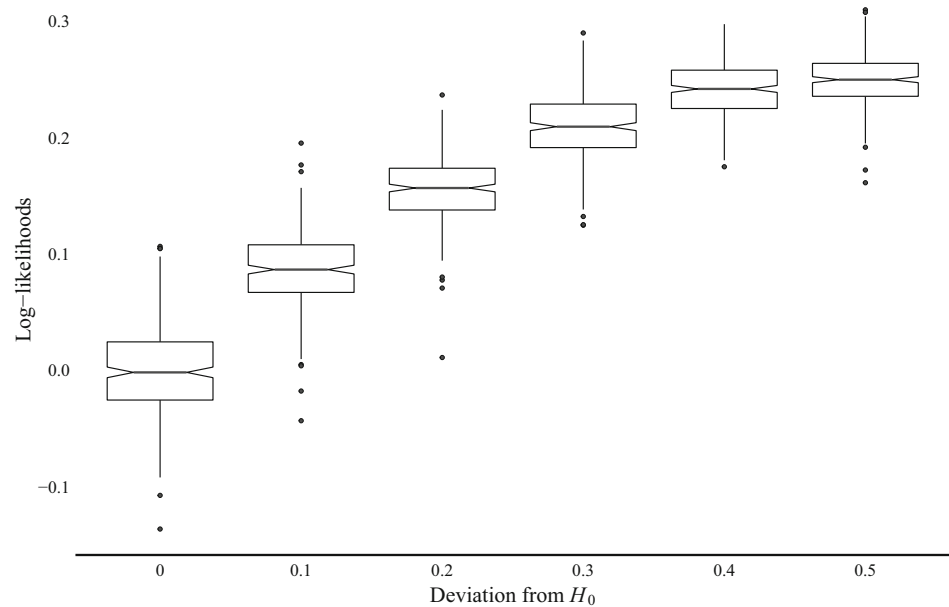
We may now compare the models with the difference of the log-likelihoods:

$$R_n(\mathbf{X}) = \ell(\hat{\mu}, \hat{\sigma}^2; \mathbf{X}) - \ell(\hat{\sigma}_0^2; \mathbf{X}).$$

Figure 2 shows the boxplots of 300 independent replications for each situation of this test statistics. The situations here considered are $n = 100$, $\mathcal{H}_0$, i.e., $N(0, 1)$ samples, and deviations from $\mathcal{H}_0$ which consist of sampling from $N(\Delta, 1)$, and $\Delta \in \{1/10, 2/10, 3/10, 4/10, 5/10\}$. Notice that the log-likelihoods ratio samples vary around zero, while central value of the others increases with Δ .

Log-Likelihood Ratio Test,

Fig. 2 R_n samples boxplots under $\mathcal{H}_0$ for $\Delta \in \{0, 1/10, 2/10, 3/10, 4/10, 5/10\}$

**Asymptotic Distribution**

The result that makes the likelihoods ratio test widely usable is the following. Consider the test statistic

$$\Lambda = 2 \log \frac{L(\hat{\theta}; \mathbf{X})}{L(\hat{\theta}_0; \mathbf{X})}.$$

Wilks (1938) proved that, under mild conditions and under $\mathcal{H}_0$, this test statistic has asymptotically a χ_r^2 distribution with r degrees of freedom, in which r is the difference of dimensions between Θ and Θ_0 .

Summary

Likelihoods ratio tests are very general and useful for contrasting a variety of hypotheses.

Bibliography

- Casella G, Berger RL (2002) Statistical inference, 2nd edn. Duxbury Press, Pacific Grove
 Lehman EL, Romano JP (2005) Testing statistical hypothesis, 3rd edn. Springer, New York
 Wilks SS (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann Math Stat 9(1):60–62. <https://doi.org/10.1214/aoms/1177732360>

Lognormal Distribution

Glòria Mateu-Figueras¹ and Ricardo A. Olea²

¹Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Girona, Spain

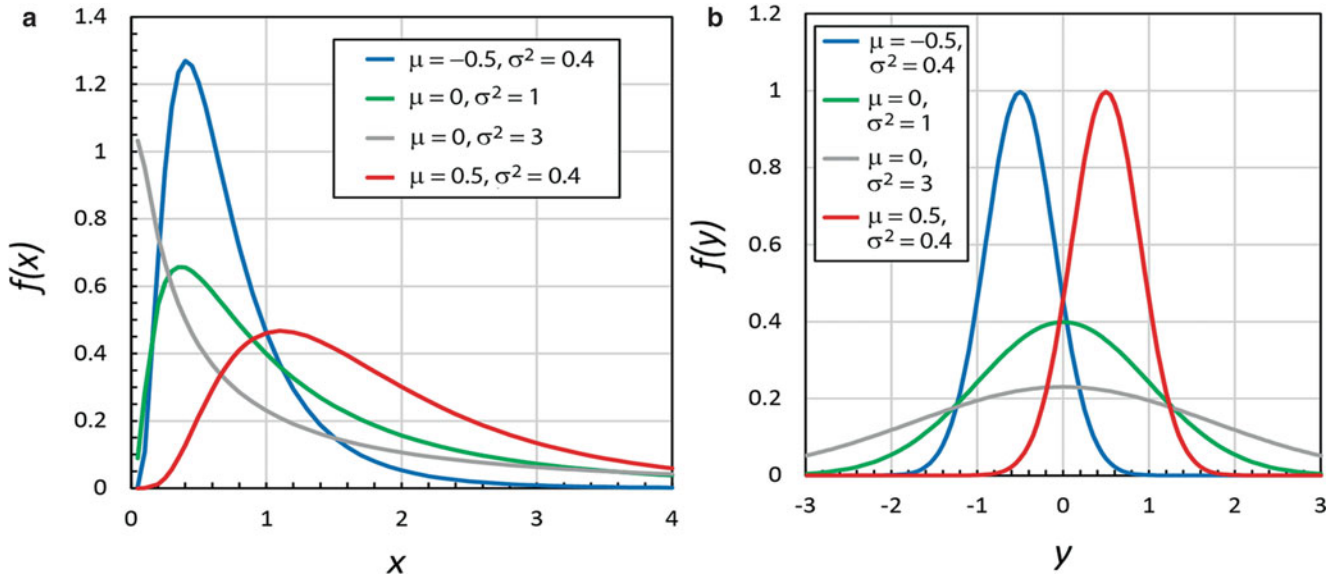
²Geology, Energy and Minerals Science Center, U.S. Geological Survey, Reston, VA, USA

Definition

A positive random variable X is lognormally distributed with two parameters μ and σ^2 if $Y = \ln(X)$ follows a normal distribution with mean μ and variance σ^2 . The lognormal distribution is denoted by $\Lambda(\mu, \sigma^2)$ and its probability density function is (Everitt and Skrondal 2010):

$$f(x) = \frac{1}{x \cdot \sigma \cdot \sqrt{2 \cdot \pi}} \cdot \exp \left[-\frac{(\ln x - \mu)^2}{2 \cdot \sigma^2} \right], 0 < x < \infty. \quad (1)$$

Note that the sampling space of X is the positive side of the real axis. By construction, the main property of the lognormal distribution is that $\ln(X)$ follows a normal distribution with mean μ and variance σ^2 , and, conversely, the exponential form $\exp(Y)$ of any normal distribution is a lognormal distribution (Fig. 1). Note that in all cases the lognormal density (Fig. 1a) is unimodal and positively skewed.



Lognormal Distribution, Fig. 1 (a) linear scale; (b) logarithmic scale

Historical Remarks and Properties

The normal distribution has long been associated with the modeling of additive errors. The roots of the formulation of the lognormal distribution go back to 1879 to the efforts of Francis Galton and Donald McAlister finding an adequate model for studying multiplicative errors (Galton 1879; McAlister 1879). In fact, given $X_1, X_2, \dots, X_n$ independent and strictly positive random variables, the product

$$\prod_{j=1}^n X_j \quad (2)$$

tends to a lognormal distribution as n tends to infinity, as confirmed by the central limit theorem applied to the respective logarithms.

Later, different properties in the theory of probability were used to explain the genesis of the lognormal distribution. The most important are the Law of Proportionate Effect and the Theory of Breakage. The Law of Proportionate effect can be formulated as (Gibrat 1930, 1931)

$$X_j = X_{j-1}(1 + \varepsilon_j), j = 1, \dots, n, \quad (3)$$

where $X_0, X_1, \dots, X_n$ is a sequence of random variables and $\varepsilon_1, \dots, \varepsilon_n$ are mutually independent and identically distributed random variables, also statistically independent of the set $\{X_j\}$. Note that, at the j th step, the change in the variate is a random proportion of the value X_{j-1} . From Eq. 3 we obtain

$$X_n = X_0 \prod_{j=1}^n (1 + \varepsilon_j). \quad (4)$$

Assuming that the absolute value of ε_j is small compared with 1, and using the Taylor expansion of $\ln(1 + x)$, we obtain

$$\ln(X_n) = \ln(X_0) + \sum_{j=1}^n \varepsilon_j. \quad (5)$$

Finally, using the central limit theorem, we conclude that $\ln(X_n)$ is asymptotically normally distributed and, consequently, X_n is asymptotically lognormally distributed.

The Theory of Breakage, used in particle size statistics, is essentially an inverse application of the Theory of Proportionate Effect (Crow and Shimizu 1988). In fact, consider that X_0 is the initial mass of a particle subject to a breakage process, that is, successive independent subdivisions. Let X_j be the mass of the particle in the j th step. Then, it can be proven that the distribution of X_n is asymptotically lognormal. This theory is very similar to the theory of classification, where the number of items classified in some homogeneity groups can be approximately lognormal distributed.

As can be observed in Fig. 1, it seems that the distribution has two shapes. For large variances the density function shows an exponential decline. Otherwise, the density function is peaked and asymmetric. In fact, the density function (Eq. 1) has two points of inflection at $\exp(\mu - \frac{3}{2}\sigma^2 \pm \frac{1}{2}\sigma\sqrt{\sigma^2 + 4})$ (Johnson et al. 1994), but for large values of the parameter σ^2 these points are too close to the origin to be observed separately.

Also, the lognormal model has been applied in an empirical way to fit many variables. So, there are both theoretical and empirical arguments in support of the existence of natural attributes and phenomena following the lognormal distribution. Aitchison and Brown (1957) and Krige (1978) also made valuable contributions to the use and understanding of the lognormal distribution.

Two characteristics of the lognormal distribution that have made it a useful tool in the earth sciences, economics, and population growth are the impossibility of taking negative values and its long upper tail (positive skewness). In the earth sciences, it is used to model the metal content of rock chips in a mineral deposit, the size of oil and gas pools, and in the lognormal kriging method for local estimation. The reader can find a large number of examples and references in Crow and Shimizu (1988), Aitchison and Brown (1957), or Johnson et al. (1994).

Statistics

Given $X \sim \Lambda(\mu, \sigma^2)$, the r -th moment of X about the origin is

$$\exp\left(r\mu + \frac{1}{2}r^2\sigma^2\right) \quad (6)$$

obtained using the properties of the moment generating function of the normal distribution, since $Y = \ln(X)$. From Eq. 6, we can obtain the mean and the variance as

$$\text{Mean} = \exp\left(\mu + \frac{1}{2}\sigma^2\right) \quad (7)$$

$$\text{Variance} = \exp(2 \cdot \mu + \sigma^2) \cdot (\exp(\sigma^2) - 1). \quad (8)$$

The shape factors, the coefficient of skewness, and the coefficient of kurtosis, are

$$\text{Skewness} = (\exp(\sigma^2) + 2) \cdot (\exp(\sigma^2) - 1)^{1/2} \quad (9)$$

$$\begin{aligned} \text{Kurtosis} &= 3 \cdot \exp(2 \cdot \sigma^2) + 2 \cdot \exp(3 \cdot \sigma^2) \\ &+ \exp(4 \cdot \sigma^2) - 3. \end{aligned} \quad (10)$$

Eq. 10 is zero when the distribution follows a normal distribution. From the expressions Eq. 9 and Eq. 10, it is clear that the skewness and the kurtosis increase with the value of σ^2 .

Other measures of interest are

$$\text{Median} = \exp(\mu) \quad (11)$$

$$\text{Mode} = \exp(\mu - \sigma^2). \quad (12)$$

As can be observed from the expressions, the relative position of the mean (Eq. 7), median (Eq. 11), and mode (Eq. 12) is $\text{Mode} < \text{Median} < \text{Mean}$, emphasizing again the positive skewness of the distribution.

Given $X \sim \Lambda(\mu, \sigma^2)$, the properties of the logarithmic function and the properties of the normal distribution allow proving that $(1/X) \sim \Lambda(-\mu, \sigma^2)$, $cX \sim \Lambda(\mu + \ln c, \sigma^2)$ for $c > 0$ and $X^a \sim \Lambda(a\mu, a^2\sigma^2)$ for $a \neq 0$.

Estimation

Given a sample from a $\Lambda(\mu, \sigma^2)$ distribution, the estimation of the parameters μ and σ^2 is relatively straightforward. As the logarithm of the variable is normally distributed, one must take logarithms of all values and proceed to estimate parameters μ and σ^2 using the well-known estimators for the mean and the variance of any sample. The same is true if we want to obtain exact confidence intervals for parameters μ and σ^2 . The main statistics of the lognormal distribution can be obtained by applying Eqs. 7–12.

Certainly, the mean and the variance of a lognormal random variable can be estimated applying directly the conventional estimators available for samples, that is, without making any logarithmic transformation. However, it can be proved that such estimators and their confidence intervals may be inefficient and suboptimal. In the literature we can find a large number of works seeking the best estimators. Finney (1941) proposes an unbiased minimum variance estimator for the mean using the Finney function. This solution is difficult to apply because it involves an infinite series, which Clark and Harper (2000) tabulated in an effort to simplify the application. Taraldsen (2005) derives a modified maximum likelihood estimator which approximates the Finney approach. Taraldsen (2005) and Tang (2014) compare several methods to estimate the mean of the lognormal distribution offering some numerical examples and simulations for appreciating the differences. Ginos (2009) does it for both the mean and the variance.

Other Forms of the Distribution

Mateu-Figueras and Pawlowsky-Glahn (2008) have proposed to consider the multiplicative structure of the positive real line and work with the density function with respect to the resulting multiplicative measure. The result is a density different from Eq. 1 but the same law of probability is obtained. Some interesting differences are obtained in the moments providing a justification for using the geometric mean to represent the first moment. This approach was implicit in the original definition of the lognormal distribution by McAlister (1879).

Two other expanded forms of the lognormal distribution have been postulated for accommodating special needs. In the three-parameter lognormal distribution, a value θ is subtracted from Eq. 1, so that now the sampling domain is shifted, starting at θ :

$$f(x) = \frac{1}{(x - \theta) \cdot \sigma \cdot \sqrt{2 \cdot \pi}} \cdot \exp \left[-\frac{(\ln(x - \theta) - \mu)^2}{2 \cdot \sigma^2} \right], \theta < x < \infty. \quad (13)$$

The three-parameter lognormal has been extensively studied by Yuan (1933). The addition of the shift parameter θ gives the model more flexibility but increases the difficulties of estimating parameters. It is an important model in many scientific disciplines. For example, in hydrology, it is used to model seasonal flow volumes, rainfall intensity or soil water retention (Singh 1998).

The other expanded form is the truncated lognormal distribution (Zaninetti 2017):

$$f(x) = \frac{\sqrt{2} \cdot \exp \left[-\frac{(\ln x - \mu)^2}{2 \cdot \sigma^2} \right]}{-\sqrt{\pi} \cdot \sigma \cdot x \cdot \left(\operatorname{erf} \left(\frac{\ln x_l - \mu}{\sqrt{2} \cdot \sigma} \right) - \operatorname{erf} \left(\frac{\ln x_u - \mu}{\sqrt{2} \cdot \sigma} \right) \right)}, x_l < x < x_u, \quad (14)$$

where x_l and x_u are the lower and upper truncation boundaries, respectively, and erf is the function:

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \cdot \int_0^x \exp(-t^2) dt. \quad (15)$$

The truncated lognormal distribution has been used in the assessment of the sizes of undiscovered oil and gas accumulations (Attanasi and Charpentier 2002).

If there are problems with the $(0, +\infty)$ boundaries of the lognormal distribution, in general it is better to work with other distributions with fixed and flexible boundaries, such as the beta distribution (Olea 2011).

The Multivariate Lognormal Distribution

Although not so popular as the univariate distribution, we have the multivariate version of the lognormal distribution. Given $\mathbf{Y} = (Y_1, \dots, Y_n) = (\ln(X_1), \dots, \ln(X_n))$ a multivariate normal distribution with parameters vector $\boldsymbol{\mu}$ and matrix $\boldsymbol{\Sigma}$, the distribution of the positive random vector $\mathbf{X} = (X_1, \dots, X_n)$ is multivariate lognormal and its density function is

$$f(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|^{1/2} x_1 \cdots x_n} \exp \left[-\frac{1}{2} (\ln \mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\ln \mathbf{x} - \boldsymbol{\mu}) \right], \mathbf{x} \in \mathbb{R}_+^n. \quad (16)$$

Summary and Conclusions

The lognormal distribution matches the tendency of numerous geological attributes that are positively defined and positively skewed, thus highlighting its importance in earth sciences modeling. Also, it is a distribution that has applications supported by theoretical arguments. It is a congenial model due to the relationship with the normal distribution. A large number of publications deal with the problem to directly find consistent estimators and exact confidence intervals for the mean and the variance.

Cross-References

- Normal Distribution
- Probability Density Function
- Random Variable
- Variance

Bibliography

- Aitchison J, Brown JAC (1957) The lognormal distribution. Cambridge University Press, Cambridge, p 176
- Attanasi ED, Charpentier RR (2002) Comparison of two probability distributions used to model sizes of undiscovered oil and gas accumulations: does the tail wag the assessment? *Math Geol* 34(6): 767–777
- Clark I, Harper WV (2000) Practical geostatistics 2000. Ecosse North America Llc, Columbus, p 342
- Crow EL, Shimizu K (1988) Lognormal distributions. Theory and Applications. Marcel Dekker, Inc, New York. 387 pp
- Everitt BS, Skrondal A (2010) The Cambridge dictionary of statistics, 4th edn. Cambridge University Press, Cambridge. 480 pp
- Finney DJ (1941) On the distribution of a variate whose logarithm is normally distributed. *J R Stat Soc Ser B* 7:155–161
- Galton F (1879) The geometric mean, in vital and social statistics. *Proc R Soc Lond* 29:365–366
- Gibrat R (1930) Une loi des répartitions économiques: l'effet proportionnel. *Bull Statist Gén Fr* 19. 460 pp
- Gibrat R (1931) Les Inégalités Économiques. Librairie du Recueil Sirey, Paris
- Ginos BF (2009) Parameter estimation for the lognormal distribution. Master thesis, Brigham Young University, <https://scholarsarchive.byu.edu/etd/1928>
- Johnson NL, Kotz S, Balakrishnan N (1994) Continuous univariate distributions, Wiley series in probability and mathematical statistics: applied probability and statistics, vol 1, 2nd edn. Wiley, New York. 756 pp
- Krige DG (1978) Lognormal-de Wijsian geostatistics for ore evaluation. South African Institute of Mining and Metallurgy; second revised edition, 50 pp

- Mateu-Figueras G, Pawlowsky-Glahn V (2008) A critical approach to probability laws in geochemistry. *Math Geosci* 40(5):489–502
- McAlister D (1879) The law of geometric mean. *Proc R Soc Lond* 29: 367–376
- Olea RA (2011) On the Use of the Beta Distribution in Probabilistic Resource Assessments. *Nat Resour Res* 20(4):377–388
- Singh VP (1998) Three-Parameter Lognormal Distribution. In: *Entropy-Based Parameter Estimation in Hydrology, Water Science and Technology Library*, vol 30. Springer, Dordrecht, 82–107
- Tang Q (2014) Comparison of different methods for estimating log-normal means. Master thesis, East Tennessee State University, <https://dc.etsu.edu/cgi/viewcontent.cgi?article=3728&context=etd>
- Taraldsen G (2005) A precise estimator for the log-normal mean. *Statistical Methodology* 2(2):111–120
- Yuan P (1933) On the logarithmic frequency distribution and semi-logarithmic correlation surface. *Ann Math Stat* 4:30–74
- Zaninetti L (2017) A left and right truncated lognormal distribution for the stars. *Adv Astrophys* 2(3):197–213

Lorenz, Edward Norton

Kerry Emanuel
Lorenz Center, Massachusetts Institute of Technology,
Cambridge, MA, USA

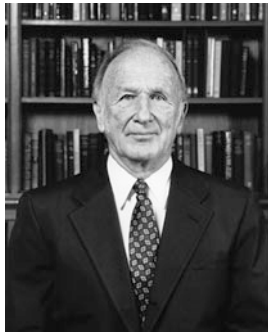


Fig. 1 Edward Norton Lorenz (1917–2008), courtesy of Prof. Kerry Emanuel

Biography

Edward Norton Lorenz, a mathematician and meteorologist, is best known for having demonstrated that the macroscopic universe is almost certainly not predictable beyond fundamental time horizons, putting an end to the notion of a clockwork universe. His work had the most immediate application to weather prediction but has also had a strong influence on mathematics and physics.

This is a condensed version of a more extensive biography of Lorenz, which the author published with the National Academy of Sciences (Emanuel 2011)

Lorenz was born in West Hartford, Connecticut, on 23 May 1917, to Edward Henry Lorenz, a mechanical engineer, and Grace Lorenz (née Norton), a school teacher. As a boy he enjoyed solving mathematical puzzles with his father and from his mother learned to love board and card games.

Lorenz entered Dartmouth College as a mathematics major in 1934. In 1938 he entered the graduate school of the mathematics department at Harvard, but his mathematical training was interrupted by World War II. He contributed to the war effort by training in meteorology at the Massachusetts Institute of Technology (MIT), and becoming a weather forecaster for the Army Air Corps (now the Air Force) operating out of Saipan, and later Okinawa.

After the war, Lorenz decided to pursue academic studies of meteorology rather than mathematics, completing his Ph.D. work at MIT in 1948 and being appointed to the MIT faculty in the mid-1950s. He became interested in the problem of forecasting the weather by using a digital computer to integrate the known equations governing the behavior of fluids like the atmosphere. At the time, a group of statisticians proposed that anything that could be forecast by integrating equations could be forecast at least as well using statistical techniques. Lorenz intuitively felt that was wrong and set out to prove it. By the early 1960s he had found a set of very simple equations that demonstrated certain properties that were then unknown in mathematics, including extreme sensitivity of forecasts to the specification of the initial state of the system. These results were published in a meteorological journal (Lorenz 1963), which later made him famous as the father of chaos theory.

Arguably, his most important contribution came in a paper published 6 years later (Lorenz 1969) in which he demonstrated, using a particular mathematical system representing turbulent flows, that the solutions could not be predicted beyond a certain definite time no matter how small one made the errors in the description of the initial state. When married with Irwin Schrodinger's uncertainty principle, this implies that there exist systems that fundamentally cannot be predicted beyond a certain time horizon.

Lorenz died in April, 2008, having made many important contributions to atmospheric science in addition to his fundamental work on chaos theory.

Bibliography

- Emanuel K (2011) Edward Norton Lorenz, 1917–2008. In: *Biographical memoir*, vol 28. National Academy of Sciences, Washington, DC
- Lorenz EN (1963) Deterministic nonperiodic flow. *J Atmos Sci* 20: 130–141
- Lorenz EN (1969) The predictability of a flow which possesses many scales of motion. *Tellus* 21:289–307. <https://doi.org/10.3402/tellusa.v21i3.10086>

Loss Function

Tanujit Chakraborty and Uttam Kumar
Spatial Computing Laboratory, Center for Data Sciences,
International Institute of Information Technology Bangalore
(IIITB), Bangalore, Karnataka, India

Definition

Over the past few decades artificial intelligence has paved the way for automating routine labor, understanding speech and images, tackling the geosciences and remote sensing problems, supporting basic scientific research through intelligent software, etc. Majority of the tasks are performed through learning from data which is popularly known as machine learning. The generalization performance of a learning method is its prediction capability on independent test data. While assessing the performance of the learning methods, one desires to achieve two separate goals: (a) model selection – estimating the performance of different models in order to choose the best one; (b) model assessment – after selecting a final model, estimating its prediction or generalization error on a new dataset (Hastie et al. 2009). Loss function acts as a building block during this learning phase and enables the algorithm to model the data with higher accuracy. Loss function (also known as cost function or an error function) specifies a penalty for an incorrect estimate from a statistical or machine learning model. Typical loss functions might specify the penalty as a function of the difference between the estimate and the true value, or simply as a binary value depending on whether the estimate is accurate within a certain range. Any loss criterion in classification actually penalizes negative margins more heavily than positive ones as positive observations have already been correctly predicted by the model.

Introduction

Learning and improving upon past experiences is the key step in the arena of research of data science and artificial intelligence. In order to understand how a machine learning algorithm learns from data to predict an outcome, it is essential to understand the underlying concepts involved in training an algorithm. The data on which a learning algorithm is trained can be broadly classified as labeled and unlabeled data and the learning is termed as supervised learning and unsupervised learning accordingly. During the training phase, an algorithm explores the data and identifies patterns for determining good values for all the weights and the bias from labeled examples. In supervised learning, a machine learning algorithm builds a model by examining many examples and attempting to find a model that minimizes loss; this process is called empirical risk

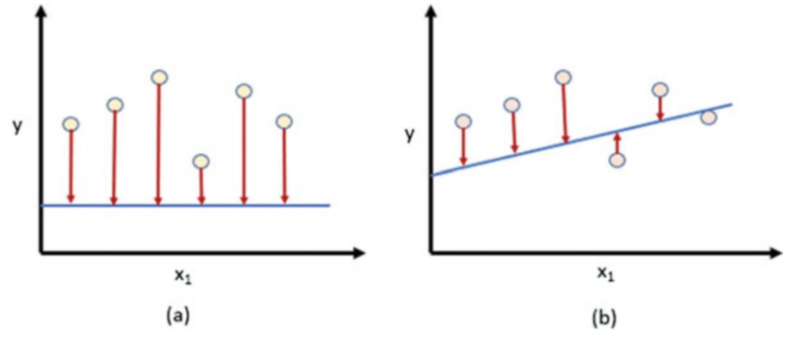
minimization (Raschka and Mirjalili 2019). Loss functions are related to model accuracy. If the current output of the algorithm deviates too much from the expected ones, loss function would stump up a very large number. If they are pretty good, it will output a lower number. Gradually, with the help of some optimization functions, loss function learns to reduce the error in prediction (Bishop and Nasrabadi 2006). In Bayesian decision theory, loss function states how costly could be - the use of more than one feature, allowing more than two states of nature for generalization for a small notable expense, allowing actions other than classification facilitating rejection, etc. It is used to convert a probability determination into a decision (Duda et al. 2000).

What Is Loss Function?

We shall begin with a quote by François Chollet, “A way to measure whether the algorithm is doing a good job — this is necessary to determine the distance between the algorithm’s current output and its expected output. The measurement is used as a feedback signal to adjust the way the algorithm works. This adjustment step is what we call learning” (Chollet 2021).

During the training phase of any supervised learning algorithm, the data-modeler computes the error function for each individual instances by calculating the difference between the actual value and the predicted value. In the next step, the loss function is evaluated as the average error over all training data points. If the model’s prediction is perfect, the loss is zero; otherwise, the loss is greater. The goal of training a model is to find a set of weights and biases that have low loss, on average, across all examples. We shall visualize the loss for a regression problem where we wish to predict the price of an apartment in New York City. We consider two models and calculate the losses for each of them as depicted in Fig. 1. Here the predictions are represented by blue lines and the red arrows indicate the loss. Depending on the problem, the loss function may be directionally agonistic, i.e., it doesn’t matter whether the loss is positive or negative. All that matters is the magnitude. This is not a feature of all loss functions; in fact, the loss function will vary significantly based on the domain and unique context of the problem that we are applying machine learning algorithms to. In some situations, it may be much worse to guess too high than to guess too low, and the loss function we select must reflect that. The machine learning problem of classifier design is usually studied from the perspective of probability elicitation in statistical learning. This shows that the standard approach of proceeding from the specification of a loss to the minimization of conditional risk is overly restrictive. Earlier literature on loss functions is highly biased toward convex loss functions. Recently, various new loss functions have been derived in the literature which

Loss Function, Fig. 1 Plot of predicted value (blue), actual output (shaded circle) and loss for each example (red arrows)



are not convex and robust to the contamination of data with outliers, for details see (Vortmeyer-Kley et al. 2021).

Different Kind of Loss Functions

There is no “one-size-fits-all loss function” in machine learning algorithms. There are various factors that influence the choice of loss function – one of the major influential factors being the type of problem under consideration. Based on that, loss function can be broadly classified as follows:

Regression Loss Function

Regression is a supervised learning method that is used to estimate the relationship between a continuous dependent variable and a set of independent variables. In this section, we will focus on some common loss functions used for the prediction of dependent variable (Hastie et al. 2009). The mathematical expressions of these loss functions are stated using the notations where Y_i indicates the target value, $h(x_i)$ gives the predicted output, and n is the total number of examples.

Mean Absolute Error (MAE) or L1 Loss Function

L1 loss function minimizes the absolute differences between the estimated values and the existing target values. Mathematically, the mean absolute error (MAE) is given by

$$MAE = \frac{1}{n} \sum_{i=1}^n |Y_i - h(x_i)|.$$

One of the major drawbacks of MAE is that it is a scale-dependent measure hence one cannot obtain the relative size of the loss from this measure. To overcome this problem, Mean Absolute Percentage Error (MAPE) was proposed to compare forecasts of different series in different scales. Mathematically, MAPE is denoted as follows:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{Y_i - h(x_i)}{Y_i} \right|.$$

Mean Squared Error (MSE) or L2 Loss Function

Mean Squared Error (MSE) is the workhorse of basic loss functions; it's easy to understand, implement, and generally works pretty well. To calculate MSE, we take the difference between model predictions and the ground truth, square it, and average it out across the whole dataset. The L2 loss function is computed as:

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - h(x_i))^2.$$

A special case of MSE is Root Mean Square Error (RMSE). It is computed as the square root of the MSE and is given by:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - h(x_i))^2}.$$

In Fig. 2 we have plotted the MAE and MSE loss values for each individual predictions in the interval $(-10, 10)$ where the actual output is assumed to be 0.

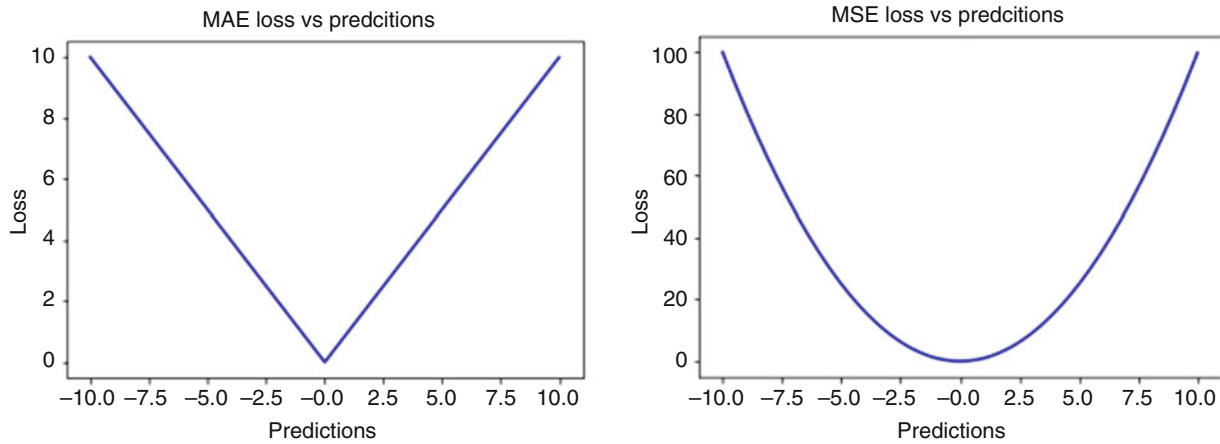
Log-Cosh Loss

Log-cosh is another loss function used in regression tasks that is smoother than L2 loss function (as evident from Fig. 3). It is the logarithm of the hyperbolic cosine of the prediction error. The mathematical formula is given by:

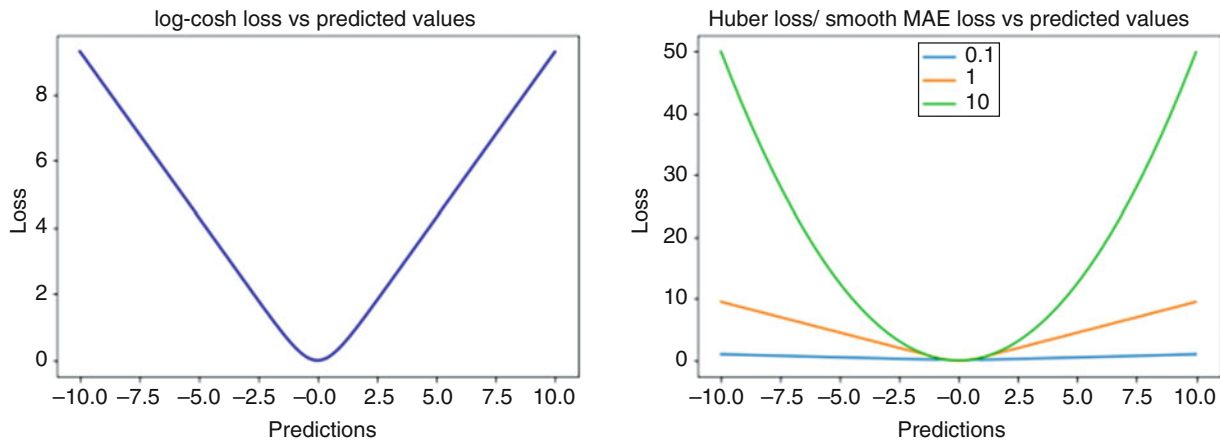
$$\log - \cosh = \sum_{i=1}^n \log(\cosh(h(x_i) - Y_i)).$$

Huber Loss or Smooth Mean Absolute Error

A combination of L1 & L2 loss functions accompanied with an additional parameter delta (δ) gives us the Huber loss. The shape of the loss function is influenced by this parameter (as demonstrated in Fig. 3). So, the delta parameter is fine-tuned by the algorithm. When the predicted values are far from the original, the function has the behavior of the MAE,



Loss Function, Fig. 2 Plot of MAE (left) and MSE (right) loss functions



Loss Function, Fig. 3 Plot of log-cosh (left) and Huber (right) loss functions

closed to the actual values, the function behaves like the MSE. The mathematical form of the Huber Loss is given by (Hastie et al. 2009):

$$L_{\delta}(y, h(x)) = \begin{cases} \frac{1}{2} (y - h(x))^2 & \text{for } |y - h(x)| \leq \delta, \\ \delta |y - h(x)| - \frac{1}{2} \delta^2 & \text{otherwise.} \end{cases}$$

Quantile Loss

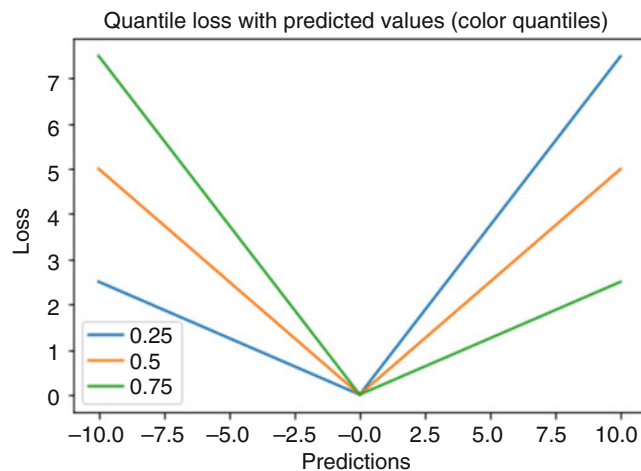
The functions discussed so far give us a point estimate of the loss, whereas in certain real-world prediction problems, interval estimate of the loss function can significantly improve decision-making processes. Quantile loss functions turn out to be useful when we are interested in predicting an interval instead of only point predictions. Quantile-based regression aims to estimate the conditional “quantile” (of order γ) of a response variable given certain values of predictor variables. Quantile loss is an extension of MAE (when the quantile is 50th percentile, it is MAE). It is computed as follows:

$$L_{\gamma}(y, h(x)) = \sum_{i=y_i < h(x_i)} (\gamma - 1) \cdot |y_i - h(x_i)| + \sum_{i=y_i \geq h(x_i)} (\gamma) \cdot |y_i - h(x_i)|.$$

The value of the quantile to be used depends on the formulation of the problem, i.e., whether we wish to penalize overestimation or underestimation more severely. For example, a quantile loss function of $\gamma = 0.25$ gives more penalty to overestimation and tries to keep prediction values a little below median. Figure 4 exhibits the shape of the quantile loss for various values of quantiles.

Some of the observations from the regression loss functions discussed above are as follows:

- MAE is more robust to outliers, whereas MSE is a differentiable function and hence more widely used than its MAE counterpart. However, both MAE and MSE penalizes the large prediction errors.



Loss Function, Fig. 4 Plot of quantile loss function

- Huber loss function is less sensitive to outliers than the MSE as it treats error as square only inside an interval.
- The predictions are little sensitive to the value of the hyperparameter chosen in case of the model with Huber loss.
- The quantile loss gives a good estimation of the corresponding confidence levels.

Classification Loss Function

Classification is a predictive learning problem where a class label is predicted for a given set of input data. The loss functions that are commonly used to compute distances between prediction and categorical target variables in classification problems are given below (Hastie et al. 2009).

Cross Entropy

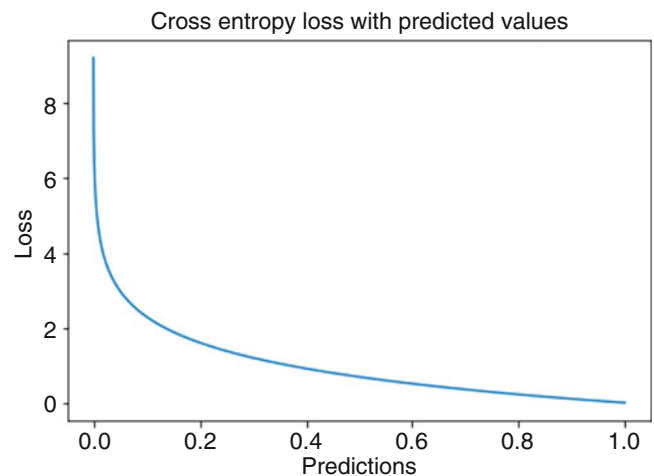
Cross-entropy compares the distance (or deviance) between predictions and the actual values. Cross entropy loss increases when the predicted value diverges from the actual. This function can be used for binary and multiclass classification problems. The cross-entropy loss is ubiquitous in modern deep neural networks. Mathematically, it is computed as follows:

Cross entropy loss

$$= \begin{cases} -(y \log(p)) - (1 - y) \log(1 - p) & \text{for } M = 2, \\ -\sum_{c=1}^M y_{o,c} \log(p_{o,c}) & \text{for } M > 2, \end{cases}$$

where M = number of classes, $\log$ = the natural logarithm, y = indicator if the class label c is the correct classification for the observation o , p = predicted probability for observation o that it belongs to class c .

There are several types of cross entropy loss functions in literature. Among them, binary cross entropy or log loss is



Loss Function, Fig. 5 Plot of cross entropy loss function

used only for binary classification problems (commonly used in logistic regression) where the target can take either of the two values 0 or 1. It measures the distance from the actual class (0 or 1) to the model's prediction, which is a probability (i.e., value in $[0-1]$). The shape of the loss function (as in Fig. 5) is convex and grows linearly for negative values (less sensitive to outliers). For any given problem, a lower log loss value indicates better prediction. Log loss function can be interpreted as the negative average of the log of the predicted probabilities. On the other hand, categorical cross entropy can be used for both binary and multiclass problems having two or more labels, the label needs to be encoded as categorical, one-hot encoding representation. Sparse categorical cross-entropy is used for both binary and multiclass problem. The number of floating-point values per feature should match the number of classes in the prediction and a single floating-point value should be present per feature for the target variable.

Logistic Loss

Logistic loss is a variant of the binary cross entropy up to a constant multiplier $1/\log(2)$. Similar to the log loss function, logistic loss is also less sensitive to outliers. It is widely used for optimizing LogitBoost algorithm (Friedman et al. 2000).

Exponential Loss

The exponential loss is a convex function and grows exponentially for negative values which makes it more sensitive to outliers. The exponential loss, widely used in optimizing AdaBoost algorithm, is defined as follows:

$$\text{Exponential loss} = \frac{1}{N} \sum_{i=1}^N e^{-h(x_i) * y_i},$$

where N is the number of examples, $h(x_i)$ is the prediction for i^{th} example, and y_i is the target for the i^{th} example.

Hinge Loss

The hinge loss is used for “maximum-margin” classification, most notably for support vector machines (SVMs) (Rosasco et al. 2004). The goal is to make different penalties at the point that are not correctly predicted or are too close to the hyper-plane. Its mathematical formula is:

$$\text{Hinge loss} = \max(0, 1 - y * h(x)),$$

y is the target vector and $h(x)$ is the predicted vector. Hinge loss function can be used for both binary classification and multi-class classification.

Kullback-Leibler Divergence Loss

Kullback–Leibler divergence, D_{KL} (also called relative entropy), measures how one probability distribution is different from a reference probability distribution. Usually, P represents the data and the predictions are encoded as Q . The Kullback–Leibler divergence is interpreted as the information gain achieved if P would be used instead of Q which is currently used. The relative entropy from Q to P is denoted by $D_{KL}(P||Q)$ and is evaluated as

$$D_{KL}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right),$$

where P and Q are discrete probability distributions on the probability space $\mathcal{X}$. For P and Q being the distributions of continuous random variables $D_{KL}(P||Q)$ is defined as

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \left(\frac{p(x)}{q(x)} \right) dx,$$

where p and q are the probability densities of P and Q , respectively (MacKay and Mac Kay 2003). Kullback-Leibler divergence is not symmetric and does not follow triangle inequality.

Loss Function in Geoscience

Geoscience is a field of great societal relevance that require solutions to several urgent problems such as predicting impact of climate change, measuring air pollution, predicting increased risks to infrastructures by disasters, modeling future availability and consumption of water, food, and mineral resources, and identifying factors responsible for natural calamities. As the deluge of big data continues to impact practically every commercial and scientific domain, geoscience has also witnessed a major revolution from being a data-poor field to a data-rich field. This has been possible with the advent of better sensing technologies, improvements in computational resources, and internet-based democratization of

data. The growing availability of geoscience big data offers immense potential for machine learning practitioners to apply state-of-the-art methodologies to model this data (Samui 2008); (Toms et al. 2020). The popular types of algorithms that are used in these applications include artificial neural network, support vector machine, self-organizing map, decision trees, ensemble methods such as random forest, neuro fuzzy, multivariate adaptive regression splines, etc. These models use various loss functions in their training phase, however, the most commonly used loss functions in geoscience applications incorporates cross-entropy loss function and Huber loss function due to higher overall accuracy and better learning efficiency.

Summary

Loss functions provide more than just a static representation of how the model is performing – they are how the algorithms fit data in the first place. Most machine learning algorithms use some sort of loss function in the process of optimization or finding the best parameters (weights) for the data. Importantly, the choice of the loss function is directly related to the formulation of the problem. As such, absolute loss in regression is same as binomial deviance in classification; it increases linearly for extreme margins. Exponential loss is more undesirable than squared-error-loss. When considering robustness in spatial data analysis, squared error loss for regression and exponential loss for classification are not idle. However, they both lead to sophisticated boosting method in forward stagewise additive modeling setting. One must understand that none of the loss functions can be treated as universal, however, choosing a loss function is a difficult proposition and depends on the choice of machine learning model for the problem in hand. We can think of selecting the algorithm as a choice about framing of the prediction problem, and the choice of the loss function as the way to calculate the error for a given framing of the problem. A future scope of research on the area of loss function in geoscience would be to work with Wasserstein loss which has got much attention in machine learning domain for the last decade.

Data and Code

The data and code used in this study are available at <https://github.com/tanujit123/loss>.

Cross-References

- [Artificial Intelligence in the Earth Sciences](#)
- [Big Data](#)

- [Cluster Analysis and Classification](#)
- [Logistic Regression, Weights of Evidence, and the Modeling Assumption of Conditional Independence](#)
- [Machine Learning](#)
- [Optimization in Geosciences](#)
- [Pattern Classification](#)
- [Probability Density Function](#)
- [Regression](#)
- [Support Vector Machines](#)

Bibliography

- Bishop CM, Nasrabadi NM (2006) Pattern recognition and machine learning, vol 4, no 4. Springer, New York
- Chollet F (2021) Deep learning with Python. Simon and Schuster
- Duda RO, Hart PE, Stork DG (2000) Pattern classification, 2nd edn. A Wiley-Interscience Publication, New York, ISBN 9814-12-602-0
- Friedman, J, Hastie T, Tibshirani R (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann Stat* 28(2):337–407
- Hastie T, Tibshirani R, Friedman J, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, New York
- MacKay D, Mac Kay D (2003) Information theory, inference and learning algorithms, 1st edn. Cambridge University Press, s.l.
- Raschka S, Mirjalili V (2019) Python machine learning: machine learning and deep learning with Python, scikit-learn, and TensorFlow 2, 3rd edn. Packt Publishing Ltd
- Rosasco, L., De Vito, E., Caponnetto A, Piana M, Verri A (2004). Are loss functions all the same?. *Neural Comput* 16(5):1063–1076
- Samui P (2008) Support vector machine applied to settlement of shallow foundations on cohesionless soils. *Comput Geotech* 35(3):419–427
- Toms B, Barnes E, Ebert-Uphoff I (2020) Physically interpretable neural networks for the geosciences: applications to earth system variability. *J Adv Model Earth Syst* 12(9):e2019MS002002
- Vortmeyer-Kley R, Nieters P, Pipa G (2021) A trajectory-based loss function to learn missing terms in bifurcating dynamical systems. *Sci Rep* 11(1):1–13

Machine Learning

Feifei Pan

Rensselaer Polytechnic Institute, Troy, NY, USA

Definition

Machine learning is the science which allows humans to design algorithms and teach computers to learn patterns from vast amounts of data and use the patterns to automatically make decisions or predictions. In this process, data can be and not limited to the following types: numeric values, text, graph, photos, audio, and more. Machine learning is considered as a subset of artificial intelligence. It is closely related to computational statistics, data science and data mining and frequently applied to other research domains such as natural language processing, computer vision, robotics, bioinformatics, and so forth.

Introduction

Machine learning automates computers to make decisions or predictions without being explicitly programmed so. With the exponential growth of data, machine learning has become a crucial component behind many great inventions and services in the past decades. For instance, machine learning enables the accurate recommendation systems for online shopping websites such as Amazon and eBay, the effective search engines such as Google, Baidu, and Bing, the convenient voice assistant systems such as Siri and Alexa.

Machine learning has attracted considerable attention since the past century. As early as 1950, Alan Turing raised the questions, “Can machines think?”, in his paper *Computing machinery and intelligence* (Turing 2009). Later in 1959, the term machine learning was first introduced by Arthur

Samuel in his paper on computer gaming *Some studies in machine learning using the game of checkers* (Samuel 1959). Nilsson pioneered the research in machine learning with the book *Learning Machines* in the 1960s, which focused on pattern classification in the discriminant machine systems. Since then, substantial efforts have been put into machine learning research. With the publication of Hinton’s highly-cited paper *Learning representations by back-propagating errors* (Rumelhart et al. 1988) in late 1980s, machine learning research welcomed the era of deep learning.

In general, a learning process in machine learning can be described as the following: (1) the model input is denoted as x ; (2) the target, can be either a decision or a prediction, is denoted as Y ; (3) a dataset D is also provided, where $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, note that y is not always necessary for input data of semi-supervised and unsupervised learning models. The goal of a learning process is to learn a function $g: X \rightarrow Y$ that closely approximates the true relation between inputs and outputs $f: X \rightarrow Y$ (Abu-Mostafa et al. 2012). The exact target function $f: X \rightarrow Y$ usually cannot be learned due to the limitation of data. When encountering unseen data x_i , a Y_i can be automatically generated by computers using the learned function $g: X \rightarrow Y$. In most cases, the function $g: X \rightarrow Y$ is chosen from a set of hypothesis functions based on the characteristics of the data and the properties of the learning tasks.

Type of Learning

There are various types of learning designed for different types of input/output data and learning problems. The learning types can be roughly categorized into four major classes: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. Some other learning types include self-learning, feature learning, association rule learning, etc.

Supervised, Unsupervised, and Semi-Supervised Learning

Supervised learning is designed for labeled data, which means the input data are data pairs (x_i, y_i) with both input x_i and ground truth output y_i (usually manually labeled). The input data for a supervised model is split into two non-overlapping datasets: training data and testing data. A supervised learning model is built on training data pairs to learn the function $g: X \rightarrow Y$ discussed in section “Introduction.” The function g is later applied to the unseen testing data for evaluation. Supervised learning includes classification and regression algorithms, such as k -nearest neighbors, support vector machines, logistic regression, naive Bayes, decision trees, Stochastic Gradient Descent, etc. Active learning and online learning are two of the main variations of supervised learning (Abu-Mostafa et al. 2012).

In contrast to supervised learning, unsupervised learning (Hinton et al. 1999) involves unlabeled data. Hence, the input data for unsupervised models only contains x , and the goal of unsupervised learning is to directly learn patterns or structures among data. The primary approach of unsupervised learning is cluster analysis, an algorithm that groups unlabeled data into N clusters based on the underlying similarities. K-means, hierarchical clustering, mixture models, and density-based spatial clustering of applications with noise (DBSCAN) are well-known algorithms for clustering. The main advantage of unsupervised learning is that it requires a minimum amount of supervision and human efforts, especially for data annotation.

Semi-supervised learning, similar to its literal meaning, is a hybrid of supervised and unsupervised learning. The model input consists of a small amount of labeled data and a sizable amount of unlabeled data. While unsupervised learning tends to produce lower accuracy, semi-supervised learning introduces a small amount of labeled data for model tuning, which considerably enhances the model performance without significantly increase the annotation expense.

Reinforcement Learning

Reinforcement learning is also built on unlabeled data but works differently from unsupervised learning. It is most similar to how humans and animals learn basic skills. Imagining a puppy is given a treat every time he/she sits after the command “sit” and is given nothing if he/she does something else. With several repetitions, the puppy gradually learns the correct action to take after the “sit” command under this reward mechanism. In reinforcement learning, the computer is trained similarly: for each input training sample, the outcome is measured accordingly, and the final goal is to maximize the reward for all samples. Reinforcement learning is particularly useful in game theory, control theory, and operations research.

Learning Models

A standard machine learning process involves model building. Commonly seen models include simple models such as linear regression, support vector machine, decision tree, k -nearest neighbors, and more complex models such as neural networks.

Linear Models

The general idea of linear models is to learn a set of weights, also called coefficients, denoted as $w = (w_1, w_2, w_3, \dots, w_n)$ for every feature x_i of the input data X . The predicted value $\hat{y}$ is the linear combination of w and X , where:

$$\hat{y} = w_0 + w_1x_1 + w_2x_2 + \dots + w_nx_n$$

with w_0 denotes a constant.

The basic linear regression model aims to minimize the sum of squared errors between the predicted values and the ground truth. Mathematically speaking, the linear regression model learns a set of coefficients to minimize the cost function, denotes as:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \left(y_i - \sum_{j=0}^k x_{ij}w_j \right)^2$$

Lasso (least absolute shrinkage and selection operator) regression (Tibshirani 1996) is a variation from the basic linear model. It aims to minimize the sum of squared errors with a tuning parameter, also known as the $L1$ regulation. The cost function of lasso regression is:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \left(y_i - \sum_{j=0}^k x_{ij}w_j \right)^2 + \lambda \sum_{j=0}^k |w_j|$$

where λ is a constant and $\sum_{j=0}^k |w_j|$ represents the $L1$ norm. With the regulation, lasso regression automatically performs feature selection and reduces overfitting.

Other widely used linear models include logistic regression, multi-task lasso, elastic-net, Bayesian regression and more.

Support Vector Machine

The Support Vector Machine(SVM) model (Cortes and Vapnik 1995) is a set of supervised algorithms that construct multiple high-dimensional hyperplanes in vector spaces for regression and classification tasks. The SVM model is very effective for high-dimensional data, especially when the datasets are relatively small. It is also very flexible: SVM’s

decision functions vary with the selection of the kernel functions. Linear, polynomial, sigmoid and Radial Basis Function (RBF) are the frequently used kernel functions for SVM models. Customized kernel functions are also possible for SVM models.

Decision Tree

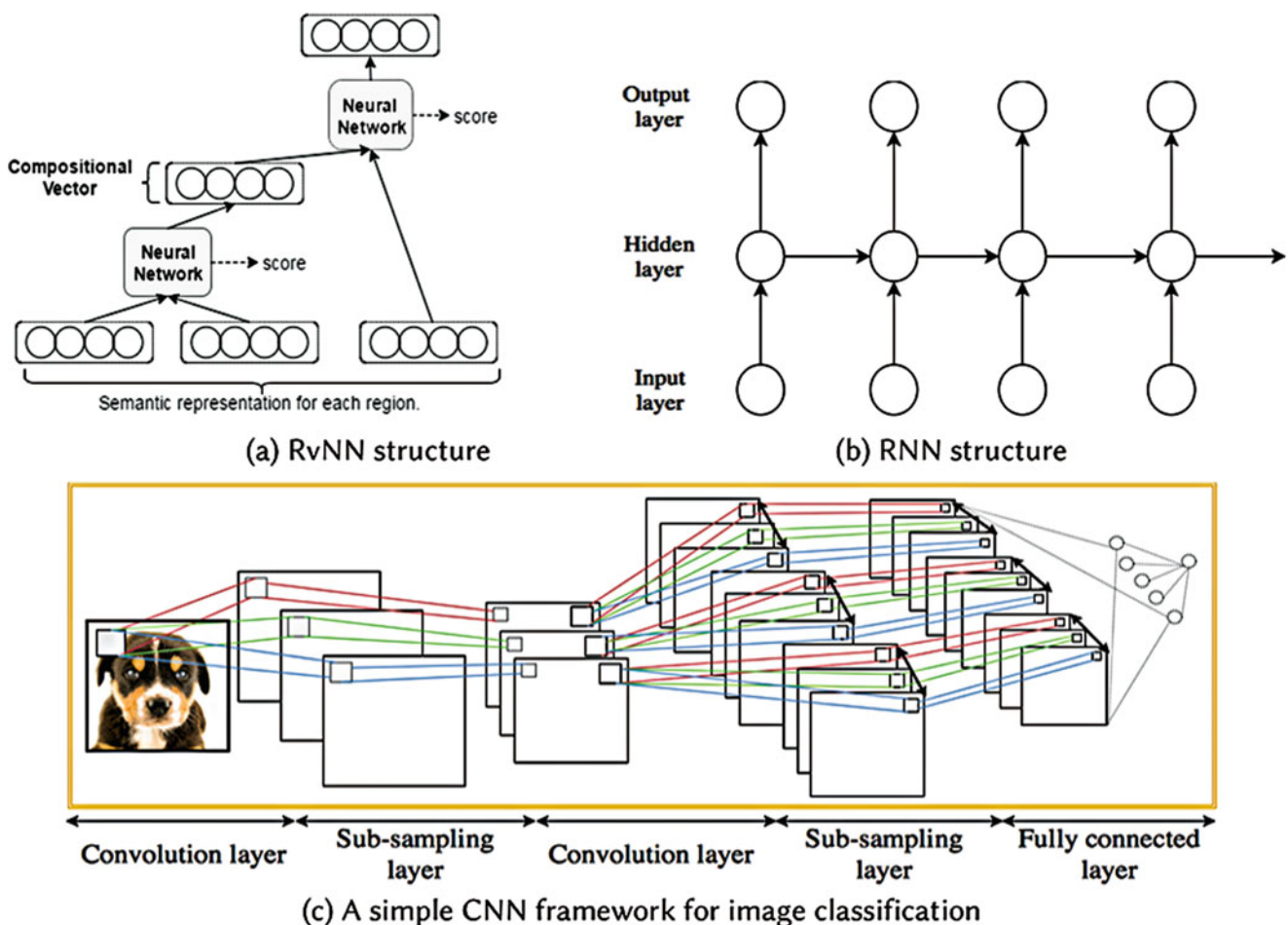
A decision tree model produces a flowchart-like, tree-structured hierarchy. Each branch represents a condition or a feature, and each bottom layer leaf node represents a class or a label. Decision tree learning is widely used in classification (Classification tree) and regression (Regression tree) tasks. To enhance the accuracy of decision tree learning, Ho Tin-kam proposed the random forest, or the random decision forests model, in her work *Random decision forests* in 1995 (Ho 1995). Random forest is an ensemble learning method that constructing multiple decision trees and outputting based on the predictions of all decision trees. Random forest model increases the accuracy of the decision tree algorithm and reduces overfitting.

Neural Networks

Neural networks, the foundation of deep learning, are learning models that designed to replicate the mechanism of human brains. A typical neural network model has an input layer, one or multiple connected, hidden layers and an output layer. Signals are transmitted through the neurons in each layer and the hidden layers can be organized and connected in different fashion, as shown in Fig. 1. Neural networks can be simple as multi-layer perceptrons and can also be complicated as RNN, CNN, and Long/Short Term Memory (LSTM) models. Neural network models can be used in supervised, unsupervised, semi-supervised, and reinforcement learning. It has a high requirement for computational power but is famous for its portability, robustness, and high accuracy.

Issues and Limitations

Overfitting is one of the most common issues for machine learning models. It occurs when a model only focuses on perfectly fitting the training data while loses generality and



Machine Learning, Fig. 1 Diagrams of three neural network architectures: Recursive Neural Network (RvNN), Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs). It is the Fig. 1 in

A Survey on Deep Learning: Algorithms, Techniques, and Applications (Pouyanfar et al. 2018)

causes low performance on unseen validation data. Overfitting is very likely to increase the computational cost and lower the portability of the models. However, overfitting can be avoided through properly-designed cross-validation.

Machine learning models highly rely on the quality of data. Lack of data, biased data, and inaccessible data are some of the significant issues that fail machine learning models. Training with low-quality data may cause algorithmic bias and requires additional human supervision and remedy. Recently, Xin et al. proposed the idea of Human-in-the-Loop Machine Learning in Xin et al. (2018), which integrates human efforts into the traditional machine learning process and achieves impressive results on typical iterative workflows.

Computers, as one of the essential components in learning processes, can also become a barrier to the development of machine learning. Deep learning was first emerged in the early twentieth century but did not fully blossom until the late 1990s when the graphics processing units (GPU) were developed. The quality of computer hardware also largely contributes to the success of learning.

Summary and Conclusions

Machine learning offers a solution for accurate and quick data analysis and knowledge discovery in this big data era and enables the advancements in artificial intelligence. Despite the minor drawbacks mentioned in section “[Issues and Limitations](#),” machine learning now tackles various type of data under different problem settings and has been successfully applied to solve data problems in diverse research domains.

Cross-References

- [Big Data](#)
- [Deep Learning in Geoscience](#)

Bibliography

- Abu-Mostafa YS, Magdon-Ismael M, Lin HT (2012) Learning from data. AMLBook, Pasadena
- Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20(3): 273–297
- Hinton GE, Sejnowski TJ, Poggio TA et al (1999) Unsupervised learning: foundations of neural computation. MIT Press, Cambridge, MA
- Ho TK (1995) Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition, vol 1, pp 278–282
- Pouyanfar S, Sadiq S, Yan Y, Tian H, Tao Y, Reyes MP, Shyu ML, Chen SC, Iyengar S (2018) A survey on deep learning: algorithms, techniques, and applications. *ACM Comput Surv* 51(5):1–36
- Rumelhart DE, Hinton GE, Williams RJ (1988) Learning representations by back-propagating errors. MIT Press, Cambridge, MA, pp 696–699

- Samuel AL (1959) Some studies in machine learning using the game of checkers. *IBM J Res Dev* 3(3):210–229
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Methodol* 58(1):267–288
- Turing AM (2009) Computing machinery and intelligence. In: *Parsing the Turing test*. Springer, Dordrecht, pp 23–65
- Xin D, Ma L, Liu J, Macke S, Song S, Parameswaran A (2018) Accelerating human-in-the-loop machine learning: challenges and opportunities. In: *Proceedings of the second workshop on data management for end-to-end machine learning. DEEM'18*, Association for Computing Machinery, New York. <https://doi.org/10.1145/3209889.3209897>

Mandelbrot, Benoit B.

Katepalli R. Sreenivasan
New York University, New York, NY, USA



Fig. 1 Benoit B. Mandelbrot (1924–2010), courtesy of Prof. Laurent Mandelbrot, son of Benoit Mandelbrot

Biography

Benoit Mandelbrot has contributed much to several branches of science, often in pioneering ways, by seeing what others could not see. I will describe two of his contributions to turbulence (but omit other topics such as vortical boundaries and the R/S analysis of hydrological data). Chapter ten of his book (Mandelbrot 1982), which he regarded as “[a plea] for a more geometric approach to turbulence and for the use of fractals,” discusses his thoughts at the time; it is fair to say,

however, that those views have been superseded in significant ways.

One of them was his speculation, on the basis of a free-hand sketch prepared by S. Corrsin (1959), that the interface bounding a dye mixed by homogeneous turbulence is a fractal. His ideas on this topic, expressed in his 1982 book (also its 1977 predecessor), were free-wheeling yet remarkable because, at that time, there was no real evidence to back up his speculations. It took me and my colleagues many years of diligent work to bring realism to them; I cite two examples that are almost three decades apart (Sreenivasan 1991; Iyer et al. 2020). A comparison of Corrsin's schematic (see p. 54 of Benoit's 1982 book) was accomplished by actual direct numerical simulation, finally, in Sreenivasan and Schumacher (2010). This comparison reveals both the fertility of past imagination and its limitations.

The second contribution described here stemmed from the fact, well-known at the time, that turbulent energy dissipation is intermittent in amplitude, in both space and time. Benoit knew that a single fractal dimension would not suffice for describing this property, and developed further ideas (Mandelbrot 1974): while geometric objects such as interfaces can be characterized meaningfully by a single fractal dimension, others such as energy dissipation require a multitude of dimensions. Benoit's ideas were not yet sharp and, to his later chagrin, he did not use the term "multifractal" until after it was first used by Frisch and Parisi (1985); the characterization of multifractal measures (without using the word multifractal) was the work of Halsey et al. (1986). Later, Benoit connected these developments to his earlier ideas of 1974 and gave a beautiful probabilistic interpretation of multifractals, which led him to devise concepts such as negative and latent dimensions (see Chhabra and Sreenivasan 1991). Meneveau and I (1987, 1991) – see also Sreenivasan (1991) – connected these concepts meaningfully to turbulence, made measurements of the multifractal spectrum for energy dissipation and other such quantities, and developed and measured joint multifractal measures.

Benoit was my colleague at Yale for many years. My students and I benefitted greatly from discussions with him, for which I am grateful. His enthusiasm for our work was remarkably uplifting. He was well aware of being described as the "father of fractals," and he once told me that fractals as a subject wouldn't exist without him; he is probably right. His most memorable lines are recorded in his magnum opus, the 1982 book: "Clouds are not spheres, mountains are not cones, coastlines are not circles, and bark is not smooth, nor does lightning travel in a straight line." Toward the end of his career, Benoit tried to recast his legacy as having brought attention to the overarching theme of "roughness" as an important fact of Nature, whereas much of the rest of science was focused on smoothness.

All in all, Benoit was a unique scientist who ventured into several areas, some of which went beyond science, and originated key concepts in each of them. He has an assured place in the annals of Science.

Bibliography

- Chhabra AB, Sreenivasan KR (1991) Negative dimensions: theory, computation, and experiment. *Phys Rev A* 43:1114(R)
- Corrsin S (1959) Outline of some topics in homogeneous turbulent flow. *J Geophys Res* 64:2134
- Frisch U, Parisi G (1985) In: Ghil M, Benzi R, Parisi G (eds) *Turbulence and Predictability in Geophysical Fluid Dynamics*, Proc. Intern. School Phys. "E. Fermi", p 84
- Halsey TC, Jensen MH, Kadanoff LP, Procaccia I, Shraiman BI (1986) Fractal measures and their singularities: the characterization of strange sets. *Phys Rev A* 33:1141
- Iyer KP, Schumacher J, Sreenivasan KR, Yeung PK (2020) Fractal iso-level sets in high-Reynolds-number scalar turbulence. *Phys Rev Fluids* 5:044501
- Mandelbrot BB (1974) Intermittent turbulence in self-similar cascades: divergence of high moments and dimension of the carrier. *J Fluid Mech* 62:331
- Mandelbrot BB (1982) *The fractal geometry of nature*. W.H. Freeman & Co.
- Meneveau C, Sreenivasan KR (1987) Simple multifractal cascade model for fully developed turbulence. *Phys Rev Lett* 59:1424
- Meneveau C, Sreenivasan KR (1991) The multifractal nature of turbulent energy dissipation. *J Fluid Mech* 224:429
- Sreenivasan KR (1991) Fractals and multifractals in fluid turbulence. *Annu Rev Fluid Mech* 23:539
- Sreenivasan KR, Schumacher J (2010) Lagrangian views on turbulent mixing of passive scalars. *Phil Trans Roy Soc Lond A* 368:1561

Markov Chain Monte Carlo

Swathi Padmanabhan and Uma Ranjan

Indian Institute of Science, Bangalore, Karnataka, India

Definition

Probabilistic inference is a way to derive the probability of a set of random variables, and it can take a set of value/values. Markov chain Monte Carlo (MCMC) is a stochastic process that aims to get the inference through efficient sampling means. Markov chain Monte Carlo (MCMC) is a class of methods that combines the Markov chains to Monte Carlo sampling for efficient sampling. This method helps construct the "most likely" distribution that gets us closer to the actual or expected distribution. The Bayesian approach helps quantify the uncertainty and thus gives a method to sample the inference. Various distributions of physical parameters can be studied with the help of this MCMC as the use of posterior distribution maximizes the expectation of the desired

parameter. Many extensive applications can be found for MCMC from analyzing rock physics to understand the weak points of rock to developing efficient deep learning algorithms to better analyze various phenomena. Some include modelling distributions of geochronological age data and inference of sea-level and sediment supply histories from 2D stratigraphic cross-sections (Gallagher et al. 2009) and so on.

Introduction

In the computation of Bayesian inverse problems, an important step is that of sampling the posterior distribution. The parameters of interest correspond to the properties of the posterior distribution. The direct calculation of this parameter is computationally intractable. Hence, the posterior probability must be approximated by other means (Svensén and Bishop 2007). Typically, the desired calculations are done as a sum of a discrete distribution of many random variables or integral of a continuous distribution of many variables and is difficult to calculate. This problem exists in both schools of probability (Bayesian and Frequentist); it is more prevalent with Bayesian probability and integrating over a posterior distribution for a model.

One common solution to extract inference is by Monte Carlo sampling. This method draws independent samples from the probability distribution, and then repeat this process many times to approximate the desired physical quantity. Markov chain is a systematic method of predicting the probability distribution of the future or next state based on the current state. Monte Carlo samples the random variables independent from each other directly from the desired probability density function but fails at higher dimensions. Markov chain Monte Carlo (MCMC) combines both these methods and allows random sampling of high-dimensional probability distributions that takes into account the probabilistic dependence (transition probability) between samples by constructing a Markov chain that comprise the Monte Carlo sample (Gelman and Lopes 2006).

Historically speaking, Monte Carlo methods have existed since World War II times. The first MCMC algorithm, now Metropolis algorithm, was published in 1953. But more practical aspects and integration with statistics took when the work of Hastings came in the 1970s. The development of Monte Carlo methods during the 1940s coincided with the development of the first computer, ENIAC. Von Neuman set the computation of the Monte Carlo method was set up by Von Neuman for analyzing thermonuclear and fission problems in 1947. Fermi invented the version without computers in the 1930s, but it did not gain any recognition. The Metropolis algorithm (first MCMC algorithm) was developed using MANIAC computer under the guidance of Nicholas Metropolis in the early 1950s. From the 1990s to the present, the

evolution involves these algorithms taking in various advances like the Gibbs sampling, perfect sampling, regeneration, and central limit theorem for convergence and so on. Over time, the algorithms have evolved and significantly advanced along with the growth of computation; these techniques have influenced many aspects and applications (Robert and Casella 2011).

Monte Carlo Sampling

Monte Carlo estimates the physical quantity desired as the expectation of random variable averaged over a large number of samples for a desired density function. MC methods are used to solve problems that are stochastic in nature (particle transport, telecommunication systems, population studies based on the statistical nature of survival and reproduction) and that are deterministic (solving integrals, complex linear algebra equations, solving complex partial differential equation (PDE)). The main advantages of using Monte Carlo sampling are to:

1. Estimate density by collecting samples to approximate the distribution of a target function.
2. Approximate a parameter (physical quantity), like the mean or variance of a distribution.
3. Optimize a function by locating a sample that maximizes or minimizes the target function or the desired distribution.
4. Estimate more than one physical quantities simultaneously by using more than one random variables.

The random numbers are generally considered to follow a uniform distribution or any probability function desired. The random number generated are independent, thus having no correlation between each sample generated. Sometimes, pseudorandom sequences are used for rerunning the simulations where the pseudorandom sequences involves a known deterministic sequence to sample the random the samples from the probability density function of the physical quantity that needs to be predicted. This method is generally known as inverse distribution method and this is used for Monte Carlo simulations in optical imaging methods (Wang and Wu 2012). The sampling process in Monte Carlo can be explained as follows (Goodfellow et al. 2016).

Consider the expectation defined over an integral of a probability density (or distribution) over a random variable x :

$$q = \int f(x)p(x) = E[f(x)] \quad (1)$$

If Eq. 1 is the integral to estimate the quantity q , then q can be approximated by drawing “ n ” number of samples and then by averaging it as (Goodfellow et al. 2016):

$$\hat{q}_n = \frac{1}{n} \sum_{i=1}^n f(x^i) \quad (2)$$

$\hat{q}_n$ is the average of n samples for a quantity, q . The **law of large numbers** in statistics states that if the samples x^i are i.i.d, then the average value, $\hat{q}_n$, will definitely converge to the expected values of the random variable, provided the variance is bounded by:

$$\text{Var}[\hat{q}_n] = \frac{\text{Var}[f(x)]}{n} \quad (3)$$

$\text{Var}[\hat{q}_n]$ is the variance of $\hat{q}_n$, $f(x)$ is the distribution of the samples and $\text{Var}[f(x)]$ is the variance of the $f(x)$. By **central limit theorem**, we can say that the distribution of $\hat{q}_n$ to a normal distribution has a finite mean and variance. In this case, the mean will be q and variance as in Eq. 3. Hence, from these outcomes, accurate estimations of $\hat{q}_n$ can be done from the corresponding cumulative distribution of the normal density function. The graphical illustration (Wang and Wu 2012) of sampling from a distribution described in Fig. 1 will give a better knowledge of generating the pseudorandom random numbers when we have specific probability density function as goal.

Drawbacks

The main drawbacks of Monte Carlo Sampling are:

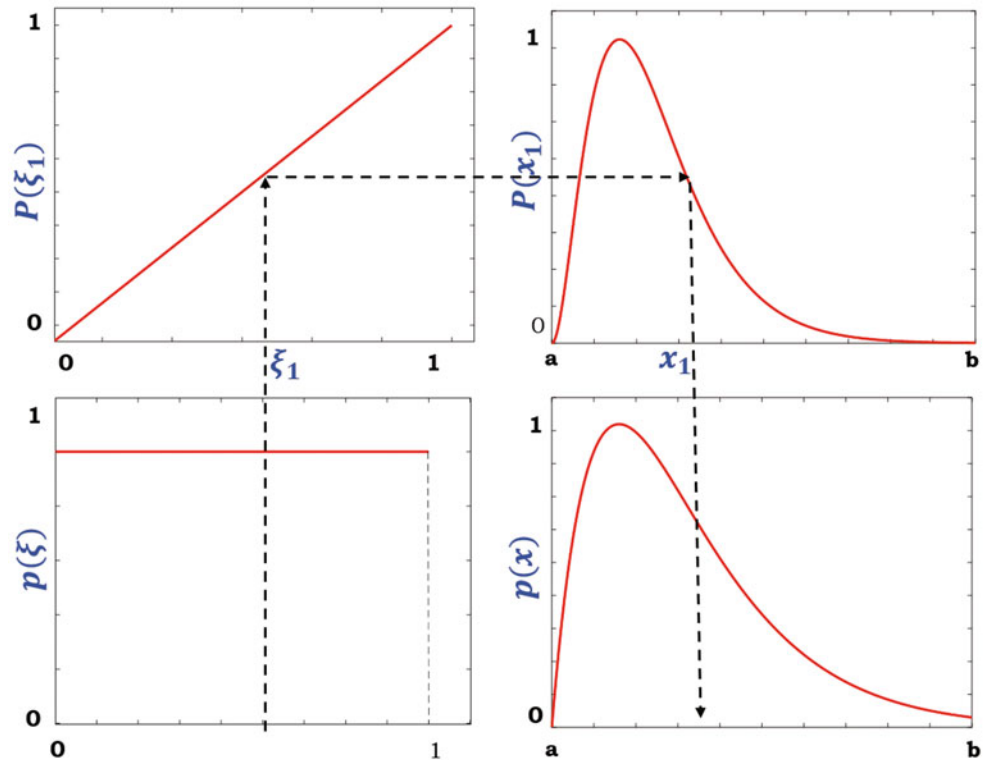
- MC sampling fails at high dimensions because the volume of the sample space increases exponentially with the number of parameters (dimensions).
- MC samples each of the random sample from the desired base probability density function and each of them are independent and is drawn independently as well. This increases the complexity of the model.

Markov Chain Monte Carlo (MCMC)

The framework for MCMC provides a general approach for generating Monte Carlo samples from the posterior distribution, in cases where we cannot efficiently sample from the posterior directly. We get the posterior distributions in Bayesian inference (Ravenzwaaij et al. 2018). In MCMC, we construct an iterative process that gradually samples from distributions that are closer and closer to the posterior. The challenge is to decide how many iterations we should perform before we can collect a sample as being (almost) generated from the posterior (Koller and Friedman 2009).

Markov Chain Monte Carlo,

Fig. 1 Example illustration of sampling a random variable x from a desired probability density function $p(x)$ based on a random number ξ generated from a uniform distribution in interval $[a, b]$. The graph also shows the mapping with the cumulative functions of the same.



Markov Chains

Markov Chain is a special type of stochastic process, which deals with characterization of sequences of random variables. It focuses on the dynamic and limiting behaviors of a sequence (Koller and Friedman, 2009). It can also be defined as a *random walk* where the next state or move is only dependent upon the current state and the randomly chosen move.

Also, the sampling can be explained using chain dynamics as follows. Consider a random sequence of states $x^0, x^1, x^2, \dots, x_k$ with weights $w_1, w_2, w_3, \dots, w_k$ is shown in Fig. 2. Let the state of the process at step “ k ” be viewed for a random variable X^k . We assume that the initial state X^0 is distributed according to some initial state distribution $P^0(X^0)$. The definition of subsequent states of P are $P^1(X^1), P^2(X^2), P^3(X^3)$.

$$P^{(k+1)}(X^{(k+1)} = x') = \sum_{x \in \text{Val}(X)} P^{(k)}(X^{(k)} = x) T(x \rightarrow x') \quad (4)$$

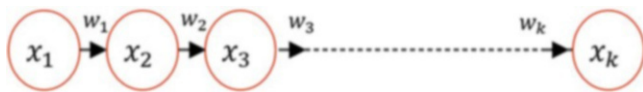
where the state transition from x to x' at $k+1$ is nothing but the sum of all the possible states of x at step “ k .” Markov chains provides good insight to random walk models, but as a standalone model, it is inefficient to reach inferences.

Designing MCMC

For implementing a Markov chain Monte Carlo, there are certain conditions of the process to be understood. Firstly, the Markov chain process is considered to be stationary, or it is sampled until it comes to the stationary state. If we arrive at the stationary distribution (invariant distribution) of Markov chain, then it continues to remain in that state for all the future samples of Monte Carlo. The condition for a Markov chain to be stationary is given by definition as (Koller and Friedman 2009):

$$\pi(X = x') = \sum_{x \in \text{Val}(X)} \pi(X = x) T(x \rightarrow x') \quad (5)$$

where the T denotes the transition probability and π denotes the probability of the states of each sample X . An



Markov Chain Monte Carlo, Fig. 2 A very simple Markov chain representing states $x_0, x_1, x_2, \dots, x_k$ with weights $w_1, w_2, w_3, \dots, w_k$

important property of a transition model of Markov chain is that, for a stationary distribution, the transition model T can be viewed as a matrix A (also known as adjacency matrix). The stationary Markov distribution is then an eigen vector of the matrix A corresponding to the eigen value 1.

The representation of a Markov chain sampling using adjacency matrix, A in linear algebra context and π as probability states:

$$\pi A = \pi \quad (6)$$

where this relation shows the eigen vector relationship between the states and π is the target distribution.

In MCMC, the samples are drawn from the probability distribution by building a Markov Chain, and the next sample is drawn from the probability distribution that dependent upon the last sample that was drawn. After an initial number of samples, the distribution will settle or converge to become stationary (equilibrium) that will help estimate the physical quantity desired. Hence, the introduction of correlation between the sampling of variables in Monte Carlo will enable efficient estimation of the density or quantity rather than wandering over a domain.

Markov Chain Monte Carlo Algorithms

Various algorithms exist for MCMC. The two commonly used for MCMC are:

- The Gibbs sampling algorithm
- Metropolis-Hastings algorithm

The idea of MCMC is that the sequences are constructed such that, although the first sample may be generated from the prior, subsequent samples are generated from distributions that conclusively gets closer and closer to the desired succeeding values for samples (Koller and Friedman 2009).

The initial samples are important for MCMC as it is during these samples, the distributions are observed to know whether it is converging or not. From the initial point, the samples undergo a *warm-up* phase where the samples are being chained to the Markov process to reach useful values. Generally, some prior samples are eliminated or discarded from the initial phase, once useful values are attained. This is referred to as *burning-in* of Markov process that marks entry to the stationary distribution of Markov chain (Murphy 2019).

For constructing a Markov chain, the detailed balance condition is necessary.

The detailed balance condition for a finite state Markov chain is reversible if there exists a unique distribution p for all states of x :

$$\pi(\mathbf{x})T(\mathbf{x} \rightarrow \mathbf{x}') = \pi(\mathbf{x}')T(\mathbf{x}' \rightarrow \mathbf{x}) \quad (7)$$

Equation 7 asserts that the probability of a transition from $\mathbf{x}$ to $\mathbf{x}'$ is the same as the probability of a transition for $\mathbf{x}'$ to $\mathbf{x}$, which implies that a random transition from state $\mathbf{x}$ to $\mathbf{x}'$ is **reversible** and the $\mathbf{p}(\mathbf{x})$ forms a **stationary distribution** to the transition probability T (Koller and Friedman 2009).

Gibbs Sampling

Gibbs sampling is a technique used when we have to consider more than one dimensions. The density functions may also be multivariate distributions. The probability of the next state is calculated from the conditional probability of the prior state. Therefore, the new states or samples are conditioned over all the values of the entire distribution (Murphy 2019).

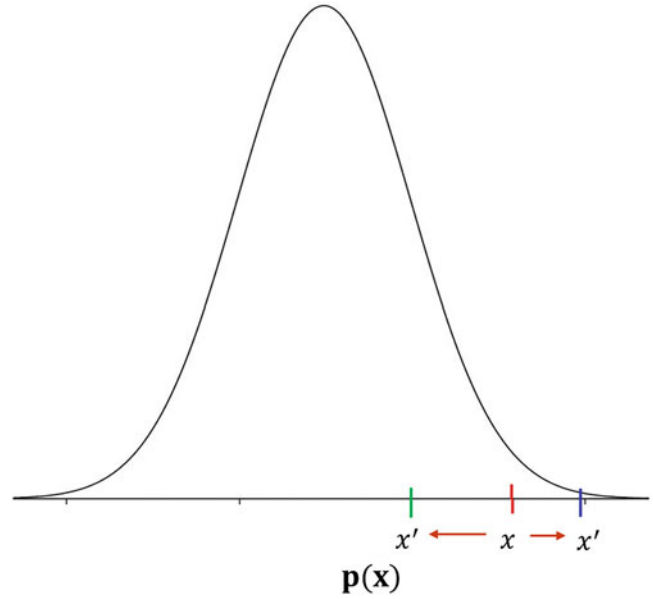
The conditional sampling method is easy and accurate when the dimensions are low and more difficult as dimensions increase. Gibbs sampling is an easy to implement model. The samples are acquired from the posterior distribution $P(Q|E = e)$ where “e” is the observations, $Q = X - E$. The states in the Markov chain will be defined from $\mathbf{x}$ to $\mathbf{x} - E$. Then, transition states are defined so as to converge to a stationary distribution $\pi(Q)$. For arriving at this condition, let the states in the Markov chain be (x_{-i}, x_i) . Transition probability of the upcoming states are given as:

$$\begin{aligned} T_i(x_{-i}, x_i) &\rightarrow (x_{-i}, x'_i) \\ T_i(x_{-i}, x_i) &= P(x'_i | x_{-i}) \end{aligned} \quad (8)$$

From Eq. 9, it can be understood that the transition probability for Gibbs sampling depends only on the remaining states x_{-i} for a posterior distribution $P_N(X|e)$ for a set of N factors. This shows that $P_N(X|e)$ is stationary for the process considered (Koller and Friedman 2009).

Metropolis-Hastings

Metropolis-Hastings was initially published in 1970, which was a paper that generalized the Metropolis algorithm and to overcome the dimensionality drawback of the MC methods. In this algorithm, the stationary distribution needed is acquired using the detailed balance condition in Eq. 7. with a particular stationary distribution. The Metropolis-Hastings algorithm requires a *proposal distribution* as we cannot sample directly from the required distribution. It uses the concept of acceptance probability to decide is the proposed state can be accepted or rejected. If we denote the *acceptance probability* as $A(\mathbf{x} \rightarrow \mathbf{x}')$, which gives the probability of the state, $\mathbf{x}'$ getting accepted over the proposal distribution τ^q defines the



Markov Chain Monte Carlo, Fig. 3 Illustration of acceptance probability using an example distribution $p(x)$ on the proposed states. If the $\mathbf{x}'$ is towards the denser part of distribution (green point), then the state is accepted else it might or not be accepted as in the state denoted by blue point

transition model for a given state space. The random states are generated using the defined proposal distribution.

When two successive states are generated, the state can be either accepted or rejected based on their acceptance probabilities. For example, if the new state $\mathbf{x}'$ is found to satisfy the desired density function $\mathbf{p}(\mathbf{x})$ and the prior state is $\mathbf{x}$, the chance of acceptance of $\mathbf{x}'$ as the next state is higher if the state is lying closer to the higher density as in Fig. 3. This will lead to the successive states to be closer to the probability values populated around that region. Suppose if the next state is farther away as in right hand side of the distribution in Fig. 3, then it may or may not get accepted. The acceptance here depends on the distance between the current state and the proposed state.

Therefore, the condition for acceptance can be defined based on the detailed balance as:

$$\pi(\mathbf{x})T(\mathbf{x} \rightarrow \mathbf{x}')\tau^q(\mathbf{x} \rightarrow \mathbf{x}') = \pi(\mathbf{x}')T(\mathbf{x}' \rightarrow \mathbf{x})\tau^q(\mathbf{x}' \rightarrow \mathbf{x}) \forall \mathbf{x} \neq \mathbf{x}' \quad (9)$$

Then, the acceptance probability can be defined for a stationary distribution as:

$$A(\mathbf{x} \rightarrow \mathbf{x}') = \min\left(1, \frac{\pi(\mathbf{x}')\tau^q(\mathbf{x}' \rightarrow \mathbf{x})}{\pi(\mathbf{x})\tau^q(\mathbf{x} \rightarrow \mathbf{x}')}\right) \quad (10)$$

Or in other words, using conditional probability, τ^q can be denoted as a function, say V , and then

$$A(\mathbf{x} \rightarrow \mathbf{x}') = \min\left(1, \frac{p(\mathbf{x}')V(\mathbf{x}|\mathbf{x}')}{p(\mathbf{x})V(\mathbf{x}'|\mathbf{x})}\right) \quad (11)$$

If the condition for Eqs. 9 and 10 is satisfied, then the process has reached stationary state. Gibbs Sampling method is a special case of the Metropolis-Hastings method. There are cases when it becomes difficult to use the Metropolis-Hastings algorithms as result of nonlinearity in the models. In applications, Metropolis-Hastings is used in conjunction with a Gibbs sampler, but when nonlinearity in model arises, it becomes difficult to sample from the full conditional distributions. In those cases, some resampling techniques like rejection method is employed in the Metropolis-Hastings algorithm (Gamerman and Lopes 2006) especially in the context of nonlinear and non-normal models.

Summary and Conclusions

In short, Markov chain Monte Carlo (MCMC) is a method for efficiently sampling the random variables for accurate or exact estimation of the parameters necessary. Monte Carlo sampling alone cannot provide very good estimations as it fails at higher dimensions particularly. The beauty of extracting inference from Bayesian statistics is what forms the basis for MCMC. It focusses on generating samples from a posterior distribution using a reversible Markov chain that will have its converging or equilibrium distribution as the posterior distribution. The dependency created due to integration of Markov chains will ensure fast convergence to the true or exact solution while combined with Monte Carlo. The conditions for a good MCMC can be defined with the help of transition probability and the stationary condition for Markov chain. The two common algorithms for MCMC are the Gibbs sampling and the Metropolis-Hastings algorithms. Metropolis-Hastings is a more general and commonly used algorithm, while Gibbs sampling can be considered as a special case of Metropolis-Hastings. MCMC based on Metropolis-Hastings is used more because of its flexibility over other algorithms. Gibbs algorithm is more often used when the next step is the conditional probability distribution as the *proposal distribution*. There are generalized algorithms when the *proposal distributions* are symmetric in nature like the Gaussian (Murphy 2019). For efficient convergence of the models, we can design the algorithms that has altered Markov chain sequences or that alter the samples generated or the transition kernels based on the complexity of the process under study (Gamerman and Lopes 2006). Different strategies can be adopted based on the requirement of the desired target distribution.

MCMC in Geosciences: Various geophysical problems involve inverse problems wherein an input parameter and an output solution are obtained via inversion of an operator. This is commonly used in seismological models and resolution analysis and model optimization (Gallagher et al. 2009), for example, finding randomly searching earth's consistent model with seismological model. MCMC provides improved reliability in GIS systems. Generally, GIS problems are ill-posed and solved linearly. Solving nonlinear cases involve non-Gaussian changes and requirement of a posterior distribution. Since complexity involved is high, MCMC improves the prediction and reliability for the nonlinear cases. SAR (synthetic aperture radar) land applications acquires images that are RAW images composed of complex data. The speckles contained can be removed by Markovian random field grouping and the method involves estimating iteratively estimation of parameter, and the conditional posterior distribution of each pixel is updated at every step of the algorithm, with only dependent on its neighboring pixels.

Cross-References

- [Markov Chains: Addition](#)
- [Monte Carlo Method](#)

References

- Gallagher K, Charvin K, Nielsen S, Sambridge M, Stephenson J (2009) Markov chain Monte Carlo (MCMC) sampling methods to determine optimal models, model resolution and model choice for Earth Science problems. *Marine Petrol Geol* 26(4): 0264–8172
- Gamerman D, Lopes HF (2006) Markov chain Monte Carlo: stochastic simulation for Bayesian inference. CRC Press, Boca Raton
- Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge
- Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. MIT Press, Cambridge
- Murphy KP (2019) Machine learning: a probabilistic perspective. MIT Press, Cambridge
- Robert C, Casella G (2011) A short history of Markov chain Monte Carlo: subjective recollections from incomplete data. *Stat Sci Inst Math Stat* 26:102–115. <https://doi.org/10.1214/10-STS351>
- Svensén M, Bishop CM (2007) Pattern recognition and machine learning. Springer, New York
- Van Ravenzwaaij D, Cassey P, Brown SD (2018) A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic Bull Rev* 25(1):143–154
- Wang LV, Wu H-i (2012) Biomedical optics: principles and imaging. Wiley, New York

Markov Chains: Addition

Adway Mitra

Centre of Excellence in Artificial Intelligence, Indian Institute of Technology Kharagpur, Kharagpur, India

Definition

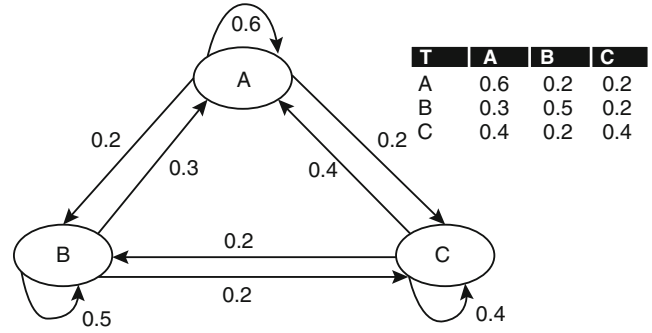
Consider a system, which can be described at any point of time t using a *state variable* X_t . This state variable may be a D -dimensional vector (where D may be 1 or higher), defined on a space X called the *state space*, and each individual value in the state space is called a *state*. This space can be either discrete or continuous, depending on the nature of the system. Time may be considered to be *discrete*. *System dynamics* refers to the process by which the state of the system changes over time, so that we have a sequence of states $\{X_1, X_2, \dots, X_{t-1}, X_t, X_{t+1}, \dots\}$. The changes of state from one time-point to the next are known as *state transitions*. If the state space is discrete and finite, we can look upon a Markov chain as a finite-state machine which shifts between the different states.

Suppose, these state variables are *random variables*, i.e., their values are assigned according to probability distributions. We describe such a system as a *Markov chain* if the state transition from X_{t-1} to X_t is governed by a probability distribution that depends only on X_{t-1} , but not any previous state or any other external information, for any t . This is known as the *Markov property*. In other words

$$p(X_t | X_{t-1}, X_{t-2}, \dots, X_1) = p(X_t | X_{t-1}) \quad (1)$$

The conditional distribution $p(X_t | X_{t-1})$ satisfies the following relation: $\text{prob}(X_t = b | X_{t-1} = a) = T_t(a, b)$ where a and b are two states and T_t is any function specific to time t that defines a probability distribution for X_t with respect to b , i.e., $T_t(a, b) \geq 0 \forall a, b$, and $\sum_{b \in X} T_t(a, b) = 1$ (if χ is discrete), or $\int_X T_t(a, b) db = 1$ (if χ is continuous). If the state space is discrete and countable, T may be looked upon as a *state transition table*, where each row denotes a possible value of the current state, each column denotes a possible value of the next state, and the corresponding entry indicates the probability of such a state transition. An example of such a Markov chain is given in Fig. 1.

We can define various properties of Markov chains using *marginal distributions* of individual state variables, i.e., $P(X_t)$ for any t , or their *joint distributions*, i.e., $P(X_{t1}, X_{t2}, \dots, X_{tn})$, n is any integer and $(t1, t2, \dots, tn)$ are arbitrary time-points. A Markov chain is said to be *stationary* or *homogeneous* if the following holds



Markov Chains: Addition, Fig. 1 A three-state discrete Markov chain showing the state transition table

$$P(X_{t1}, X_{t2}, \dots, X_{tn}) = P(X_{t1+k}, X_{t2+k}, \dots, X_{tn+k}) \forall n, k \quad (2)$$

In a homogeneous Markov chain, T_t is independent of t . Further, by setting $n = 1$, it follows that the marginal distributions of all state variables of a stationary Markov chain are identical, and this distribution is called the *stationary distribution* of the Markov chain, denoted by π . This basically represents the probability distribution over the state space that the state variable can take at any time. In case of discrete state spaces, where T is the state transition matrix, it can be shown that $\pi T = \pi$.

Introduction

Markov chains are widely used in various domains of science and engineering to represent stochastic processes. They are specifically useful in applications where there is a system whose state is continuously evolving, but the evolution has a randomness associated with it. Typically, the system state cannot be directly measured. However, there are measurable quantities that are functions of the system state, and these functions may be either deterministic or stochastic. Markov chains are often used to model the system state in such situations. The main reason for choosing them are as follows: (i) They can take care of the stochasticity, and (ii) the Markov assumption is simplifying and hence computationally tractable. In most cases, the task is to solve the *inverse problem*, i.e., estimate the system state based on the observations, even though it is really the system state which produces the observations. A good example is in atmospheric sciences, where the state of the atmosphere cannot be measured directly, but other dependent variables such as temperature, precipitation, and wind speed can be measured. Solving the inverse problem also enables us to estimate the dynamics of state transition, which in turn enables forecasting.

Markov models find applications in a very wide range of domains. They are used in natural language processing for assigning part-of-speech tags to words in a sentence; in

speech processing for identifying the speakers in a long audio stream; in activity recognition to identify a sequence of activities by a person based on readings from wearable sensors; in bioinformatics to find the most likely alignment between sequences of DNA, RNA, or proteins; and many others. Though the initial applications were mostly in the domains of computer science, signal processing, and computational biology, the past decade has seen their uses in other disciplines including earth system sciences.

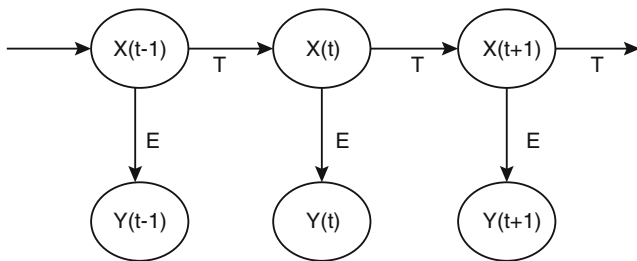
Variants and Derived Concepts

Various concepts related to or derived from Markov chains are used frequently to solve various problems related to geosciences. Some of the major variants of Markov chains and concepts based on them are discussed below.

Hidden Markov Models

Consider a sequence of observations $\{Y_1, \dots, Y_t, \dots, Y_T\}$. Let us consider this as a random variable which may take values from an *observation space* denoted by y . Y_t is a P -dimensional vector ($P \geq 1$) that may be continuous or discrete. According to the *hidden Markov model*, the unobserved state variable X is a Markov chain over a *discrete state space*, and the observation Y_t follows a distribution over the observation space that is a function of X_t , i.e., $P(Y_t = y | X_t = x) = E(x, y)$. Here $E(x, y)$ is called the *emission distribution* that satisfies $E(x, y) \geq 0 \forall (x, y)$ and $\sum_{y \in \mathcal{Y}} E(x, y) = 1$ (if Y is discrete) and $\int_{\mathcal{Y}} E(x, y) dy = 1$ (if Y is continuous).

The salient feature of this model is that the current system state depends only on the previous state, and the current observation depends only on the current state. This can be succinctly represented by a graphical model, as shown in Fig. 2. These assumptions may be simplistic, but they have an important benefit: they allow efficient and exact inference, i.e., we can exactly compute the distribution $p(X_t | Y_1, \dots, Y_T)$ for any t , and also we can compute the most likely sequence of values of the unobserved state variable, given a sequence of observations.



Markov Chains: Addition, Fig. 2 Graphical representation of hidden Markov model

For specialized applications, various extensions of hidden Markov models are used. The most common one for earth science applications are the *non-homogeneous hidden Markov model*, where the state variable X_t is driven by an additional external variable U_t in addition to its past value denoted by X_{t-1} . In other words, $Prob(X_t = b | X_{t-1} = a, U_t = c) = T(a, b, c)$. $-1 = a, U_t = c) = T(a, b, c)$. This allows the state variable's marginal distribution to vary over time. Another well-known variant is the *factorial hidden Markov model*, where the state space is decomposed into smaller spaces, each represented by separate state variables ($X^1, X^2, \dots, X^K$) in which individual Markov chains are defined. The observation Y is a function of all of these K variables, i.e., $prob(Y_t = y | X_t^1 = x_1, X_t^2 = x_2, \dots, X_t^K = x_K) = E(x_1, x_2, \dots, x_K, y)$.

Markov Chain Monte Carlo

Another important concept based on Markov chains, which is frequently used in the domain of geosciences is *Markov chain Monte Carlo (MCMC) sampling*. This is essentially a sampling technique, which enables us to estimate the posterior distribution of an unknown quantity X , such as the state variable. It is essentially a *Monte Carlo* method of computation through random sampling that is based on the *law of large number* of statistics. The idea is to consider different possible values of the variable X as the states of a Markov chain and estimate the posterior distribution of X as the stationary distribution of such a Markov chain. To estimate this, MCMC methods aim to perform *random walk* over the state space. But this is possible only if there is a transition function from one state to another, and this matter is handled in different ways.

Consider the special case where the conditional distribution of each dimension of X , conditioned on all other dimensions, i.e., $P(X_i | X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_D)$, is known for all $i \in \{1, D\}$, where D is the dimensionality of the variable X . In this situation, we may use *Gibbs sampling*, where we choose any one dimension and sample its value from the corresponding conditional distribution, keeping the values of all other dimensions unchanged. Consider an initial value x_0 of X . Then using the above procedure, we can sample a new value x_1 , where at most one dimension will have a different value from x_0 . This is essentially a transition of the Markov chain. Next we again choose another dimension, sample its value conditioned on the rest, and hence get another sample x_2 . This process is repeated over and over, till we get a large enough number of samples of X that effectively represent the stationary distribution of this Markov chain and hence the posterior distribution of X .

In a more general case where these conditional distributions are not known, but we do know a function $f(X)$ that is proportional to the unknown posterior distribution $P(X)$, we may use another Markov chain Monte Carlo method known

as *Metropolis-Hastings algorithm*. Here too we start with an initial estimate x_0 of X . We choose another candidate value y according to a *proposal distribution* $Q(y|x)$. However, this value is accepted with a probability that is related to its relative likelihood with respect to the current value x_0 , i.e., according to the ratio $\frac{f(y)}{f(x_0)}$. If accepted, we set $x_1 = y$ and then proceed in the same way to get the next sample x_2 . However if y is rejected, then we again sample more values and test them, until one of them is accepted and set as x_1 . Repeating this process enables us to collect a sample set of X which effectively represents the stationary distribution of this Markov chain and hence the posterior distribution of X .

In either case, while collecting the samples, it is necessary to consider samples at regular intervals only (as successive samples are highly correlated to each other). There are also many other algorithms based on the concept of Markov chain Monte Carlo, such as slice sampling. There are also various hybrid methods to enable faster navigation of the state space and focusing on high-density areas (sample with higher likelihood under the posterior P). For a more detailed understanding of these algorithms, the reader is referred to the excellent tutorial paper by Neal (1993).

Applications in Geosciences

Now, we come to case studies of various applications of the above models and concepts in various domains of geosciences, such as climate, hydrology, geology, seismology, and remote sensing.

Hydroclimatology

Hidden Markov models have been used quite extensively for modeling rainfall at hourly or daily spells. The main reason for this is that rainfall usually occurs in spells that may last for a few hours (in case of convective events like thunderstorms), or may hang around for a few days (in case of depressions), and hence the rainfall occurrence and intensity in adjacent time-steps are likely to be related. However, the process of rainfall is dependent on many complex processes in the atmosphere, not all of which can be observed. Hence, it is a common practice in statistical hydro-meteorology to represent the daily/hourly *weather state* as a latent state variable that follows a Markov chain, and the rainfall at any time-step depends only on the weather state at the corresponding time. However, this weather state may be affected by other atmospheric variables like geopotential height, temperature, and relative humidity, which are introduced as external variables affecting the latent state, as in a *non-homogeneous hidden Markov model (NHMM)*. This modeling style was first

introduced by Bellone et al. (2000) and has continued for the next two decades. In these models, the emission distribution of rainfall are usually hybrid: First, a Bernoulli distribution decides whether there is any rain or not, and if so, then the volume is generated by another distribution which may be gamma, exponential, log-normal, etc. In some studies, the parameters of the emission distribution too depend on the external variables.

A more sophisticated application of the same idea is developed by Kwon et al. (2018), where the observations are measurements of soil moisture at multiple sites. The latent variable represents the spatial profile of soil moisture. The covariates or external variables of the NHMM are weather observations like temperature and rainfall. The challenge here is to learn the optimal number of values that the state variable can take, and the parameters related to transition and emission distributions. Successful estimation of the model enables the prediction of soil moisture based on weather forecasts.

Hydrology

In the domain of hydrology, hidden Markov models have been used for modeling daily streamflow sequences in many studies such as Pender et al. (2016). Here, the innovation lies in the emission distribution, which has been specifically designed to model *extreme values* using extreme value distributions like generalized Pareto (GP) and generalized extreme value (GEV). Some of the states signify extreme conditions of streamflow, for which these distributions are used for emission.

Zhang et al. (2018) applied MCMC-based methods in the domain of hydrology, for the purpose of estimating parameters of a groundwater model. They used a Bayesian approach to quantify the uncertainties of the model parameters, conditioned on the magnetotelluric measurements for subsurface salinity of groundwater across a vertical profile. Different types of MCMC algorithms were used to estimate a posterior distribution of the model parameters. In the simple Metropolis-Hastings, a number of sample parameter values were drawn from the prior distribution (based on empirical studies) and accepted or rejected based on the ratio of likelihoods with respect to the hydrological model. In the second approach, an adaptive sampling strategy was considered to navigate the sample space more efficiently, as all accepted samples were used to progressively update the prior distribution. This ensured that samples that are closer to the high-density regions of the posterior are picked, and hence fewer samples are rejected.

Geology and Geophysics

Geostatistics is an important concept in the domain of geology and geophysics which aims to estimate geological values

at any arbitrary spatial location, with the help of a *geostatistical equation*. The geostatistical equation has parameters which may be estimated based on a few locations where the desired variable is measured, using algorithms like *kriging*. However, Oh and Kwon (2001) use a Bayesian approach for estimation of these parameters based on Schlumberger inverse and well log observations of dipole-dipole resistivity. For this, they need a posterior distribution over the parameter space. They first build a *prior* distribution using simulations conditioned on the observations, after which they carry out MCMC sampling of parameter values. They draw samples from the prior distribution, use them to run simulations, compute the error at the points where observations are available, and subsequently accept or discard these samples based on the error. The prior distribution gets updated as more samples are accepted.

Another work by Jasra et al. (2006) considers the use of MCMC approaches to identify the most suitable model parameters, including structural parameters such as number of components in a mixture distribution. They analyze *geo-chronological data* to identify age groups of minerals. The geological properties of minerals from each age group are represented by a component of the mixture distribution. However, the number of mixture components is not fixed but itself a random variable, for which a distribution is estimated using a variant of MCMC sampling called reverse-jump MCMC, which allows for *birth* and *death* of components.

Remote Sensing

Markov chain-related concepts have been considered in the domain of remote sensing too. Harrison et al. (2012) used Markov chain Monte Carlo sampling to estimate posterior distributions on the parameters of land surface models and the uncertainty in simulations of soil moisture by them, by calibrating these models against passive microwave remote sensing estimates of soil moisture. Liu et al. (2021) consider a spatiotemporal HMM for classifying the pixels of satellite image sequences according to land cover types. The spatial consistency of adjacent pixels and the dynamics of temporal changes in land cover are captured through this version of HMM.

Summary and Conclusions

The concept of Markov chain has been developed into models like hidden Markov models and its variants and algorithmic frameworks like Markov chain Monte Carlo. These models and methods find applications in many problems related to mathematical geosciences, especially for forecasting, simulation, and uncertainty estimations. Recent challenges being

explored by the research community are to include spatial and temporal characteristics of processes at multiple scales into these models and applying them in conjunction with deep learning – the state-of-the-art approach being increasingly adopted by researchers of all scientific domain, including geosciences.

Cross-References

- [Bayesian Inversion in Geoscience](#)
- [Computational Geoscience](#)
- [Geostatistics](#)
- [High-Order Spatial Stochastic Models](#)
- [Interpolation](#)
- [Machine Learning](#)
- [Markov Chain Monte Carlo](#)
- [Markov Random Fields](#)
- [Remote Sensing](#)
- [Sequential Gaussian Simulation](#)
- [Spatial Data](#)
- [Spatial Statistics](#)
- [Spatiotemporal Analysis](#)
- [Spatiotemporal Modeling](#)

Bibliography

- Bellone E, Hughes JP, Guttorp P (2000) A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. *Clim Res* 15(1):1–12
- Harrison KW, Kumar SV, Peters-Lidard CD, Santanello JA (2012) Quantifying the change in soil moisture modeling uncertainty from remote sensing observations using Bayesian inference techniques. *Water Resour Res* 48(11):11514
- Jasra A, Stephens DA, Gallagher K, Holmes CC (2006) Bayesian mixture modelling in geochronology via Markov chain Monte Carlo. *Math Geol* 38(3):269–300
- Kwon M, Kwon HH, Han D (2018) A spatial downscaling of soil moisture from rainfall, temperature, and amsr2 using a Gaussian-mixture nonstationary hidden Markov model. *J Hydrol* 564: 1194–1207
- Liu C, Song W, Lu C, Xia J (2021) Spatial-temporal hidden Markov model for land cover classification using multitemporal satellite images. *IEEE Access* 9:76493–76502
- Neal RM (1993) Probabilistic inference using Markov chain Monte Carlo methods. Department of Computer Science, University of Toronto, Toronto
- Oh SH, Kwon BD (2001) Geostatistical approach to Bayesian inversion of geophysical data: Markov chain Monte Carlo method. *Earth Planets Space* 53(8):777–791
- Pender D, Patidar S, Pender G, Haynes H (2016) Stochastic simulation of daily streamflow sequences using a hidden Markov model. *Hydrol Res* 47(1):75–88
- Zhang J, Man J, Lin G, Wu L, Zeng L (2018) Inverse modeling of hydrologic systems with adaptive multifidelity Markov chain Monte Carlo simulations. *Water Resour Res* 54(7):4867–4886

Markov Random Fields

Uma Ranjan

Indian Institute of Science, Bengaluru, Karnataka, India

Definition

Most physical phenomena result in structures which have short-range continuity and long-range changes. Such structures are seen in the distribution of rock structures, soil characteristics, temperature and humidity patterns, ore distributions, etc. Even within areas of uniformity or same rock type, there are changes which are random and do not have a spatial structure. Markov random fields are an effective way of characterizing the joint probability distribution of such structures.

Markov random fields are typically defined over a discrete lattice, with an associated neighborhood structure. The neighborhood structure dictates how the value at a lattice point is influenced by the values at nearby sites. Hence, a Markov random fields can also be viewed as a graph whose nodes correspond to the lattice points and whose edges correspond to the neighborhood structure. Such a model, also known as a graph model, combines the power of stochastic models with the computational ease and visualization offered by graph models. MRF are used in several areas ranging from computer graphics to computer vision, classification of images to generating textures. MRFs form an essential part of inverse problems, where characterizing a solution as an MRF allows attractive computational benefits to estimating joint probability distributions.

Introduction

Markov random fields are mathematical models used to solve problems of estimation of quantities. Mathematical models of acquisition of images are well-studied. Digital images of terrain, rock, or earth samples are obtained through different modes of acquisition such as synthetic aperture radar, electron microscopy, or micro-CT. Each of these use certain properties of the object such as surface reflectance, scattering, or absorption to create a digital image, typically representing the spatial distribution of a certain quantity on a discrete lattice. Inverse problems in image processing are those where the properties of an object are inferred from the images. For instance, a camera may record intensities that are reflected, but these observations are modified by blurring and noise. Hence, the true reflectance is the parameter to be estimated from the noisy observations. In satellite imagery, one is interested in identifying the vegetation type at each pixel, which represents

a spatial location. In images of rocks, one is interested in identifying different segments such as rock and pore and sub-types of rocks. Most problems of interest in geosciences are inverse problems.

Inverse problems suffer from the problem of ill-posedness. According to Hadamard (1923), a well-posed solution is one which is possible, accurate, and smooth (shows gradual variations for small variations of input). Most natural systems are smoothly varying; hence, this is a reasonable condition to impose on a solution. Moreover, smooth solutions are easier to handle computationally and represent stability with respect to variations of input, a significant advantage in engineering systems. However, a number of factors beyond our control cause a problem to become ill-posed. Limitations of acquisition systems, different types of noise, and functions used for modeling may cause solutions to be non-smooth. One way to make a problem well-posed is through the use of additional constraints of smoothness which force the solutions to correspond to our assumptions of a well-behaved system. This process is called regularization.

Markov random fields can be represented through probability distributions. This leads to powerful frameworks such as Bayesian inference which can be used to accurately estimate the distribution of the unknown quantities.

Bayesian Inference

One of the most popular methods to obtain regularized solutions to inverse problems is through a probabilistic framework. The variables of interest are assumed to follow a choice of probability distributions.

The notion behind the use of probabilistic distribution is the assumption that the underlying quantity is a well-posed approximation of what is actually seen or obtained from a numerical solution. This uncertainty is modeled as a random variable and forms the fundamentals of the Bayesian approach to inverse problems.

The quantity which needs to be modelled is estimated as a parameter of the probability distribution. In the context of image restoration, the parameter to be estimated is the actual (de-noised) intensity value at each pixel. In an image segmentation problem, the parameter is the class to which each pixel belongs.

The four aspects of Bayesian inverse solutions are:

1. Prior probability distribution
2. An estimated likelihood of the quantity of interest
3. Posterior probability, obtained as a product of the prior probability and the likelihood
4. A loss function, which penalizes an error in the estimate of the parameter of interest

The prior, likelihood and posterior probability are related through the Bayes theorem. If f is the quantity that is to be estimated and g are the observations (such as the acquired digital image), we have

$$Pr\{f|g\} = Pr\{g|f\}Pr\{f\}/Pr\{g\}$$

where $Pr(\cdot)$ refers to the probability function. $Pr\{f\}$ is the prior, reflecting our prior knowledge of the quantity to be estimated. It represents our belief about the distribution of the quantity of interest prior to observing the actual data. The prior probability is defined on the space of values which the quantity to be estimated can take.

$Pr\{g|f\}$ is the likelihood that the observed data fits the prior probability. This represents a “goodness of fit” of the assumed prior with the actual observations. Since the assumed prior is a model of the acquisition system, $Pr\{g|f\}$ is the likelihood function which reflects the fit of the observations to the model. This is also referred to as the degradation model.

The quantity $Pr\{f|g\}Pr\{g\}$ represents the posterior probability distribution. In usual practice, the normalizing constant $Pr\{g\}$ is neglected, since all observations are assumed to be equally likely. Hence, the conditional probability $Pr\{g|f\}$ is often referred to as the posterior probability.

The posterior probability distribution represents an update of our belief about the distribution of the quantity of interest over the lattice. Since the value to be estimated at each point is considered to be a random variable, its actual value is estimated as a parameter of the posterior distribution.

This estimate depends on the loss function that is chosen. The loss function indicates the penalty of choosing a parameter value different from the true value. Different loss functions lead us to different estimators.

For example, minimizing the squared loss function leads to estimating the posterior mean. Minimizing the zero-one loss leads to an estimate of the mode of the posterior (also known as the maximum a posteriori estimate (MAP)). If no information about the prior is available, only the likelihood is maximized, leading to maximum likelihood (ML) estimate.

Ripley (Ripley 1988) showed that the MAP estimate minimizes a variational energy which consists of two terms – one related to data fidelity and the second to a smoothness. The smoothness condition acts as a regularizer. The desired solution corresponds to the minimum of such an energy.

In image processing applications, the posterior probability needs to be estimated over the entire pixel array. When the image sizes are large as in most applications, the posterior probability is computationally intractable. However, when the posterior probability can be modeled as a combination of local interactions, it simplifies the computation of the posterior. This is captured in the notion of a Markov random field, where the assumption is that the information at a point is

dependent only on the information in a neighborhood of a fixed size around a point.

Markov Random Field

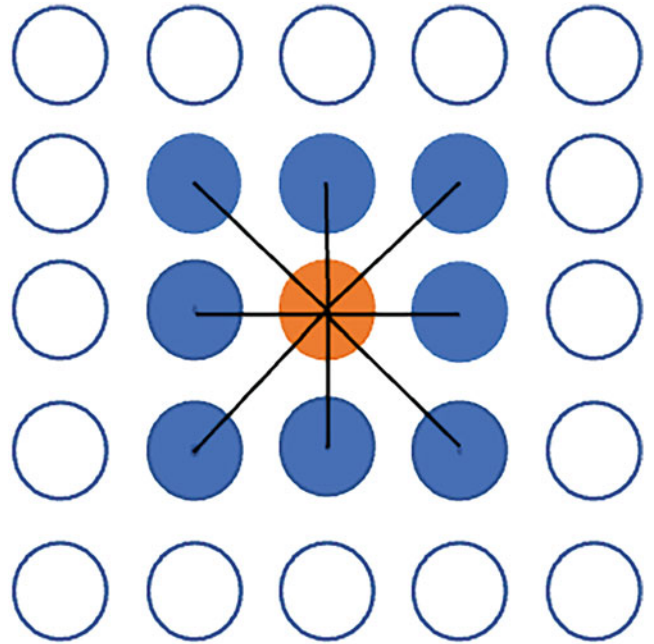
If X_s is a quantity of interest (such as intensity, color, texture) that is to be estimated at pixel s in an image S , the Markov property states that the conditional probability of value at a pixel s , estimated from the knowledge of the entire lattice, is the same as the probability of its value, estimated from its neighbors.

$$Pr\{X_s = x_s | X_t = x_t, t \in S\} = Pr\{X_s = x_s | X_t = x_t, t \in N(s)\}$$

X_s is a random variable at pixel location s , and x_s is a specific value that X_s can take. The Markov property thus states that the probability of the value at a pixel s in the image, given the value of X over the rest of the image, is the same as the probability of x_s , given only its neighbors.

The MRF is then a model of the quantity, together with a probability measure which represents the distribution of the quantity of interest. The probability measure is assumed to take values from a pre-determined set $x_s \in \Omega$. These values represent the digitized values of the property (such as color, texture, gradient, or more complex quantities such as class to which the pixel belongs).

The notion of a local neighborhood induces a connectivity structure. For example, in Fig. 1, the central orange pixel is connected to eight neighbors. This configuration is referred to



Markov Random Fields, Fig. 1 Eight neighborhood

as an eight-neighborhood structure. This neighborhood structure is commonly considered to be symmetric, i.e. x is a neighbor of y , and then y is a neighbor of x . Hence, an MRF can also be represented as an undirected graph S, G , where S consists of the set of nodes and G represents the set of edges between the nodes. The graph structure enables the use of efficient computational methods such as graph coloring to perform segmentation tasks on images.

The blue lattice cells are the neighbors of the central orange lattice cells, and the black edges denote the connectivity structure.

The nondirectional nature of the graph representation does not mean that the underlying process which gave rise to the observations is nondirectional. In some cases, the observations may correspond to an underlying drift or diffusion process which could have a direction. The Markov random field in such a case represents the (final) evolutionary state of the underlying process. Hence, an MRF could also be viewed as a snapshot in time of an underlying evolutionary process. This viewpoint has been used in some cases to define hidden factors which correspond to actual physical entities and their evolution, and these factors could in turn influence the observed MRF.

This formulation of the MRF, however, does not directly give us a procedure to actually construct a probability distribution which satisfies the MRF property. A practical way to construct an MRF is given by the Clifford-Hammersley theorem of MRF-Gibbs equivalence:

If an MRF also satisfies the following properties, then it has a joint distribution which is a Gibbs distribution (Rangarajan and Chellappa 1995):

- Positivity: $Pr\{X = x\} > 0$
- Locality: $Pr\{X_s = x_s | X_t = x_t, t \in S\} = Pr\{X_s = x_s | X_t = x_t, t \in N(s)\}$
- Homogeneity: $Pr\{X_s = x_s | X_t = x_t, t \in N(s)\}$ is the same for all sites s in the lattice.

Positivity implies that the values at all pixels are positive, locality is the Markovian property, and Homogeneity implies that the probabilities are independent of spatial location.

The Gibbs distribution is a global property over the entire lattice, given by the following property:

$$Pr(X) = \frac{1}{Z} \exp -\frac{1}{T} U(X)$$

f is a configuration in the sample space, and Z is a normalization constant given by

$$Z = \sum_{X \in S} \exp -\frac{1}{T} U(X)$$

The energy $U(X)$ over the entire configuration X can be factored into energies over smaller, localized set of nodes called cliques.

A clique is defined by a neighborhood configuration in such a way that all the sites in the cliques are neighbors of each other. This offers attractive possibilities of creating computationally efficient means of estimating the MRF. Figure 2 illustrates the possible cliques of two sites in a four-neighborhood and eight-neighborhood of a regular lattice. In the four-neighborhood case, two-site cliques can only be of two orientations, while in the eight-neighborhood case, there are four possible orientations. It is to be noted that Fig. 2b shows only the unique orientations.

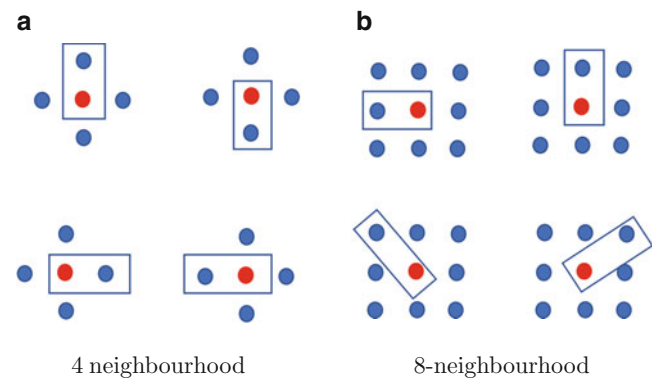
Each clique c is associated with an energy V_c , such that the energy of the entire configuration is given by the sum of the clique potentials:

$$\sum_c V_c(X)$$

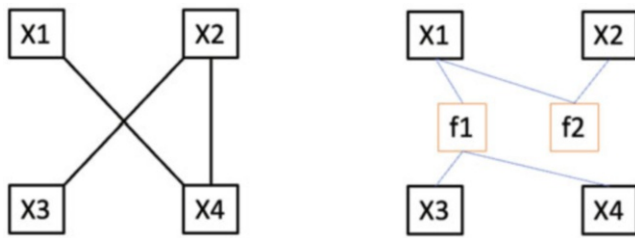
Gibbs distributions are widely used in statistical physics and have the advantage that they correspond to the minimum value of an energy functional. Thus, the problem of estimating the maximum a posteriori (MAP) estimate is now converted to the problem of estimating the minimum of an energy functional. This enables us to define the energy functional using contextual spatial knowledge for the specific problem.

Factor Graphs

A modification over the MRF is when the variables are not directly observed through a measurement but are inferred from other variables that are observed. For example, fault planes, lithospheric boundaries, or mineral distributions are not directly observed but derived from intermediate variables



Markov Random Fields, Fig. 2 Two-node cliques in different neighborhoods



Markov Random Fields, Fig. 3 MRFs vs. factor graphs. (Reproduced from Bagnell 2021)

such as Potts spins. These intermediate variables are known as latent variables. Latent variables are modeled through the use of factor graphs. Factor graphs make explicit use of latent factors to represent unobserved variables. The neighborhood structure in MRF and Gibbs fields does not specify whether the neighborhood pixels must all be considered simultaneously or pairwise. Factor graphs resolve this ambiguity. In Fig. 3 (reproduced from Bagnell 2021), it can be seen that the ambiguity related to diagonal pixels is easily resolved in the factor graph representation.

Solutions of MRF

Finding the optimal joint distribution of an MRF which minimizes a cost function is in general intractable.

The Gibbs distribution, while offering a way to compute distributions over a local neighborhood, includes a normalization constant which is NP-hard to compute, since it involves a summation over all possible clique energies. However, a number of approximations exist which sample the Gibbs distribution in a manner which does not need a normalization.

An efficient way to estimate the Gibbs distribution on a discrete lattice is through the Gibbs sampler, which is a special case of a Markov chain. This is the basis of the Markov chain Monte Carlo (MCMC) method.

The choice of the prior distribution also has a role in determining the complexity of the solution. In general, it is desirable that the prior and posterior probability distributions have the same functional form. This is guaranteed through the use of conjugate prior functions for the given likelihood function. Normal distributions are a popular choice to ensure conjugate priors.

Some of the methods that have been used for the MAP estimate are:

Simulated Annealing (Geman and Geman 1984): The Gibbs sampler was used to obtain the global minimum of

the energy function which corresponded to the maximum a posteriori estimate. The energy functional is defined over an entire image configuration, and the Gibbs sampler traverses the state space, accepting configurations of a lower energy according to a probability dependent on the temperature. The temperature follows an annealing schedule, which allows the Gibbs sampler to converge in probability to the global optimum solution. This method, while giving very good results, is extremely slow. Some of the later modifications (Ranjan et al. 1998) have focused on reformulating the state-space to accelerate convergence.

Iterated Conditional Modes (Besag 1986): An iterative algorithm was presented, where the value at a point is updated one parameter at a time, while holding the other parameters fixed. This corresponds to a local steepest descent of the conditional probability. This results in a local minimum solution of the posterior probability and has the disadvantage of being sensitive to initial conditions. However, its rate of convergence is fast.

The highest confidence first (HCF) method introduces a confidence measure (stability) for each node. The algorithm starts with a null state at each pixel and terminates at a configuration which minimizes the variational energy. At each step, the stability of a node is computed as an estimate of the local posterior function, i.e., as a combined measure of the observable evidence and a priori knowledge about the preferences of the current state over the other alternatives. The ICM updates only the node with the highest stability. A higher stability denotes an estimate in higher consistency with the observations and a priori knowledge. In this way, the HCF algorithm makes maximal progress towards a final solution based on current knowledge about the MRF.

Belief Propagation: In belief propagation, the interactions between the nodes are modeled as messages being passed between them. The messages represent a state of “belief” of each pixel about the marginal distribution of its neighbor. The messages are updated in a single direction, and the marginal distribution is estimated from a normalized set of beliefs after the messages have converged.

A number of improvements to accelerate convergence and improve memory efficiency have been proposed in recent literature.

Multidimensional dependencies and inter-class relationships have been modeled in the context of geosciences through Markov chain random fields (MCRF) (Li 2007). These are a special category of Markov Random Fields (MRF) where the neighborhood structure is not fixed but may vary in different directions.

Although initial use of the MRF formulation in geosciences was focused on images, it is being increasingly used in more general field settings such as characterization of using the Rayleigh scatter of the Earth's microseismic field (Tsukanov and Gorbatnikov 2018). Thus, the MRF formulation and its accurate solutions are central to developing an increased understanding of the structural and environmental properties of the Earth.

Summary and Conclusions

Spatial data such as images can be modeled through Markov random fields. Obtaining an MRF that satisfies optimality conditions together with constraints that contribute to regularized solutions is a rich field. Many of the approaches that have been used for estimating MRFs have also been presented here. Many of these estimations use stochastic optimization methods to find global minimum solutions.

Recently, deep learning methods have been combined with Markov random field models to provide deterministic solutions to inverse problems. These frameworks provide the advantage of computational scaling along with providing contextual information and capturing higher-order relationships.

MRFs are hence an active area of research in the foreseeable future.

References

- Besag J (1986) On the statistical analysis of dirty pictures. *J Royal Stat Soc* 48(3):259–302
- D Bagnell (2021) Gibbs Fields and Markov Random Fields. In: Statistical techniques in robotics (16-831, F10), Carnegie Mellon University. http://www.cs.cmu.edu/~16831-f14/notes/F11/16831_lecture07_bneuman.pdf. Accessed 24 Sept 2021.
- Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell* 6:721–741
- Hadamard J (1923) Lectures on Cauchy's problem in linear partial differential equations. Yale University Press, New Haven
- Li W (2007) Markov chain random fields for estimation of categorical variables. *Math Geol* 39:321–335
- Rangarajan A, Chellappa R (1995) Markov random field models in image processing. In: Arbib M (ed) *The handbook of brain theory and neural networks*. MIT press, Cambridge
- Ranjan US, Borkar VS, Sastry PS (1998), Edge detection through a time-homogeneous Markov system. *J Indian Inst Sci* 78:31–43
- Ripley BD (1988) *Statistical inference for spatial processes*. Cambridge University Press, Cambridge
- Tsukanov AA, Gorbatnikov AV (2018) Influence of embedded inhomogeneities on the spectral ratio of the horizontal components of a random field of rayleigh waves. *Acoust Phys* 64:70–76

Marsily, Ghislain de

Craig T. Simmons

College of Science and Engineering & National Centre for Groundwater Research and Training, Flinders University, Adelaide, SA, Australia



Fig. 1 Professor Ghislain de Marsily. (Photo by Antoine Meyssonnier 2014)

Biography

Professor Ghislain de Marsily is an internationally renowned scientist famed for his contributions to groundwater hydrology and water management. A highly active member of the French Academy of Sciences, he is currently Professor Emeritus at both the Sorbonne University (Pierre et Marie Curie) and Paris School of Mines, France.

Career Overview

Graduating from the École des Mines as an engineer in 1963, de Marsily went on to complete a doctoral degree in science at Pierre et Marie Curie in 1978. He then proceeded to build a long and distinguished career in hydrology characterized by energy, innovation, and technical mastery. As a hydrologist, Professor de Marsily has researched areas including basin analysis, fluid mechanics, solute transport, waste disposal, river ecology, surface hydrology as well as hydrogeology, water resources management, and global food production. He is a pioneer in the development of stochastic hydrogeology and inverse methods and is the originator of the “hydrogeological national parks” concept to protect the world's dwindling supplies of groundwater. He has collaborated extensively to study groundwater systems throughout the world and has been influential in exploring the hydrology of North Africa together with his African colleagues. He has also worked in India with his former students.

In addition to his personal science, de Marsily has been internationally important in his capacity to build networks and collaborations to address global issues at local levels. In 1989, he established the UMR CNRS SISYPHE, an interdisciplinary research unit involving the National Centre for Scientific Research and the University Pierre and Marie Curie, directing the initiative until 2000. From 1989 to 1999, he established and directed the PIREN-SEINE program at the National Centre for Scientific Research, studying the hydrology of the whole of the Seine catchment area. From 2000 to 2004, he was the founder and director of the Postgraduate School of Geosciences and Natural Resources at Pierre et Marie Curie. From 1994 to 2006, he was a member of the national commission which evaluated research into the management of radioactive waste in France.

Scientific Contribution

Professor de Marsily has made significant and sustained contributions to mathematical geosciences and porous media flow and modeling. He is recognized as a pioneer in geostatistics, groundwater modeling, and approaches for dealing with hydrogeologic spatial heterogeneity. He worked closely with geostatistician Georges Matheron, particularly on a paper dealing with transport in porous media (Matheron and de Marsily 1980). Their paper showed, for the very first time that for the special case of a stratified porous medium with flow parallel to the bedding, that solute transport cannot, in general, be represented by the usual convection-diffusion equation, even for large time. They went on to demonstrate that when flow is not exactly parallel to the stratification, diffusive behavior will eventually occur at large times. They prompted and emphasized the need for further work on the mechanism of solute transport in porous media. This paper was an extraordinary breakthrough triggering a major line of critical scientific inquiry that is of profound importance to groundwater science and to mathematical geosciences to this day. It has altered our understanding of groundwater flow and solute transport behavior, demonstrating for the first time the ways in which diffusion and dispersion act to control those processes. The Matheron and de Marsily (1980) paper has become a seminal work in hydrogeological science. Cited 731 times (Google Scholar 25 March 2021), it remains one of the most cited papers in groundwater science to this day.

de Marsily's impact extends well beyond his own research. He has been a driving force in communicating and teaching groundwater science to students and professionals around the world. His textbook *Quantitative Hydrogeology: Groundwater Hydrology for Engineers* (de Marsily 1986) is a classic text in groundwater science. It is a unique book, placing great emphasis on both qualitative and quantitative hydrogeology.

As noted in its preface: "the qualitative and the quantitative aspects are not treated separately but combined and blended together, just as geology and hydrology are woven together in hydrogeology." This was, and remains, a simple but profound observation that has done more to shape the field of hydrogeology than many research papers. It reflects de Marsily's keen interest and success in understanding concepts and to ensure that they are treated physically, mathematically, and philosophically in a robust and rigorous manner. The textbook is still in its first and only edition and remains in print to this day, with 3161 citations (Google Scholar 25 March 2021) – and growing.

Professor de Marsily remains an energetic, inspirational, and provocative scientist, a visionary thinker, and a world leader in groundwater research. He has inspired, mentored, and taught generations of researchers and students while making a remarkable impact on all aspects of hydrogeology research, education, policy, and management. Emphasizing the vital relevance of science to society, he has both made and inspired important contributions to some of the most critical environmental debates of our time, including critical planetary-scale insights into water, climate change, food and population growth, and nuclear waste disposal.

Awards and Honors

Recognizing his outstanding contributions to hydrogeology, including his legendary lectures and numerous books on hydrogeology, fluid transport, and geostatistics, de Marsily has received many awards, including the O. E. Meinzer Award of the Geological Society of America, Robert E. Horton Medal of the American Geophysical Union, and the President's Award of the International Association of Hydrogeologists. In addition to these honors, de Marsily is Chevalier de la Légion d'Honneur, a Member of the French Academy of Sciences, French Academy of Technologies, French Academy of Agriculture, and a foreign Member of the US National Academy of Engineering. In recognition of his scientific achievements and significant contributions to hydrogeology, de Marsily was awarded the degree of Doctor of Science *honoris causa* from Flinders University (Australia), University of Neuchâtel (Switzerland), and University of Québec (Canada).

Bibliography

- de Marsily G (1986) Quantitative hydrogeology. Groundwater hydrology for engineers. Academic, New York, 440p
- Matheron G, de Marsily G (1980) Is transport in porous media always diffusive? A counterexample. Water Resour Res 16(5):901–917

Mathematical Geosciences

Qiuming Cheng

School of Earth Science and Engineering, Sun Yat-Sen University, Zhuhai, China

State Key lab of Geological Processes and Mineral Resources, China University of Geosciences, Beijing, China

Definition

Mathematical geosciences or geomathematics (MG or GM) is an interdisciplinary field involving application of mathematics, statistics, and informatics in earth sciences. The International Association for Mathematical Geosciences (IAMG) is a unique international academic society with mathematical geosciences as part of its title and the mission of which is to promote, worldwide, the advancement of mathematics, statistics, and informatics in the Geosciences. However, the lack of a unified definition of mathematical geosciences or geomathematics (MG or GM) as an interdisciplinary field of natural science may lead to misunderstanding of the subject or not even treating it as an independent discipline. This has, to some extent, affected the development of the field of mathematical geosciences. Since late 1990s, the author of the current chapter has served IAMG as capacities of, council member, vice president, and president. Thus, the author has witnessed and contributed to the transformation of IAMG from Mathematical Geology to Mathematical Geosciences, as well as various updates of IAMG associated publications including journals and conferences proceedings. In 2014, in his IAMG president column in the newsletter, the author discussed about a definition and connotations of mathematical geosciences as well as demonstration of contributions of mathematical geoscientists and frontiers of mathematical geosciences. Mathematical Geosciences was defined as an interdisciplinary scientific subject that integrates mathematics, computer science, and earth science, and it is the study of the mathematical characteristics and processes of the earth (and other planets) and the prediction and analysis of its resources and evaluation of environment.

Introduction

The International Association for Mathematical Geosciences (IAMG) is a unique international association with the term of Mathematical Geosciences as its title and the mission of promoting application of mathematics, statistics, and informatics in earth science. IAMG was initially established with its title of International Association for Mathematical Geology in 1968 at the 23rd International Geological Congress

(IGC) and in affiliation to the International Union of Geological Sciences (IUGS) and the International Statistical Institute (ISI). In addition to Mathematical Geology, Geomathematics and Geostatistics were also suggested to be used as its name. Some books and series of monographs published by the IAMG members in the field of mathematical geology also used these names. For example, the books authored by Frits Agterberg and published by Springer use geomathematics as book titles: the book “Geomathematics: Mathematical Background and Geo-science Applications” in 1974 and the book “Geomathematics: Theoretical Foundations, Applications and Future Developments” in 2014. Since mathematical geology could be literally interpreted as a branch of geology, some government Natural Science Foundations also group mathematical geology as a subdiscipline of geology. With the continuous development in the field, the applications of mathematical theory and methods as well as relevant software systems are no longer limited to geology. They are also widely used in other fields of geosciences including but not limited to hydrology, geophysics, and geochemistry. To serve the community better and to represent the complete connotation of the discipline, during the General Assembly of IAMG at the 33rd IGC held in Oslo in 2008, the name of the Association was changed from International Association for Mathematical Geology to International Association for Mathematical Geosciences with the same abbreviation IAMG. Accordingly, it revised the names of journals, for example, the flagship journal “Mathematical Geology” was renamed as “Mathematical Geosciences.” Thus, the three journals of IAMG that time include *Mathematical Geosciences*, *Computers & Geosciences*, and *Natural Resources Research*. Since then, these journals and IAMG annual conference proceedings have published works with more variety of mathematical geoscience topics. The name change of the association must be important in broadening and expanding the application fields of mathematical geosciences. One example worth mentioning is that IAMG established an affiliation with the International Union of Geodesy and Geophysics (IUGG). In this way, the international alliance to which IAMG is affiliated has expanded to include the International Union of Geological Sciences (IUGS), the International Union of Geodesy and Geophysics (IUGG), and International Statistics Institute (ISI). These three international organizations are large and active union members of the International Science Council (ISC). They cover most of the disciplines of earth sciences, mathematics, and statistics. Establishing affiliations and actively engaging in activities with these Unions have broadened the IAMG’s international collaboration which in turn enhanced its international visibility and influence in the mainstream earth science community. Now IAMG fully represents a branch of geosciences rather than only geology.

It should be noted that Mathematical Geosciences (or Geomathematics), as a relatively young earth science discipline, still has not been widely accepted by the mainstream field of geosciences and is even often ignored. On the one hand, it may be due to the organization of the association which has relatively small number of registered members. On the other hand, it may be because the society does not have a clear subject definition and connotative boundaries, especially the unclear subject boundaries with other related disciplines. Without a clear definition of the Mathematical Geosciences subject, those who indeed work in the fields of mathematical and statistical geosciences are not considered as mathematical geoscientists rather as geodesists, geophysicists, etc., and their work and academic contributions are not regarded as mathematical geosciences. For example, the section of this chapter will introduce an internationally renowned scientist Professor Dan McKenzie of Cambridge University in the United Kingdom. He is a physicist and mathematician by education and training. In the 1960s, he combined mathematical thermodynamics with geoscience and systematically established thermal structural models for describing plate tectonics and mantle dynamics. Because of his pioneering and groundbreaking work in the formation of plate tectonics theory, Professor Dan McKenzie is praised as one of the four major contributors to the development of the modern plate tectonics theory (<https://geolsoc.org.uk/Plate-Tectonics/Chap1-Pioneers-of-Plate-Tectonics/Dan-McKenzie>). McKenzie's work is well known in the fields of geophysics and tectonics, and he is also considered as a geophysicist rather than a mathematical geoscientist.

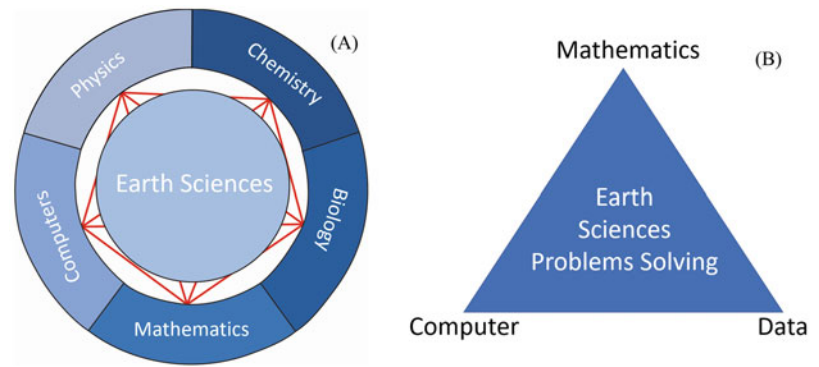
Although several descriptive disciplinary definitions and terminologies about mathematical geosciences have been proposed in the literature, there is still a lack of systematic, inclusive, and relatively standard disciplinary definition. For example, mathematical geosciences are often referred to simply as the application of mathematical and statistical methods in geology (or in earth science), or the application of geological data analysis and quantitative prediction model development (Howarth 2017). The IAMG website once defines the mission of IAMG as: to promote the development and application of mathematics, statistics, and informatics in earth sciences. In the author's view to define Mathematical Geosciences (MG) as a formal subject, or to simply define it as the application of mathematics, statistics, and informatics in the Earth sciences, is a fundamental question that has a vital impact on the development of the subject. For this reason, during his presidency of IAMG, the author of current chapter attempted to propose a definition of mathematical geosciences. It was published in as president forum on Newsletter (Issue 76–79) (Cheng 2014). Here I will elaborate on the definition of MG and give several examples to demonstrate the contributions MG made to geosciences.

What Is Mathematical Geoscience (MG)?

As mentioned earlier, Vistelius (1962) gave an earlier definition of mathematical geology and used it as the name of the Association. Geostatistics is a remarkable new subject of mathematical geology developed by IAMG scholars. Now the geostatistical theory and methods have been used not only in earth sciences but also in many other scientific fields. Geostatistics was once refereed as the application of statistical methods in geology or other geosciences (e.g., Merriam 1970; McCammon 1975), and in many cases this simple definition seems to be still in use. As mentioned in previous section, the term Geomathematics has also been used by several authors including Agterberg (1974), who used the term as the title of his two books (Agterberg 1974, 2014). Since the IAMG changed its name from Mathematical Geology to Mathematical Geosciences in 2008, the term Mathematical Geosciences has often appeared in IAMG's literature including books, conference proceedings, and journals. The difference between mathematical geology and mathematical geosciences is not only in terms but also in the connotation and scope of the subject. If mathematical geology is a branch of geology, mathematical geosciences should be an interdisciplinary subject of geological sciences, and it must include mathematical geology as one of the branches of mathematical geosciences. Compared with geology, geosciences also cover other related disciplines including but not limited to geochemistry, geophysics, geobiology, and hydrology. Mathematical geosciences should be a branch of geosciences parallel to other cross-disciplines in earth sciences such as geochemistry, geophysics, and geobiology, rather than a branch of geology (Fig. 1). In the author's opinion, this distinction is very important for the development of the subject. For example, under the concept of mathematical geology, the subject is limited to the application of mathematics, statistics, and informatics in geology, but as a subject of mathematical geosciences, similar to geochemistry and geophysics, MG serves the entire geosciences. So, what should be the definition of mathematical geosciences or geomathematics? What role should mathematical geosciences play in the earth science family? What are the major contributions of mathematical geoscientists to the development of earth science? What are the frontiers of mathematical geosciences today? In the future, how will mathematical geosciences develop in the context of the rapid development of big data, artificial intelligence, and other new mathematical theories? These questions are of interest to mathematical geoscientists who must provide answers, especially for the younger generation of students and researchers in geosciences. The following is a summary of the author's thoughts during the past years about the above questions related to mathematical geosciences.

Mathematical Geosciences,

Fig. 1 Schematic diagrams showing the relationship between natural sciences and earth science (a) and how mathematical geosciences works (b)



To give a definition of mathematical geosciences, we might as well understand the situation of other related interdisciplinary subjects, such as geochemistry, geophysics, and geobiology.

According to the descriptions in Collins English Dictionary, Geophysics is referred as “a science of studying the earth’s physical properties and the physical processes acting upon, above, and within the Earth.” Similarly, Geochemistry is a science that deals with the chemical composition of and chemical changes in the solid matter of the earth or a celestial body (Unabridged dictionary). Finally, Biogeosciences is referred as an interdisciplinary field of integrating geoscience and biological science: the study of the interaction of biological and geological processes (Unabridged dictionary). The above definitions emphasize on interdisciplinarity of intersection of basic natural sciences and earth sciences. Similarly, we can define mathematical geosciences as an interdisciplinary scientific subject that integrates mathematics, computer science, and earth science, and it is the study of the mathematical characteristics and processes of the earth (and other planets) and the prediction and analysis of its resources and evaluation of environment (Cheng 2014). The core issues of the new definition are the characteristics and processes of the earth. The concepts of chemical and physical properties of the earth, and chemical, physical, and biological processes of the earth are relatively easy to understand and so do the definitions of geochemistry, geophysics, and geobiology. But what are the mathematical characteristics, properties, and processes of the Earth? And why is the prediction and evaluation of the earth’s resources and environment inherently related to mathematics? These are key questions to answer to convince that mathematical geosciences can be regarded as an independent but interdisciplinary discipline which distinguishes it from other related disciplines. So, let’s first review the main differences among mathematics, physics, and chemistry. Most people started to learn these basic natural courses in elementary or middle schools. Chemistry is mainly about studies of composition, properties, structure, and change laws of matter at the molecular and atomic microscopic level; physics is mainly concerned with studies of physical properties and processes

of matter, including the most general laws of motion and basic structure of matter; and mathematics can be referred as studies of quantity, structure, change, space, information, etc. The integrations of these basic disciplines and earth sciences form various interdisciplinary geoscience subjects such as geochemistry, geophysics, geomathematics, and geobiology. Actual integration of multiple branches of basic sciences such as chemistry with earth science has formed various sub-disciplines of geochemistry, including major element geochemistry, trace element geochemistry, stable isotope geochemistry, and isotope chronology. Various analytical and testing techniques developed in the field of chemistry have been adopted into geochemistry such as mass spectrometry, fluorescence analysis, and laser ablation imaging technique. Similarly, multiple branches of physics are integrated with earth sciences to form multiple geophysical directions, such as gravity, magnetism, electricity, semiology, radioactivity, superconductivity, and other geophysical directions. Accordingly, various types of physical detection and survey technologies have been developed and applied in geosciences, including gravitational measurement, nuclear magnetic resonance, magnetic measurement, seismic measurement and electrical measurement. Similarly, various mathematical disciplines such as geometry, calculus, power spectrum analysis, morphology, topology, probability and statistics, and fuzzy mathematics can provide indispensable theories and methods for quantitative studies of relevant properties of the Earth. The characteristics and properties of the Earth that must be studied by means of mathematics include but not limited to the geometric characteristics, dynamic characteristics, the uncertainty of earth observation and measurement, the prediction error of earth events, and so on. These important issues related to the Earth are all mathematical problems in nature which cannot be studied without mathematics. Therefore, the integration of these subdisciplines of mathematics with geoscience forms different directions of mathematical geosciences. For example, the integration of probabilistic statistics and geosciences produces results in a new subject of geostatistics and the integration of calculus and geosciences creates a subject of

mathematical geodynamics. Theories and methods developed in Geometry can be used in earth science research to solve various geometric problems of the earth, such as the shape, coordinates, angle, length, volume, quantity, distance, and direction of the Earth. Many examples can be cited to demonstrate effective utilization of various mathematical theories and methods in solving geoscientific problems. In the next section, some selected examples will be introduced to show how mathematical geoscientists have played an important role in earth science research, what significant contributions that mathematical geoscientists have made in the development and advancement of modern earth science theories and solving applied problems related to sustainable development, and how they will continue to play an important and irreplaceable role in the innovation of earth science in the future.

Examples of Important Contributions of Mathematical Geoscientists to Earth Science

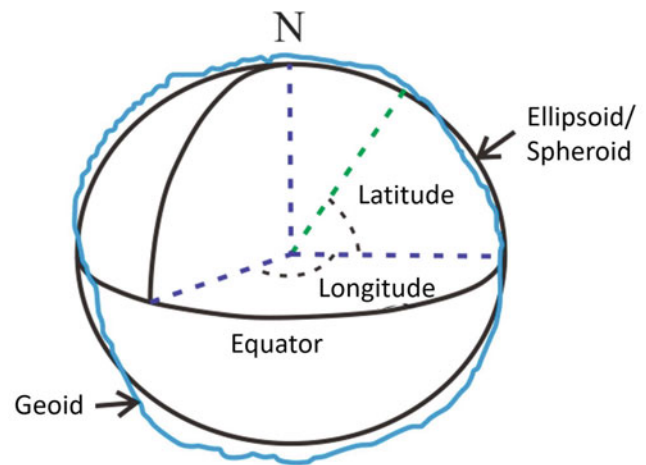
There are many examples to show how mathematical geosciences and mathematical geoscientists have made indispensable and fundamental contributions to the development of modern earth sciences. Here just name a few examples.

Earth Geometry

The geometric shape of the Earth is an important property of the Earth which serves the base for quantitative earth science. Mathematical models of the shape of the earth, such as Clark ellipsoid and Hayford ellipsoid, are the basic geodetic models for positioning and navigating systems (such as GPS and BDS) and other types of remote sensing technology (RS) for spatial measurement and spatial analysis. The shape of the earth is irregular and complicated, but in order to establish an earth coordinate system, a mathematical model needs to be used to approximately represent the shape of the earth, which is the earth ellipse (Ellipsoid or Spheroid) (Fig. 2). The equation of the ellipsoid in the Cartesian coordinate system is:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 \quad (1)$$

where a and b are the equatorial radius along the x and y axis, and c is the polar radius along the z axis. These three parameters determine the shape of the ellipsoid, including the flatness or flatness rate of the earth's ellipse. To elaborate on the relationship between the ellipse and the Geoid is beyond the scope of this chapter. The point worth mentioning is that the shape of the Earth is indeed a mathematical feature of the Earth. As a matter of fact, the first to establish the earth ellipse model and name it Geodesy was the British scholar Alexander Ross Clarke (1828–1914). He is an internationally renowned mathematician and geodesist. His important contributions



Mathematical Geosciences, Fig. 2 The mathematical model of the geometric shape of the earth as an ellipsoid

include the calculation of the British Geodetic Triangulation (1858), the calculation of the shape of the Earth (1858–1880), and the publication of a monograph on Geodesy (1880). This is an excellent example done by a pioneer mathematical geoscientist for modeling the geometrical property of the Earth.

Age of the Earth

The age of the Earth is a curious question of the earth science. Before the invention of isotope geochemistry there was no consistent answer about the age of the Earth. One of the well-known estimates of the age of the earth was given by William Thomson, 1st Baron Kelvin (1824–1907), a British mathematician, mathematical physicist, and engineer. Kelvin estimated the age of the Earth by investigating the Earth's cooling and making historical inferences of the Earth's age from his calculations. Kelvin calculated how long it would have taken Earth to cool if it had begun as a molten mass. Kelvin made several assumptions about the earth to set his conductive model for estimating the age of the Earth. These assumptions include that the Earth's heat was originally generated by gravitational energy; for a short period of time, say 50,000 years the Earth cooled from a temperature of about 3700 °C to the present temperature of about 0 °C; since then, the average temperature of the Earth's surface has not changed significantly; and the interior of the Earth is solid and a secular loss of heat from the whole Earth through heat conduction. Kelvin established a heat conduction equation from which he derived an estimate of 20–40 million years of age of the Earth. While his estimate was proved wrong by the discovery of radioactivity and the radioactive dating technology which yields an estimate of 4.54 billion years old, his heat conduction model of the Earth based on observations and calculations was an accurate and similar conduction models have been applied in other fields of geodynamics including in

simulating heat flow processes in mid-ocean ridges as to be introduced in later section of the chapter. William Thomson is also well known for his work in the mathematical analysis of electricity and formulation of the first and second laws of thermodynamics. Kelvin's work in relevance to the topic of the current chapter is also due to the term of "Mathematical Geology" was used, might have been for the first time in the literature, as the title of review paper introducing Kelvin's work in Nature by Lodge (1894).

Geodynamics of the Plate Tectonics

The modern plate tectonics theory being advanced in the 1960s has become a profound revolution in the field of earth science, which provides scientific explanations about evolution of the Earth systems and formations and distributions of major geological events such as earthquakes, volcanoes, magma, and orogeny as well as minerals, water, and energy resources. It is believed that the theory of plate tectonics has completely revolutionized the global view of the solid earth and promoted the formation and development of the concept of the earth system sciences.

The pioneers who have made important foundational contributions to modern plate tectonics theory include but not limited to Dan McKenzie of Cambridge University in the United Kingdom, Jason Morgan of Princeton University in the United States, Professor Xavier Le Pichon of France and Professor Tuzo Wilson of the University of Toronto in Canada. Among them, Dan McKenzie is an eminent geophysicist and geomathematician. Professor McKenzie received a scholarship in mathematics and entered the natural sciences in university in his early age. He had systematically studied mathematics, physics, chemistry, and geology, and graduated with a doctorate at the age of 23. He had creatively applied Euler vector theorem and calculus to describe geodynamics of earth plate tectonics and achieved remarkable results which demonstrate the essential role of mathematics in the establishment of plate tectonics theory.

Mathematical Principles of Plate Tectonics

Dan McKenzie's most influential result was the paper he published in collaboration with Robert Parker (McKenzie and Parker 1967). This paper uses Euler's rotation theorem to prove the conjugate relationship of the three sides of rigid plates: mid-ocean ridges, subduction zones, and transform faults. Euler's rotation theorem shows that a rigid body (such as a plate) is displaced in three-dimensional space, and at least one point inside it is fixed. The displacement of a rigid body is equivalent to a rotation around a fixed axis containing the fixed point (Fig. 3). This achievement is praised by colleagues as the mathematical principle that defines plate tectonics on a sphere, the most theoretical work of the plate tectonics.

Mathematical Model of Mantle Convection

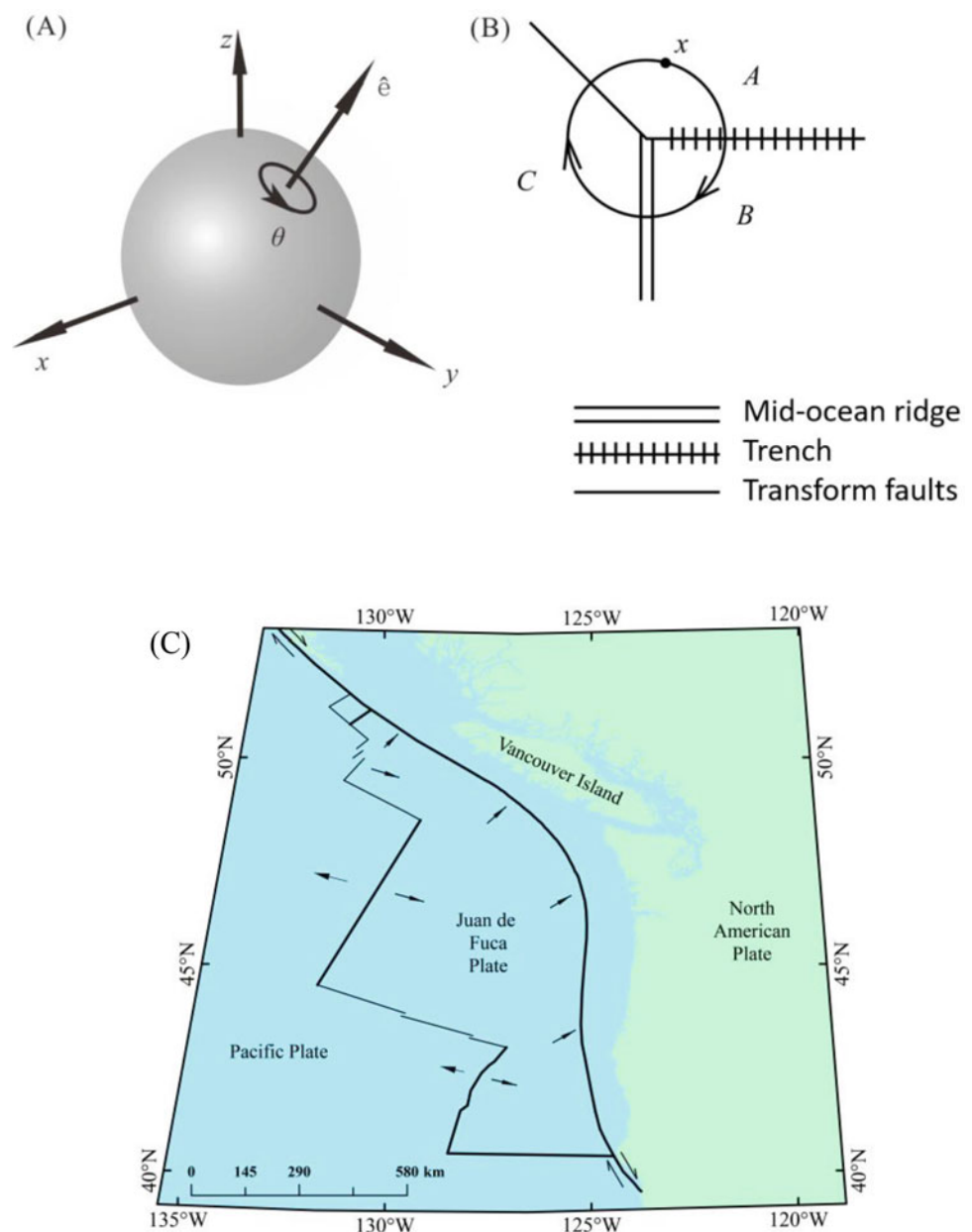
Another influential paper by McKenzie (1967) is the result of the heat flow and gravity anomalies in mid-ocean ridges. This model modified the seafloor expansion model established by the American scholar Harry Hess (1960) and established the mid-ocean ridge heat flow diffusion model and the mantle convection mathematical model. This work had laid the foundation for a series of geodynamic models subsequently developed in the field of plate tectonics for characterizing plate tectonics including thermal structures of plates, subducting slab dynamics and dynamics of basin formation. Figure 4 shows the parameters and boundary conditions of the McKenzie model. The plates on both sides of the mid-ocean ridge are regarded as constant rigid bodies with regular boundaries, with constant expansion speed (V) and gradual change of thickness (a), and the boundary conditions of temperature (T_1) on bottom boundary and temperature (T_0) on top boundary of the plate (lithosphere). Under some simplifications, the heat flux can be expressed by the following diffusion equation:

$$\frac{\partial[\rho C_P(T)T]}{\partial t} = \frac{\partial}{\partial z} \left(k(T) \frac{\partial T}{\partial z} \right) \quad (2)$$

where ρ the mass density of the lithosphere; T temperature; C_P specific heat capacity; t time; k thermal diffusivity; and z depth. The upper left picture in Fig. 4 shows the mid-ocean ridge heat flow diffusion model and boundary and initial conditions established by McKenzie (1967). The upper right picture in Fig. 4 shows the comparison between the actual measured heat flow results and the theoretical prediction results using the McKenzie model. The difference between the two results is relatively large. Many authors believed that the difference is due to the heat loss caused by the convection of hydrothermal fluid near the mid-ocean ridge. Starting from the complexity and fractal structure of the mid-ocean ridge lithosphere, the author of this book chapter substituted the lithospheric density in the McKenzie model with the fractal density proposed by the author and re-established the diffusion model. The re-established diffusion model significantly improved the results for prediction of heat flow (Cheng 2016). McKenzie and many later authors established basin formation models based on plate thermal diffusion models, slab thermal structure, subduction models, etc. Cheng (2016) and subsequent series of results have demonstrated that classical mathematical models such as thermal diffusion are effective for smooth, gradual and linear geological processes, but they are less effective in simulation of singular, abrupt, and nonlinear geological processes and geological events. For the latter, the fractal density and the principle of singularity need to be used. This will be introduced in Sect. 3 of the chapter.

Mathematical Geosciences,

Fig. 3 (a) Euler geometry applied to plate tectonics; (b) explanation of the conjugate relationship among the mid-ocean ridge, subduction zone, and transformation fault (adapted from McKenzie and Parker 1967) according to Euler vector theorem; (c) map showing the conjugate relationship among mid-ocean ridge, subduction zone, and transform faults in the Pacific/North America subduction zone. (Adapted from Johnson and Embley 1990)

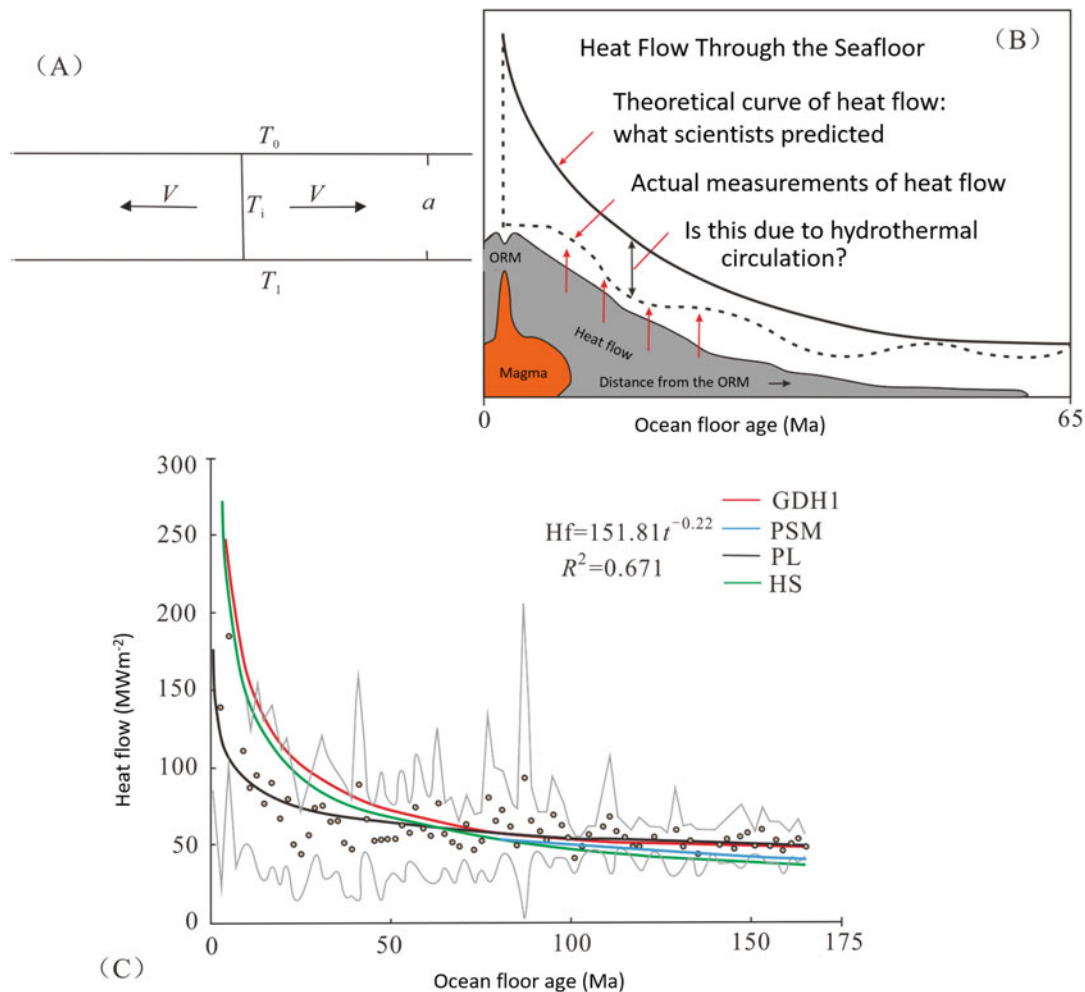
**Mineral Crystals**

Mineralogy is a basic subject of geological sciences. Chemical composition and crystal structure determine the physical and chemical properties of a mineral. Crystallography is a subject of study of the crystal structure of the atomic arrangement of minerals. At the microscopic atomic level, the crystal structure of a mineral can be expressed as a crystal lattice, the smallest unit of a crystal at the molecular level. Therefore, the geometric shape and mathematical properties of the crystal lattice are important parameters for determining the properties of minerals and for minerals classification. Mathematics has played an irreplaceable role in studying the crystal lattice.

Due to the significant contributions of scientists in the field of crystals, crystallography has become a new branch of mathematics. Mathematical crystallography is still a very active field of research today (Nespolo 2008).

Mathematical System of Lattice and Crystals

There are many international distinguished scientists and mathematicians who work in the field of crystallography. Here just name a few examples: German mineralogist, physicist, and mathematician Christian Samuel Weiss (1780–1856) was a pioneer who made fundamental contribution to the modern crystallography. He established crystal



Mathematical Geosciences, Fig. 4 The mid-ocean ridge heat flow model by McKenzie. (Adapted from McKenzie 1967). (a) Model boundary conditions; (b) comparison of McKenzie model theoretical curve and

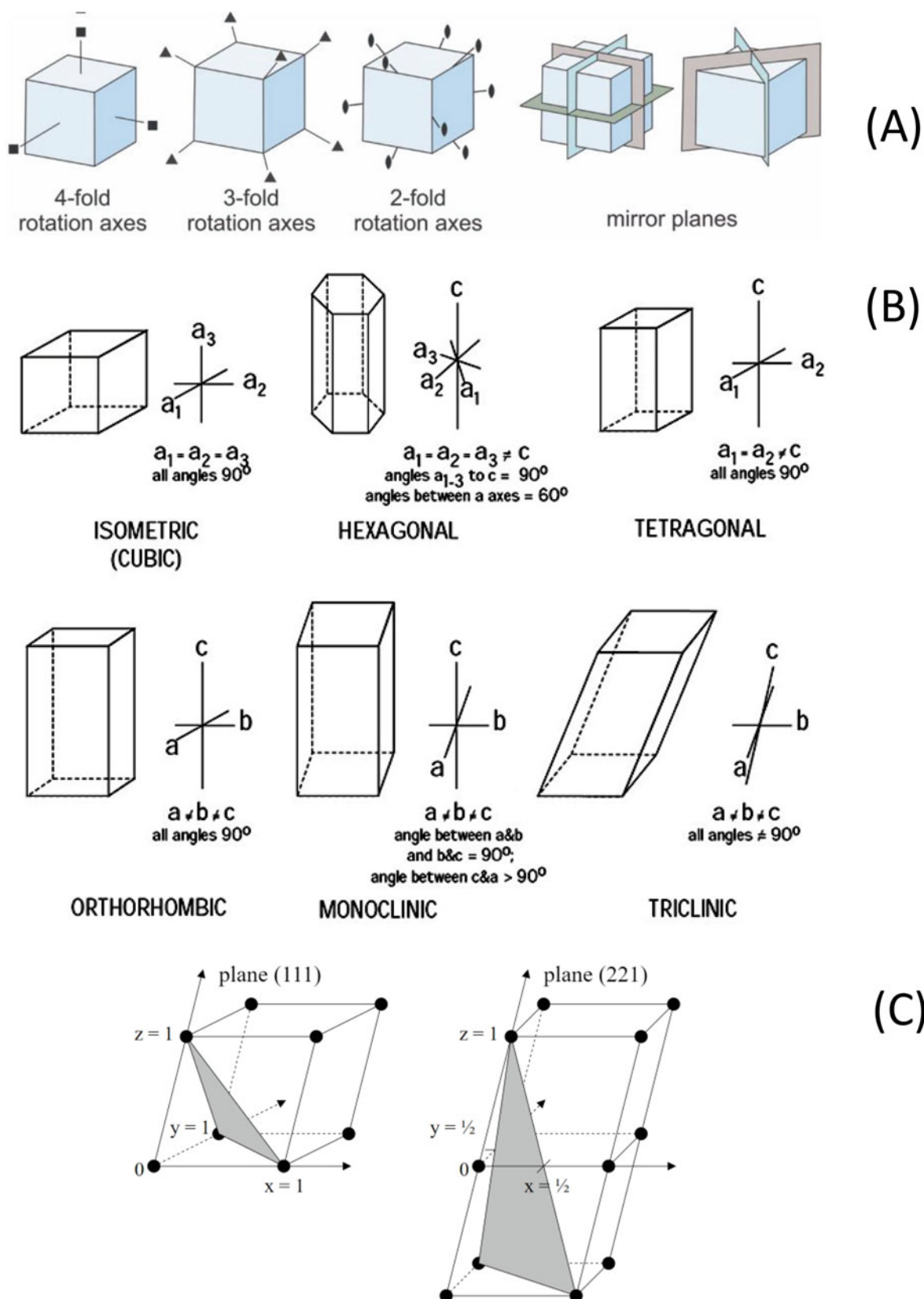
observed data, showing large deviations between the two; and (c) heat-flow profile showing much improved agreement with a new model (by author) using fractal density (black line)

systems and a classification scheme based on crystallographic axes (Fig. 5). He established the Weiss crystal face index or later called the Miller crystal face index (Weiss indices or Miller indices) and their mathematical algebraic relationships (Weiss 1820), which has been termed Weiss's zone law that has been used till today (Fig. 5). Another giant scientist whom I cannot fail to mention is the French physicist and statistician Auguste Bravais (1811–1863). He is not only a great physicist but also an outstanding statistician. He has taught a course in applied mathematics to students of the Department of Astronomy in the Faculty of Natural Sciences since 1840. Bravais published a paper on statistical correlation in 1844 which later became known as the famous Karl Pearson correlation coefficient. Today the Pearson correlation coefficient is still the most used statistical method to measure correlation in natural sciences, social sciences, and economics. Bravais also studied the observation error theory and published the research results of the mathematical analysis of error probability in 1946. Of

course, the most known academic contribution of Bravais is the work on crystallography, especially the lattice system established in 1848 (Fig. 5). He mathematically proved that there were 14 unique lattices in three-dimensional crystalline systems, which were later called Bravais lattices.

Crystal Lattice Symmetry and the Mathematical System of Mineral Classification

The physical and chemical properties of minerals are not only dependent on their chemical compositions but also constrained by their crystal structures. Crystal symmetry refers to the property of the repetition of the same part of a unit cell. It must be achieved through a certain symmetry operation. Crystal symmetry is a primary geometric characteristic of the crystal lattice that has been found as an essential attribute that affects the physical and chemical properties of minerals, such as the optical properties of minerals. There are three types of symmetry operations: rotation, reflection, and



Mathematical Geosciences, Fig. 5 Crystallographic systems. (a) Symmetry of crystal unit cells; (b) crystal systems; (c) Weiss or Miller indices

inversion. Each kind of symmetry can be expressed by a symmetry element, and various combinations of symmetry elements form multiple but not independent combination types. In 1830, the German mineralogist Johan Friedrich Hesse used mathematical deduction to study the symmetry operation and mathematically proved that there are 32 independent crystal symmetries. Thus, mineral crystals should be divided into 32 groups (Bradley and Cracknell 2009). However, the 32 groups of minerals were not completely observed at that time, and it was gradually conformed later. This indeed indicates that answers for complex issues could be predicted when proper mathematical models and laws are established.

Sedimentology-Mathematical Index of Sediment Grain Size Classification

Sedimentology is a basic geological discipline for studying sediments including exploring the origin, transport, deposition, and diagenetic alterations of the materials that compose sediments and sedimentary rocks. The grain size and its distribution of sediments can reflect the lithofacies and paleogeographic environment as well as other mechanisms, such as transportation, sorting, suspension, and deposition. The grain size may also affect other properties of sediments such as porosity, permeability, load intensity, and chemical reactivity. Therefore, grain size analysis is a primary task in sedimentology. According to grain size of sediment, sedimentary rocks are classified as: mudstone (<0.0039 mm), siltstone ($0.0039 \sim 0.0625$ mm), sandstone ($0.0625 \sim 2$ mm), and conglomerate ($2 \sim 256$ mm). Such a classification scheme was originally established based on the Udden classification (Udden 1898) and Wentworth classification (Wentworth 1922). Later, William Christian Krumbein proposed in 1934 the well-known phi (ϕ) scale based on logarithmic transformation of particle size (Krumbein 1934). The particle scale phi (ϕ) provides a quantitative index for the Udden-Wentworth scale grain size classification and laid an important foundation for quantitative classification of sedimentary rocks. Krumbein also published a series of papers in mathematical geology including various indexes for shape measurement and characterization of sediment particles. These works have played a fundamental role in the establishment of the discipline of mathematical geology. Therefore, William Christian Krumbein owned his name as the father of mathematical geology. The International Association for Mathematical Geosciences has established the highest award in the field of mathematical geosciences under the name of William Christian Krumbein.

Foundation of Geographic Information System: Topological Data Model

Topology was established in the mid-seventeenth century by the Swedish mathematician and physicist Leonhard Euler. Topology is a branch of mathematics that studies geometrical

objects that are preserved under continuous deformation. Simple topological relations include orientational relations (such as inside and outside, up and down, left and right), connectivity and directionality, etc. Topological theories and methods are widely used in many fields of natural science (Richeson 2008). For example, the concept of topology has been used in geographic information system (GIS) for organizing spatial data in topological data model and for spatial topological relationship analysis. The capability of handling topological relations for spatial information analysis distinguishes GIS from other computer graphics and image processing systems.

Fundamental Laws of Geochemical Element Distribution

Geochemistry is an interdisciplinary subject of earth science that has developed rapidly in recent decades. The concept of geochemistry was first developed by the German chemist Christian Friedrich Schönbein in 1838, but in the following decades, people still used to call it chemical geology. When it comes to modern geochemistry, we have to mention two founders who are very familiar to all. One is the American chemist Frank Wigglesworth Clarke (1847–1931), and the other is Norwegian mineralogist Victor Goldschmidt (1888–1947). Clark discovered the relationship between the abundance and content of elements in the earth's crust through statistical analysis of a large amount of data, and published a Book entitled "The data of geochemistry" in which the term "Clark Values" was soon widely accepted (Clarke 1920). Clark himself is thus praised as the father of modern geochemistry. The Geochemical Society established its Young Geochemist Award under Clark's name (Mason 1992). Clark also used statistical theory and the least square method in his other research.

Victor Goldschmidt studied inorganic chemistry, physical chemistry, geology, mineralogy, physics, mathematics, and botany during his time at the University of Norway. He proposed rules and standards for the classification of elements. His findings published in the series of books "Geochemical Laws of the Distribution of Elements" have a long-term impact on modern geochemistry. Victor Goldschmidt is praised as the father of modern geochemistry. The Geochemical Society established the highest award in the field of geochemistry under the name of Goldschmidt (Mason 1992). The study of the distribution law of geochemical elements has opened a new research direction for mathematical geosciences. The research on this topic boomed in the 1950s and 1960s. For example, the normal distribution and/or lognormal distribution were believed to be the fundamental law of determination of rock geochemical elements (Ahrens 1953). This widely accepted statistical distribution model lays the foundation for the applications of statistical theories and methods in geochemical data analysis. For more than half a century, statistical theory, statistical models, and statistical

methods have been commonly utilized in the field of geochemistry. These include but not limited to geochemical data error analysis ($\pm 2\sigma$), geochemical element correlation analysis (R^2), statistical inferential analysis, exploratory analysis, statistical reasoning, statistical discrimination, multivariate statistical analysis, anomaly and background separation, geochemical pattern recognition, and geochemical process or mechanism modeling. However, whether normal distribution or lognormal distribution is proper for all situations of the geochemical elements has also been questioned (Aubrey 1954). Aubrey (1954) found that the distribution of certain elements in coal samples deviated from normal or lognormal distribution. What should be the basic law of the distribution of geochemical elements has also become a basic scientific question in the field that has been unresolved for a long time. The author of the chapter and his PhD supervisor Frits Agterberg analyzed the geochemical data from normal rock samples and from mineralized rock samples and they found that the element concentration in normal rock samples can be described by normal or lognormal distribution, while data from samples collected in the mineralized alteration zones deviate from the lognormal distribution (Cheng et al. 1994). Further it was revealed that the relationship between intensity of ore elements anomaly in the alteration zones and the areas of the anomaly with various thresholds depicts self-similarity which can be described by the following “Concentration-Area” (C-A model) fractal model (Cheng 1994).

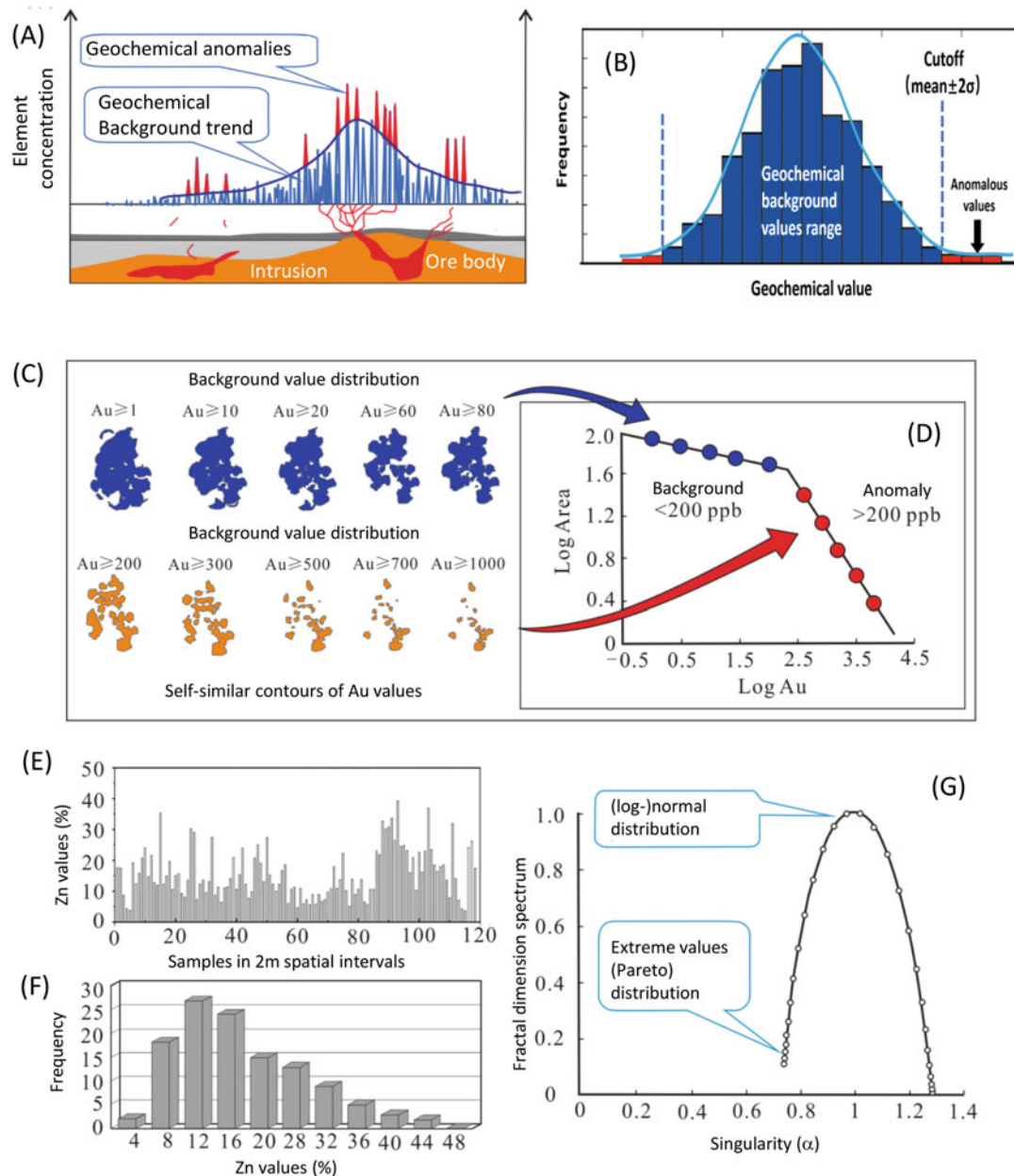
$$C(A) = cA^{-\Delta x/2} \quad (3)$$

where A represents the abnormal area, and $C(A)$ represents the average element concentration within the area A . A variety of mineralization-related elements (such as Au, Cu, As, Pb/Zn) depict the above power law relationship between the area and the average concentration values. The model provides a mathematical formula quantitatively describing the abnormal enrichment law of ore-forming elements in the mineralization district for the first time. The idea and power law model were utilized to describe geochemical distribution and to effectively distinguish between “geochemical anomaly” and “geochemical background.” This original idea and C-A fractal model has opened important new direction of quantitative research on identification of geochemical anomalies and separation of anomaly from background (Cheng 1994). The theoretical background of the C-A model is the extreme value statistics and singularity of multifractal theory. Multifractals can be considered as a generalized model which describes more complex distributions than normal or lognormal distributions of values with multiple scaling properties. Multifractals can simultaneously describe a normal distribution or a lognormal distribution of values around the mean and power-law distribution or fractal long-tailed distribution

of values along two tails. Therefore, multifractal distribution model unifies the normal distribution, lognormal distribution and Pareto’s distribution or Zipf’s law distribution (Fig. 6). Multifractals thus can be used as a general and prime distribution model for geochemical values. So far, the C-A method and the expended forms have wide applications in geochemical data analysis for mineral resources assessment and mineral deposits prediction. This method has been demonstrated as one of the most effective methods in mineral resource prospecting and environmental assessments. The paper (Cheng et al. 1994) is one of the most cited papers in *Journal of Geochemical Exploration*.

Quantitative Prediction and Assessment of Mineral and Energy Resources

Quantitative description, simulation, and prediction in geosciences are the fundamental objectives of mathematical geosciences, which require reasonable mathematical models as foundation. Mineral, water, and energy resources are the basic elements for social and economic development. The increasing demands for natural resources for the sustainable development and improving human life quality require solutions for supplies of mineral, water, and energy resources and for green utilization of natural resources in protecting environment and ecological systems. The main tasks of quantitative prediction and assessments of natural resources (minerals, water, and energy) may include but are not limited to determining potential locations of resources, estimating amounts of resources, and evaluating economic values and development feasibility. These results can provide supports for the government or industry in informative decision-making in planning and development of mineral, water, and energy resources. Due to the complexity of formation mechanisms and the distribution of natural resources and the limitations of human intelligence, there are still strong uncertainties in the quantitative prediction of natural resources (Singer and Menzie 2010). Since the establishment of IAMG, the quantitative assessments of mineral and energy resources have always been an important research field of the society. A series of predicting theories and methodologies have been successively developed and widely used. For example, researchers at the Geological Survey of Canada (GSC) have developed theories and methods including weight of evidence model, fuzzy logic, neural network, and logistic regression for comprehensive data integration for mineral potential prediction (Agterberg 1989); the researchers at the US Geological Survey (USGS) have developed and applied a “three-step” mineral resource assessment method [34]; and Chinese scholars have developed predictive methods on the basis of the principle of genetic association of geological anomaly and metallogeny (Zhao 1998, 2002), synthesis of comprehensive information for mineral resource



Mathematical Geosciences, Fig. 6 Law of geochemical distribution of the geochemical elements. (a) Separation of anomalies ($>2\sigma$) based on normal or log-normal distributions; (b) pattern delineated according

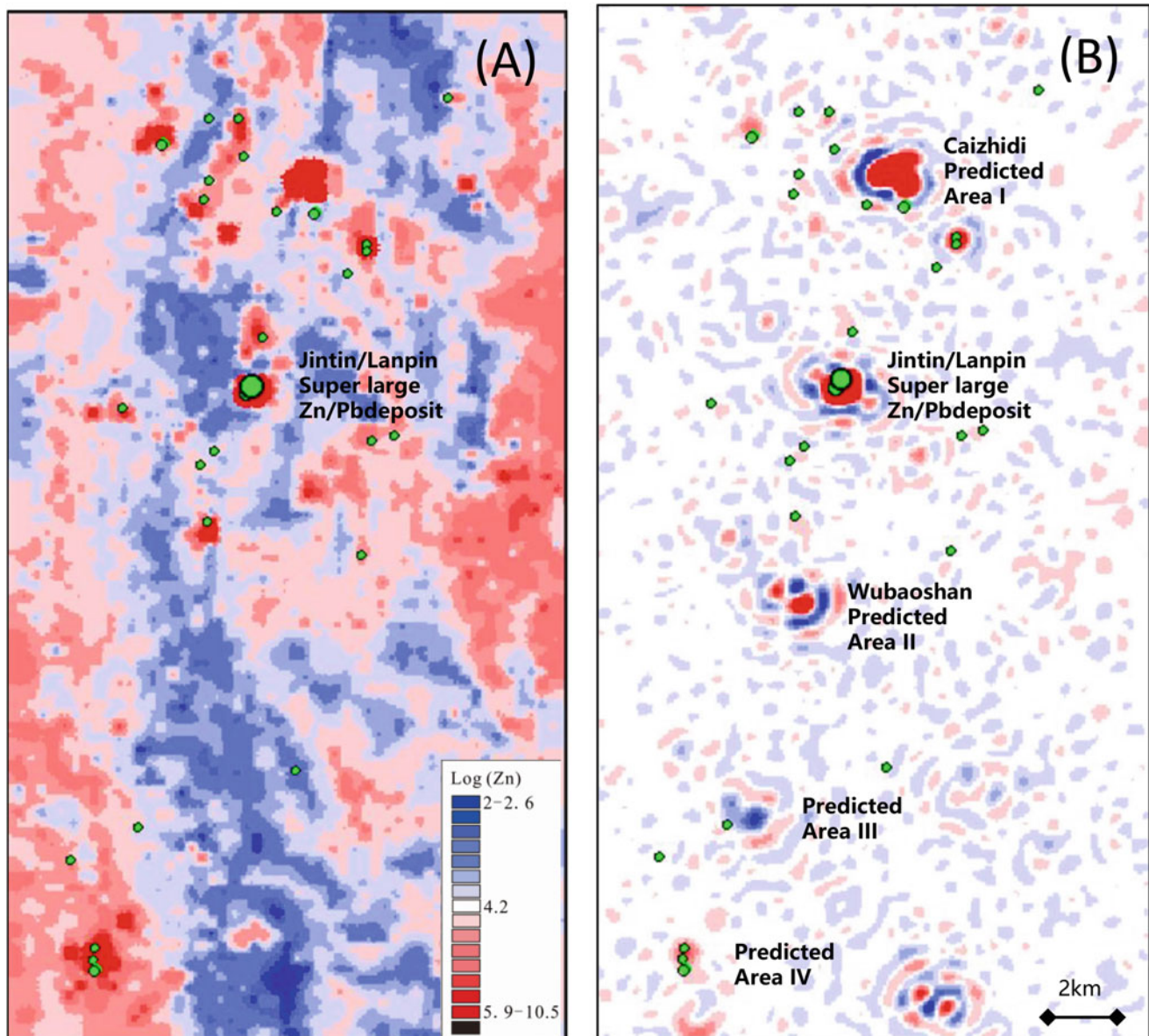
to varying elemental concentrations in a mineral district depicts self-similarity (adapted from Cheng et al. 1994); and (c) the multifractal model (adapted from Cheng 2007a, 2015a)

prediction (Wang et al. 1990), and nonlinear methods for mineral resources prediction on the basis of fractal singularity theory (Cheng 2007b, 2008). In recent years, exploration for mineral, water, and energy resources has become the forefront of geoscience due to the shift from traditional areas for shallow and low-cost resources to nontraditional regions such as deep earth, remote areas, covered areas, deep ocean, and deep space for strategic resources including critical minerals. Here are just a few examples to show the recent activities of resources assessments in nontraditional areas, and USGS

conducted a statistical assessment of oil and gas resources in the Arctic in 2009. The results of the study show that about 30% of the world's undiscovered gas and 13% of the world's undiscovered oil may exist in Arctic (Gautier et al. 2009). In 2018, Japanese scholars reported discoveries of considerable amounts of rare-earth elements and yttrium (REY) resources in deep sea mud in the western North Pacific Ocean (Takaya et al. 2018). Italian scientists claimed they found evidence of liquid water in a subglacial lake with a diameter of 20 km below the ice of the South Polar Layered Deposits on Mars

(Orosei et al. 2018). For the past decade, the author of this chapter and his research team have been focusing on quantitative prediction and evaluation of mineral resources in the regions covered by transported materials in several provinces of China. Figure 7 shows a successful example of mapping mineral potential for Pb/Zn/Cu and other metals in the Lanping-Jinding area of Yunnan Province. Stream sediment geochemical data (Fig. 7a) were processed using the Spectrum-Area (S-A) fractal filtering method (Cheng 2012a). Use of the S-A method could effectively decompose the weak anomalous from variable background signals (Cheng and Xu 1998; Cheng et al. 2000). The S-A model successfully delineated mineralization-induced anomalies by eliminating

the influence of complex background, especially the interference from the basalt containing high concentration of metals. Five copper-lead-zinc prospecting target areas with equal spatial intervals conforming self-similar distributions have been delineated (Fig. 7b). Based on the above findings, the Geological Survey Institute affiliated to Yunnan Geological Survey carried further exploration and successfully discovered several mineral deposits in the target areas. The same method was further utilized by the Yunnan Geological Survey for mineral prospecting in the entire Sanjiang area and more mineral deposits were discovered in the delineated target areas.



Mathematical Geosciences, Fig. 7 Application of spectral energy density – area method (S-A) to Zn anomaly and related background analyses in the Lanping-Jinding area in Yunnan Province. (a) Distribution of Zn anomalies; and (b) newly identified Cu-Pb-Zn prospecting areas

Frontiers of Mathematical Geosciences

Some contributions of mathematical geoscientists to related fields of geosciences have been introduced in the previous section. Next, we will briefly discuss the frontiers of mathematical geosciences, aiming to answer the question are mathematical geoscientists at the earth science frontiers? To answer the question, one approach is to summarize current scientific research in mathematical geosciences, and the other is to study the frontiers of geosciences and the strategic programs of some large international scientific organizations, such as the Future Earth 2025 vision, 2021 Strategic Challenge Domains of ISC, the Resourcing Future Generations (RFG) (Lambert et al. 2013) and Deep-time Digital Earth (DDE) (Cheng et al. 2020) programs led by International Union of Geological Sciences (IUGS), the newly released “Earth in Time, A Vision for NSF Earth Sciences” 2020–2030 by National Academies of Sciences, Engineering and Medicine (2020), US Geological Survey -Century Science Strategy for 2020–2030 (USGS 2021), and the new 2050 Science Framework: Exploring Earth by Scientific Ocean Drilling proposed by the International Ocean Discovery Program (IODP) (Koppers and Coggon 2020). These strategic reports and big science programs have revealed the developing trend of geosciences in next decade from different perspectives. Furthermore, the author has maintained close communication with many international associations through international programs and meetings, including ISC-CSP, IUGG, AGU, EGU, SGA, PDAC, and CGU. Some key topics that can be extracted from above sources of information can reflect the current trends and frontiers of the earth sciences, such as data science, data analysis, computing, interdisciplinary, uncertainty of measurement and prediction, geometry and dynamics of the Earth, climate change and the Arctic, Antarctic, and Tibet Plateau. It also reflects the three major challenges faced by geoscientists in understanding the interaction between the earth system and the human system, understanding the past, present, and future changes of the livable earth’s environment and predicting extreme events on the Earth: (1) The complexity of multilayer interaction of earth system; (2) the chaotic nature of earth processes and the singularity and predictability of extreme earth events; and (3) the effectiveness of observation and monitoring of multi-scale nonlinear mixing processes. These challenges are closely related to mathematical geosciences. Innovations on new mathematical theories, models, and computer software are the prerequisite for dealing with these challenges. Therefore, mathematical geoscientists, confronting great challenges and opportunities, are at the earth science frontiers. Many important progresses have been made by mathematical geoscientists in the earth science frontiers and two examples will be detailed as follows: quantitative simulation of

geocomplexity and applications of big data and artificial intelligence (AI).

Modelling Geocomplexity of the Earth: A New Growing Field of Mathematical Geosciences

Understanding and simulating the complexity and chaotic behaviors of the earth systems is a long-term task for scientists. However, there is no unique scientific definition of complexity of the earth system. Some authors regard modeling geocomplexity as a new kind of science (Wolfarm 2002; Turcotte 2006). These authors emphasize that the geocomplexity refers to nonlinear geological properties that cannot be described or characterized by classical mathematical calculus equations. Simple geological processes can be modeled by means of classical mathematical equations with proper boundary conditions and initial conditions. The temporal and spatial distribution of geological processes can be deterministically predicted, for example, the heat flow at the plate margin over the mid-ocean ridges can be estimated by setting up a diffusion model assuming the plate near the mid-ocean ridge as a regular rigid body, its upper and lower surface temperatures are used as boundary conditions, and the mid-ocean ridge is set as the initial condition. Under several simplifications the analytical or numerical solutions of the established thermal diffusion equation can be derived to estimate the heat flux at any position of the plate. However, many complex geological processes cannot be precisely modeled by the above classical differential equations. For example, it is impossible to precisely describe turbulence by deterministic mathematical model due to statistical uncertainty. In order to study such processes and phenomena, American mathematicians and meteorologists Lorenz (1963) proposed Chaos theory and French American mathematician Mandelbrot (1967) proposed fractal theory in 1960s. For a chaotic system or chaotic process, the output of prediction model behaves very differently if a small disturbance in the initial conditions or boundary conditions. Chaotic systems often appear random and unpredictable, but in fact have regular patterns with scale-invariant properties or fractality (Valentine et al. 2002). Fractal geometry is a new branch of mathematics that describes geometry from the perspective of fractional dimensions, including new geometries between traditional points, lines, surfaces, and volumes (Lovejoy et al. 2009). Fractal geometry is irregular geometry or patterns such as Mandelbrot set and Menger sponge with scale invariant property or self-similarity. Obviously, the concept of fractal geometry has greatly generalized the scope of traditional geometry which provides new effective mathematical tools for studying geocomplexity (Mandelbrot 1967). In recent years, fractal geometry theory and fractal models have been used in almost all disciplines of earth sciences. Just to name a few examples, frequency–size power law models were utilized to

characterize earthquake magnitude-number frequency relationship (Gutenberg and Richter 1944), area-number of plates (Sornette and Pisarenko 2003), plate boundary morphology distribution (Ouillon et al. 1996), magmatic rock area-frequency distribution (Pelletier 1999), volcanic eruption interval distribution (Cannavò and Nunnari 2016), deposit size-number distribution (Agterberg 1995), mineralized alteration zone area-perimeter distribution (Cheng 1995), and geochemical anomalies concentration and area distribution (Cheng 2007b).

With the development of fractal geometry and its application in various fields, the concept of fractal and fractal theory have also been further developed. For example, the proposal of multifractal theory (Mandelbrot 1977) has extended fractals from fractal geometry to self-similar distribution of measures or patterns defined on geometric supports (it can also be fractal geometry) which marks an important leap-forward development of fractal theory (Mandelbrot 1989). Multifractal model can be considered as a new distribution model unifying normal, lognormal, and Pareto distributions. Theory of multifractals provides not only a new way to describe complex spatial patterns with self-similarity but also a basis for further development of fractal calculus and fractal dynamics system (Lovejoy 1991). Since geometric set theory is the foundation of mathematical calculus including integral and differential operations. Only when the mathematical system of geometric measures and calculus operations is established can geometry play a greater role in mathematics. How to build mathematical calculus operations and analysis of other physical quantities related to fractals such as specific gravity, density, and thermal conductivity. Establish calculus based on fractal geometry is a unsolved question which has not been systematically explored in the literature. At the present time, the research progress in this area is still very slow, and more efforts especially by young scholars are encouraged.

The author has been working on the above problems for a long time and first discovered that the concentration values of ore elements in mineralized rocks may follow the multifractal distribution with self-similarity between average concentration levels (density) and areas of mineralized rocks. Further a fractal model (C-A model) was derived to associate the average concentration values and the irregular areas (fractals) delineated with the concentration values around concentration anomalies. This model shows that the concentration (density) of elements on a fractal depicts power law relations between density and area with a negative fractional exponent, implying the concentration within mineralized areas depicts singularity – infinity large of concentration vs infinity small area. The author termed this type of density as “fractal density,” the density defined on fractals or a density with fractal dimension (Fig. 8a). Accordingly proper mathematical model and the local singularity calculation method have been developed to describe fractal density caused by nonlinear

geological processes (earthquakes, mineralization, magmatic events, volcanic events, etc.).

Various extreme events (earthquakes, volcanoes, magmatism, mineralization, tsunamis, landslides, etc.) caused by sudden changes of geological systems were studied to explore their common physical mechanism of formation (Fig. 8b, c) and the singularity of the system outputs. The main types of mechanism that explain most of these extreme events may include but not limited to phase transition mechanisms (supercritical fluid, Moho discontinuity, subduction slab interface, interaction surface between deep magma and surface water, etc.), self-organized criticality (lithospheric delamination, rock fracture, slab fragmentation, etc.), and multiplicative cascading cyclic processes, dissipative structures, and chaotic processes (turbidity current, magma convection, mantle flow, metamorphism, etc.) (Fig. 8b, c) (Cheng 2007b, c, 2008, 2018a).

The author refers to such extreme geological events as singular geological events or events with singularity. The metric fractal density concept and the local singularity analysis (LSA) method developed have been successfully applied to the quantitative simulation and prediction of a variety of extreme geological events (Cheng 2007b). These include the identification of geochemical anomalies caused by concealed ore bodies in the coverage area and the delineation of prospecting target areas (Cheng 2007b, 2015a), seismic probability density-magnitude model (Cheng and Sun 2018), magma flare-up events due to collision of Indian continent and Eurasia continent (Cheng 2018b), anomalous heat flow over the mid-ocean ridges (Cheng 2016), large-scale magma episodic activity and prediction of plates tectonic future evolution (Cheng 2018b, c); and the singularity of lithospheric phase transformation (Cheng 2018a).

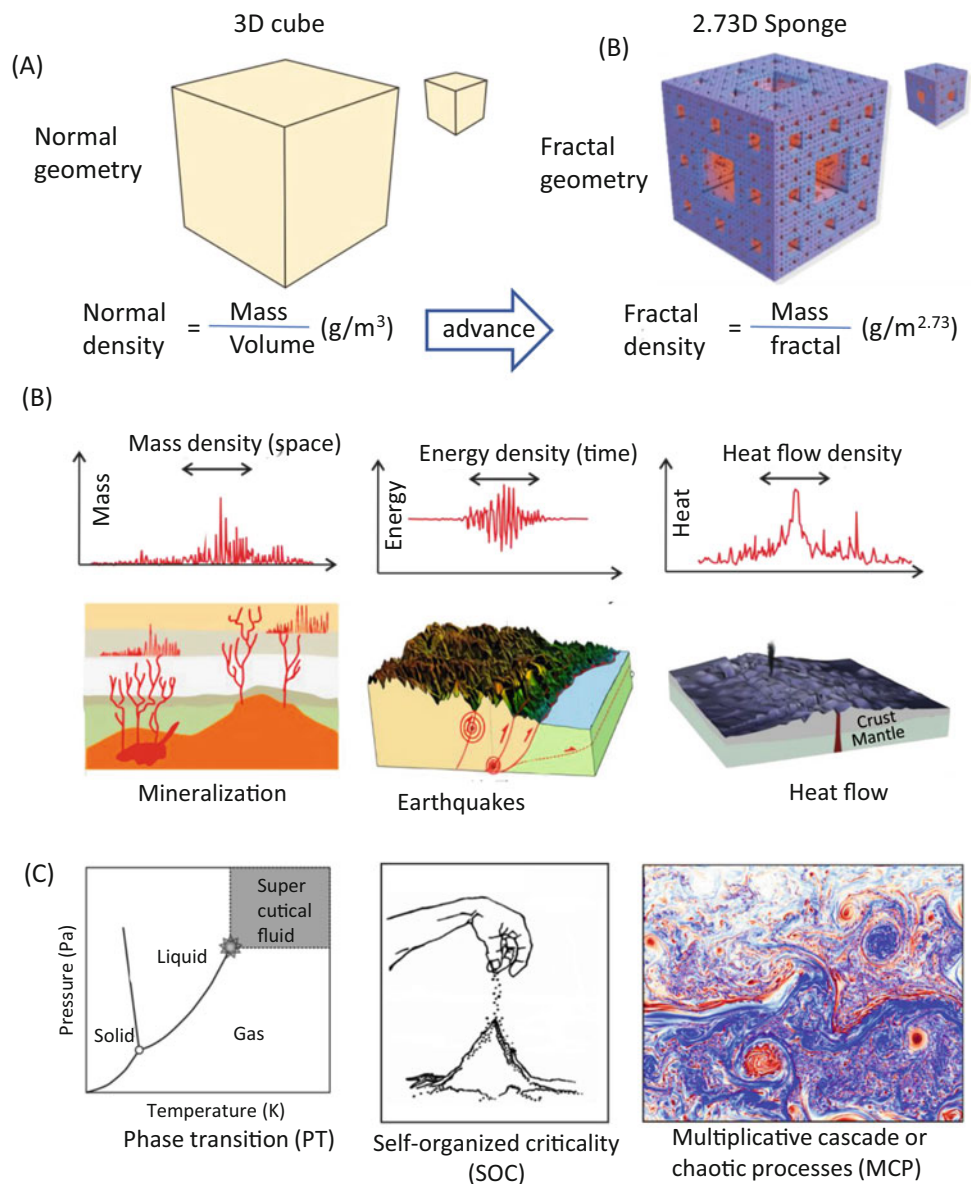
Big Data Analytics and Complex Artificial Intelligence

Big data and artificial intelligence (AI) have significantly influenced our life and gradually became research hotspots in many fields of natural science, social science, and economics. However, scientists from different fields may have different opinions about the impact of big data and AI in their research fields. For mathematical geoscientists, they are both the contributors to the development of the data science and the users taking full advantage of big data and AI technology and using them for solving practical problems. One of the good questions for mathematical geoscientists is how to further develop big data analytics and AI technology to enhance the applications.

AI and ML are based on mathematical models, algorithms, and computer performance. Some common ML techniques include linear regression, logistic regression, artificial neural network, naïve Bayes, weights of evidence, random forests, adaBoost, and principal component analysis. Besides, some traditional mathematical methods such as Monte Carlo

Mathematical Geosciences,

Fig. 8 Concepts of fractal density and its mechanisms and possible geological processes that may result in fractal density. **(a)** Normal density and fractal density; **(b)** examples of geological processes that may result in fractal density; and **(c)** mechanisms of fractal density generation. (The figure describing self-organized criticality was adapted from Wiesenfeld et al. 1989)



simulation and least-square fitting are often used in ML. For example, the research algorithm for AlphaGo combined Monte Carlo simulation with value and policy networks (Silver et al. 2016). The abovementioned ML methods are common in mathematical geosciences research. Over the past few decades, these methods have demonstrated great potential for problem solving in mathematical geosciences, particularly for mineral potential mapping and natural resource assessment (Agterberg 1989; Bonham-Carter 1994). In addition, mathematical geoscientists have developed many new models, methods as well as software which have enriched and consummated the function of AI. For example, various variations of the weight of evidence models based on the Bayesian principle have been established ranging from the original

model that requires complete conditional independence of predictors (Naive Bayes, Naïve Bayes, or normal weight of evidence, Weights of Evidence) (Agterberg 1989), to a new model based on predictors with a weak conditional independence (AdaBoost-weight of evidence, AdaBoost Weights of Evidence) (Cheng 2008, 2012b, 2015b) and then to a modified model that does not require conditional independence (weighted weights of evidence) (Journel 2002; Cheng 2008; Zhang et al. 2009; Agterberg 2011).

With the rapid development of deep learning, the research in the field of machine learning using artificial neural network (ANN) models has increased dramatically in recent years. However, even though the current ANN can theoretically approximate all Lebesgue integral function per universal

approximation theorem, it is not capable of approximating complex non-integrable functions at present due to singularity, non-integrability, and non-differentiability. It often needs to set a large number of neurons and more layers (depth) of neurons to approximate complex functions. This of course may increase the number of model parameters, which in turn requires more training data and computational power. Otherwise, the results would be of low stability and even divergent, leading to lower repeatability and precision. How to develop ANN models for solving fractal density related problem caused by singular extreme geological events remains as a question which still requires further investigation. This kind of AI should probably be defined as complex artificial intelligence or local precision artificial intelligence.

Conclusions

Mathematical geosciences is an interdisciplinary science discipline of studying earth characteristics and geological processes, as well as quantitative prediction and evaluation of resources and environment including disasters. Together with other geoscience disciplines MG plays a dispensable role in studying the evolution of the livable earth system, predicting and evaluating various extreme geological events and serving the sustainable development of society. Mathematical geoscientists have made and will continue to make important contributions to the establishment of the modern earth science system and the development of innovative theories and methods to meet the major needs of mankind. Mathematical geoscientists are already at the forefront of earth sciences. It is necessary to publicize the scientific content of mathematical geosciences comprehensively and correctly. It needs to motivate the interest of school and university students in integrating mathematics and geosciences and to encourage and cultivate more young talents to devote them to the MG. Earth science frontiers in the twenty-first century with supports of big data and artificial intelligence require that earth science moves from simple to complex, from local to global, from description to mechanism, from qualitative to quantitative, and from modeling to prediction. Mathematical Geosciences is an emerging discipline full of hope and prospects.

Bibliography

- Agterberg F (1974) *Geomathematics, mathematical background and geo-science application*. Elsevier, Amsterdam
- Agterberg F (1989) Computer programs for mineral exploration. *Science* 245(4913):76–81
- Agterberg F (1995) Multifractal modeling of the sizes and grades of giant and supergiant deposits. *Int Geol Rev* 37(1):1–8
- Agterberg F (2011) A modified weights-of-evidence method for regional mineral resource estimation. *Nat Resour Res* 20(2):95–101
- Agterberg F (2014) *Geomathematics: theoretical foundations, applications and future developments*. Springer, Cham
- Ahrens L (1953) A fundamental law of geochemistry. *Nature* 172(4390):1148–1148
- Aubrey K (1954) Frequency distribution of the concentrations of elements in rocks. *Nature* 174(4420):141–142
- Bonham-Carter G (1994) Geographic information systems for geoscientists-modeling with GIS. *Comput Methods Geosci* 13:398
- Bradley C, Cracknell A (2009) *The mathematical theory of symmetry in solids: representation theory for point groups and space groups*. Oxford University Press, Oxford
- Cannavò F, Nunnari G (2016) On a possible unified scaling law for volcanic eruption durations. *Sci Rep-UK* 6(1):22289
- Cheng Q (1995) The perimeter-area fractal model and its application to geology. *Math Geol* 27(1):69–82
- Cheng Q (2007a) Multifractal imaging filtering and decomposition methods in space, Fourier frequency, and Eigen domains. *Nonlinear Proc Geophys* 14(3):293–303
- Cheng Q (2007b) Mapping singularities with stream sediment geochemical data for prediction of undiscovered mineral deposits in Gejiu, Yunnan Province, China. *Ore Geol Rev* 32(1/2):314–324
- Cheng Q (2007c) Singular mineralization processes and mineral resources quantitative prediction: new theories and methods. *Earth Sci Front* 14(5):42–53. (In Chinese with English Abstract)
- Cheng Q (2008) Non-linear theory and power-law models for information integration and mineral resources quantitative assessments. *Math Geosci* 40(5):503–532
- Cheng Q (2012a) Singularity theory and methods for mapping geochemical anomalies caused by buried sources and for predicting undiscovered mineral deposits in covered areas. *J Geochem Explor* 122:55–70
- Cheng Q (2012b) Application of a newly developed boost Weights of Evidence model (BoostWoE) for mineral resources quantitative assessments. *J Jilin Univ* 42(6):1976–1985
- Cheng Q (2014) Generalized binomial multiplicative cascade processes and asymmetrical multifractal distributions. *Nonlinear Proc Geoph* 21(2):477–487
- Cheng Q (2015a) Multifractal interpolation method for spatial data with singularities. *J S Afr Min Metall* 115(3):235–240
- Cheng Q (2015b) BoostWoE: a new sequential weights of evidence model reducing the effect of conditional dependency. *Math Geosci* 47(5):591–621
- Cheng Q (2016) Fractal density and singularity analysis of heat flow over ocean ridges. *Sci Rep-UK* 6(1):19167
- Cheng Q (2018a) Mathematical geosciences: local singularity analysis of nonlinear earth processes and extreme geo-events. In: *Handbook of mathematical geosciences*. Springer, Cham
- Cheng Q (2018b) Singularity analysis of magmatic flare-ups caused by India-Asia collisions. *J Geochem Explor* 189:25–31
- Cheng Q (2018c) Extrapolations of secular trends in magmatic intensity and mantle cooling: implications for future evolution of plate tectonics. *Gondwana Res* 63:268–273
- Cheng Q, Sun H (2018) Variation of singularity of earthquake-size distribution with respect to tectonic regime. *Geosci Front* 9(2):453–458
- Cheng Q, Xu Y (1998) Geophysical data processing and interpreting and for mineral potential mapping in GIS environment. In: *Proceedings of the fourth annual conference of the international association for mathematical geology*. De Frede Editore, Napoli, Italy, pp 394–399
- Cheng Q, Agterberg F, Ballantyne S (1994) The separation of geochemical anomalies from background by fractal methods. *J Geochem Explor* 51(2):109–130

- Cheng Q, Xu Y, Grunsky E (2000) Integrated spatial and spectrum method for geochemical anomaly separation. *Nat Resour Res* 9(1): 43–52
- Cheng Q, Oberhänsli R, Zhao M (2020) A new international initiative for facilitating data – driven Earth science transformation. *Geol Soc Lond, Spec Publ* 499(1):225–240
- Clarke FW (1920) *The Data of Geochemistry*, 4th Edition, U. S. Geological Survey, Washington, D. C. pp. 832
- Gautier D, Bird K, Charpentier R, Grantz A, Houseknecht D, Klett T, Moore T, Pitman J, Schenk C, Schuenemeyer J, Sørensen K, Tennyson M, Valin Z, Wandrey C (2009) Assessment of undiscovered oil and gas in the Arctic. *Science* 324(5931): 1175–1179
- Gutenberg B, Richter C (1944) Frequency of earthquakes in California. *B Seismol Soc Am* 34(4):185–188
- Hess H (1960) *The evolution of ocean basins*. Department of Geology, Princeton University, Princeton
- Howarth R (2017) *Dictionary of mathematical geosciences: with historical notes*. Springer, Cham
- Johnson HP, Embley RW (1990) Axial seamount: an active ridge axis volcano on the central Juan de Fuca Ridge. *J Geophys Res-Solid Earth* 95:12689–12696
- Journel A (2002) Combining knowledge from diverse sources: an alternative to traditional data independence hypotheses. *Math Geol* 34(5): 573–596
- Koppers A, Coggon R (2020) Exploring earth by scientific ocean drilling: 2050 science framework. UC San Diego Library Digital Collections, La Jolla
- Krumbein W (1934) Size frequency distributions of sediments. *J Sediment Res* 4(2):65–77
- Lambert I, Durrheim R, Godoy M, Kota K, Leahy P, Ludden J, Nickless E, Oberhänsli R, Wang A, Williams N (2013) Resourcing future generations: a proposed new IUGS initiative. *Episodes* 36(2): 82–86
- Lodge O (1894) Mathematical geology. *Nature* 50:289–293
- Lorenz E (1963) Deterministic nonperiodic flow. *J Atmos Sci* 20(2): 130–141
- Lovejoy W (1991) A survey of algorithmic methods for partially observed Markov decision processes. *Ann Oper Res* 28(1): 47–65
- Lovejoy S, Agterberg F, Carsteanu A, Cheng Q, Davidsen J, Gaonac’h H, Gupta V, L’Heureux I, Liu W, Morris S, Sharma S, Shcherbakov R, Tarquis A, Turcotte D, Uritsky V (2009) Nonlinear geophysics: why we need it. *EOS Trans Am Geophys Union* 90(48): 455–456
- Mandelbrot B (1967) How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science* 156(3775):636–638
- Mandelbrot B (1977) *Fractals: form, chance and dimension*. W.H. Freeman, San Francisco
- Mandelbrot B (1989) *Multifractal measures, especially for the geophysicist. Fractals in geophysics*. Birkhäuser, Basel, pp 5–42
- Mason B (1992) Victor Moritz Goldschmidt: father of modern geochemistry. *Geochemical Society, Washington, DC*
- McCammon R (1975) *Concepts in geostatistics*. Springer, Cham
- Mckenzie D (1967) Some remarks on heat flow and gravity anomalies. *J Geophys Res* 72(24):6261–6273
- Mckenzie D, Parker R (1967) The North Pacific: an example of tectonics on a sphere. *Nature* 216(5122):1276–1280
- Merriam D (1970) *Geostatistics: A Colloquium (Computer Applications in the Earth Sciences)*. Plenum Press, New York, pp. 177
- National Academies of Sciences, Engineering and Medicine (2020) *A vision for NSF Earth Sciences 2020–2030: Earth in time*. The National Academies Press, Washington, DC
- Nespolo M (2008) Does mathematical crystallography still have a role in the XXI century? *Acta Cryst A* 64(1):96–111
- Orosei R, Lauro S, Pettinelli E, Cicchetti A, Coradini M, Cosciotti B, Paolo F, Flamini E, Mattei E, Pajola M, Soldovieri F, Cartacci M, Cassenti F, Frigeri A, Giuppi S, Martufi R, Masdea A, Mitri G, Nenna C, Noschese R, Restano M, Seu R (2018) Radar evidence of subglacial liquid water on Mars. *Science* 361(6401):490–493
- Ouillon G, Castaing C, Sornette D (1996) Hierarchical geometry of faulting. *J Geophys Res-Solid Earth* 101(B3):5477–5487
- Pelletier J (1999) Statistical self-similarity of magmatism and volcanism. *J Geophys Res-Solid Earth* 104(B7):15425–15438
- Richeson D (2008) *Euler’s Gem: the polyhedron formula and the birth of topology*. Princeton University Press, Princeton
- Silver D, Huang A, Maddison C, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T, Hassabis D (2016) Mastering the game of Go with deep neural networks and tree search. *Nature* 529(7587):484–489
- Singer D, Menzie W (2010) *Quantitative mineral resource assessments: an integrated approach*. Oxford University Press, Oxford
- Sornette D, Pisarenko V (2003) Fractal plate tectonics. *Geophys Res Lett* 30(3):1105
- Takaya Y, Yasukawa K, Kawasaki T, Fujinaga K, Ohta J, Usu Y, Nakamura K, Kimura J, Chang Q, Hamada M, Dodbiba G, Nozaki T, Iijima K, Morisawa T, Kuwahara T, Ishida Y, Ichimura T, Kitazume M, Fujita T, Kato Y (2018) The tremendous potential of deep-sea mud as a source of rare-earth elements. *Sci Rep-UK* 8:5763. <https://doi.org/10.1038/s41598-018-23,948-5>
- Turcotte D (2006) Modeling geocomplexity: “a new kind of science”. *Special papers-Geological Society of America* 413. p 39
- Udden J (1898) *The mechanical composition of wind deposits*. Lutheran Augustana Book Concern, Lindsborg
- United States Geological Survey (2021) *Geological survey 21st-century science strategy 2020–2030*. Geological Survey, Reston
- Valentine G, Zhang D, Robinson B (2002) Modeling complex, nonlinear geological processes. *Annu Rev Earth Planet Sci* 30(1):35–64
- Vistelius A (1962) Problems of mathematical geology: a contribution to the history of the problem. *Geol Geophys* 12(7):3–9
- Wang S, Cheng Q, Fan J (1990) *Integrated evaluation methods of information on gold resources*. Jilin Science and Technology Press, Changchun. (In Chinese)
- Weiss C (1820) *Über die theorie des epidotsystems*. *Abhandlungen der Königlichen Akademie der Wissenschaften zu Berlin* pp 242–269
- Wentworth C (1922) A scale of grade and class terms for clastic sediments. *J Geol* 30(5):377–392
- Wiesenfeld K, Chao T, Bak P (1989) A physicist’s sandbox. *J Stat Phys* 54(5):1441–1458
- Wolfram S (2002) *A new kind of science*. Champaign. Wolfram Media, Champaign
- Zhang S, Cheng Q, Zhang S, Xia Q (2009) Weighted weights of evidence and stepwise weights of evidence and their application in Sn–Cu mineral potential mapping in Gejiu, Yunnan Province, China. *Earth Sci-J China Univ Geosci* 34:281–286. (in Chinese with English abstract)
- Zhao P (1998) *Geological anomaly and mineral prediction: modern theory and methods for mineral resources evaluation*. The Geological Publishing House, Beijing
- Zhao P (2002) “Three-component” quantitative resource prediction and assessments: theory and practice of digital mineral prospecting. *Earth Sci-J China Univ Geosci* 27(5):482–489. (In Chinese with English Abstract)

Mathematical Minerals

John H. Doveton

Kansas Geological Survey, University of Kansas, Lawrence, KS, USA

Definition

Mathematical minerals are calculated from chemical or physical measurements as theoretical realizations of actual mineral assemblages in rocks. The concept was first introduced as “normative” minerals in igneous rocks, where oxide analyses were assigned to minerals in a protocol that follows crystallization history. While there is only a general match with observed “modal” minerals, the intent of the procedure is to create hypothetical mineral assemblages which can then be compared between igneous rocks. The concept of normative minerals was extended to sedimentary rocks. However, in this case, the mathematical procedures used attempt to match as closely as possible the volumes of mineral species observed. In the most widely used modern application, mathematical mineral compositions are estimated in subsurface rocks from downhole measurements of physical properties or elements.

Normative Minerals in Igneous Rocks

The Cross-Iddings-Pirsson-Washington (CIPW) norm has been a standard method for many years to transform oxides measured in igneous rocks to theoretical “normative minerals” (Cross et al. 1902). These norms represent hypothetically the minerals that would crystallize if the igneous source was cooled at low pressure under perfect dry equilibrium conditions. The procedure is governed by simplifying assumptions regarding both the order of mineral formation and phase relationships in the melt. These constraints result in a simple model that deviates from the course of naturally occurring igneous mineral differentiation. However, norms are useful for comparison and classification of igneous rocks, particularly for glassy and finely crystalline samples. Normative minerals differ from “modal” minerals observed in the rock, whose volumes are estimated typically from point-counting procedures applied to thin sections.

Normative Minerals in Sedimentary Rocks

As applied to sedimentary rocks, the calculation of normative minerals does not generally follow a genetic model, but simply aims to estimate mineral assemblages that are a close match with those observed. The earliest methods drew on

molecular ratios to calculate the mineral components of a rock, using oxide percentages from analysis (Krumbein and Pettijohn 1938). The suite of minerals to be solved was established from rock sample observation. A stepwise process was then applied to assign the molecular ratios to the mineral set. Unique associations between oxides and certain minerals were resolved first in this procedure, and then the remainder allocated between other components.

Although estimates of standard rock-forming minerals such as quartz or calcite are generally close to volumes observed visually, clay mineral estimation is often problematic when matched with results from X-ray diffraction analysis. The discrepancy is also not surprising because clay-mineral compositions are highly variable, as a result of isomorphous substitution (Imbrie and Poldervaart 1959).

Mineral Compositions Computed from Wireline Petrophysical Logs

The calculation of mineral volumes in the subsurface was originally a byproduct of methods to improve estimates of porosity in multimineral reservoir rocks. Lithologies composed of several minerals required several porosity logs to be run in combination in order to estimate volumetric porosity. In the simplest solution model, the proportions of multiple components together with porosity could be estimated from a set of simultaneous equations for the measured log responses. The earliest application of this technique was in Permian carbonate reservoirs in West Texas where measurements from density, neutron, and sonic logs were used to solve for volumes of dolomite, anhydrite, gypsum, and porosity (Savre 1963).

Compositional analysis by the inverse solution is simple and robust. The unity equation dictates that computed proportions must collectively sum to unity. However, individual proportions can be negative or take values greater than unity. These unreasonable numbers are not caused by estimation error, but simply reflect the location of the sample outside the bounds of the composition space as set by the endmember mineral coordinates. Such results can be used as an informal feedback that may suggest the presence of unaccounted minerals or a need to adjust mineral endmember properties. By the same token, a satisfactory solution does not “prove” a composition, but rather a solution that is consistent with the measurements.

The inversion model can be written as a set of simultaneous equations where the number of knowns and unknowns are matched in a “determined system.” However, when the number of logs is not sufficient for a unique solution, then the result is an underdetermined system. Alternatively, if the number of logs equals or exceeds the number of components, this becomes an overdetermined system. Both

underdetermined and overdetermined systems can be resolved by an expansion of the matrix algebra formulation of the determined system solution (Doveton 1994).

Modern compositional analysis methods have advanced beyond basic inversion models, with the incorporation of tool error functions and constraints as integral components of the solution strategy. In particular, optimal solutions were set with the goal of minimizing the incoherence between log measurements and their values predicted from compositional analysis (Mayer and Sibbit 1980). Not all component log responses need to be estimated since their properties are restricted to a limited range. However, some mineral components have ambiguous properties. The most example of such components are clay minerals, whose composition is widely variable. Calibration to clay minerals in local core is the best practice to reduce estimation uncertainties (Quirein et al. 1986).

Estimation of Minerals in the Subsurface from Geochemical Logging

The use of conventional logs for compositional analysis is implicit in the sense that physical properties are used to deduce mineral proportions. In contrast, elements measured by geochemical logs can be tied directly to mineral compositions. Natural gamma-ray spectroscopy measurements provide potassium, uranium, and thorium while neutron capture spectroscopy supplies silicon, calcium, iron, sulfur, titanium, and the rare earth, gadolinium. While not measured directly, magnesium is generally inferred from the photoelectric factor.

Because there will generally be more minerals than elements to resolve them, the situation is mostly underdetermined. But in reality, most sedimentary rocks are dominated by only ten minerals: quartz, two carbonates, three feldspars, and four clays. Consequently, acceptable compositional solutions can be created with relatively small assemblages of minerals, as long as they have been identified appropriately and that the compositions used are consistent and reasonably accurate (Herron 1988). The solution is aided when some of the elements are associated explicitly with specific minerals. The partition of minerals between species then follows an ordered protocol of assignments. However, the application of simultaneous equations to resolve mineral compositions from elemental measurements provides a more general and effective procedure (Harvey et al. 1990).

As a result of the explicit relations between elemental measures and mineral compositions, much improved mineral transforms have been developed than those of inference from mineral physical properties. The fundamental goal of an effective transform is to provide a close match between normative mineral solutions and modal mineral suites. This aim distinguishes it from the traditional norm of igneous rocks

which is based on hypothetical crystallization and only loosely tied to modal observations (Cross et al. 1902). However, a unique mineral transform from element concentrations may not always be possible in situations of compositional colinearity. If the model is exactly colinear, then there are an infinite range of solutions, creating a matrix singularity and an inversion breakdown (Harvey et al. 1990).

In recent years, methods to resolve subsurface mineralogy estimation have increasingly relied on machine learning algorithms, building on pioneer work that applied simple artificial neural networks (Doveton 2014).

Summary

Mathematical minerals are calculated from chemical or physical measurements of a rock sample as contrasted with actual minerals that are observed. The earliest methods were developed in igneous petrology, where “normative” minerals are computed from oxide analyses, based on a hypothetical and idealized protocol that mimics the order of mineral differentiation. These norms are only a general approximation of “modal” minerals, but are useful for classification purposes. By contrast, normative minerals in sedimentary rocks are computed to be a best match with observed mineral suites. The most common applications today occur in the subsurface, where downhole petrophysical measurements are used to estimate mineralogy by a variety of statistical procedures. The old terminology of “normative minerals” has been mostly replaced by the term of “mathematical minerals.”

Cross-References

- [Artificial Neural Network](#)
- [Compositional Data](#)
- [Forward and Inverse Stratigraphic Models](#)
- [Machine Learning](#)
- [Porosity](#)
- [Statistical Rock Physics](#)

Bibliography

- Cross W, Iddings JP, Pirsson LV, Washington HS (1902) A quantitative chemico-mineralogical classification and nomenclature of igneous rocks. *J Geol* 10:555–690
- Doveton JH (1994) Geological log analysis using computer methods: computer applications in geology, No. 2. American Association of Petroleum Geologists, Tulsa, 169 pp
- Doveton JH (2014) Principles of mathematical petrophysics. Oxford University Press, 267 pp
- Harvey PK, Bristow JF, Lovell MA (1990) Mineral transforms and downhole geochemical measurements. *Sci Drill* 1(4):163–176

- Herron MM (1988) Geochemical classification of terrigenous sands and shales from core or log data. *J Sediment Petrol* 58(5):820–829
- Imbrie J, Poldervaart A (1959) Mineral compositions calculated from chemical analyses of sedimentary rocks. *J Sediment Petrol* 29(4): 588–595
- Krumbein WC, Pettijohn FJ (1938) *Manual of sedimentary petrography*. Appleton-Century-Crofts, New York, 549 pp
- Mayer C, Sibbit A (1980) GLOBAL: a new approach to computer-processed log interpretation. SPE Paper 9341, 12 pp
- Quirein J, Kimminau S, Lavigne J, Singer J, Wendel F (1986) A coherent framework for developing and applying multiple formation evaluation models, paper DD. In: 27th annual logging symposium transactions. Society of Professional Well Log Analysts, 16 pp
- Savre WC (1963) Determination of a more accurate porosity and mineral composition in complex lithologies with the use of the sonic, neutron, and density surveys. *J Pet Technol* 15:945–959

Mathematical Morphology

Jean Serra

Centre de morphologie mathématique, Ecoles des Mines,
Paristech, Paris, France

Definition

Mathematical morphology is a discipline of image analysis that was introduced in the mid-1960s. From the very start it included not only theoretical results but also research on the links between geometrical textures and physical properties of the objects under study. Initially conceived as a set theory, with deterministic and random versions, it rapidly developed its concepts in the framework of complete lattices. In geosciences and in image processing, theoretical approaches may rest, or not, on the idea of reversibility. In the first case, addition is the core operation. It leads to convolution, wavelets, Fourier transformation, etc. But the visual universe can be grasped differently, by remarking that the objects are solid and not transparent. An object may cover another, or be included in it, or hit it, or be smaller than it, etc., notions which all refer to set geometry, and which all imply some loss of reversibility. Indeed, the morphological operations, whose core is the order relation, always loose information, and the main issue of the method consists in managing this loss.

Introduction

We will begin by a comment. The petrographic example of Fig. 4 is directly the matter of geosciences. Other ones can easily be transposed. The method for detecting the changes of direction in a piece of oak (Fig. 6) also applies to problems of tectonics, and the way to estimate the roughness of a road

before and after wear (Fig. 7) can serve to quantify the erosion process after an orogeny.

Mathematical morphology, which covers both a theory and a practice, appeared in the middle of the years 1960 with the pioneer works of G. Matheron and J. Serra. They aimed at the description of petrographic structures seen under the microscope (porous media, crystals, alloys, etc.), in order to link them to their physical properties (Matheron and Serra 2002). Objects were considered as sets, that were described either by transforming them (e.g., granulometries), or by modeling them in a probabilistic way (e.g., Boolean closed sets). We present here only the first approach, which extended itself through the years 1970 to numerical functions, under the impulse of medical imaging (Serra 1982). During the years 1980, morphological operations acting on more and more varied objects, the need for unification brought out the concept of *complete lattice* as a common framework (see Serra 1988 in the article “Morphological Filtering”). This algebraic structure lies at the basis of lattice theory, a branch of algebra introduced by the end of the nineteenth century by mathematicians such as E.H. Moore, to whom we owe the crucial notion of a set-theoretical closing, and G. Birkhoff, whose fundamental book has constantly been republished from 1940 until the present time.

Algebraists have much studied the structure of lattices, but have nevertheless given scant attention to the concepts of complete lattice and of lattices of operators: for example in Birkhoff’s book, only one (short) chapter among 17 is devoted to them. Though more recent works written by theoretical computer scientists tackle complete lattices and basic operators such as dilations and openings, nevertheless the idea of describing geometrical features by means of lattices never appears in these studies. The chief notions on which Mathematical Morphology rests, such as morphological filters, segmentation by connections, flat operators, the watershed, levelings, etc., are totally absent from them. In fact, the common logical basis of the lattice has been used in mathematical morphology as starting point for exploring new directions. Since the decade of the 1990s, following the works by H. Heijmans, G. Matheron, Ch. Ronse, and J. Serra, among others (Serra 1988; Heijmans and Ronse 1990; Heijmans 1994; Matheron 1996), the theoretical core of mathematical morphology is developing in an autonomous way. Here we base ourselves on these latter references. A comprehensive bibliography for general morphology is given in Najman and Talbot (2010), and for old mathematical references refer Kiselman (2020) and Ronse and Serra (2010).

Notation: The generic symbol for a complete lattice is $\mathcal{L}$, with some curvilinear variants for the most usual lattices: $\mathcal{P}(E)$ for the family of subsets of the set E , $\mathcal{D}(E)$ for that of partitions of E , $\mathcal{F}$ for the lattice of functions, among others. More generally, curvilinear capital letters designate families of

sets, such as $\mathcal{B}$ for that of invariants of an opening, or a connection $\mathcal{C}$ on $\mathcal{P}(E)$.

Elements of a lattice are generally denoted by lowercase letters; however, we will use capital letters for the elements of the lattices $\mathcal{P}(E)$ and T^E (respectively of parts of E and of numerical functions $E \rightarrow T$), in order to distinguish them from the elements of E (points) or of T (numerical values), denoted by lowercase letters.

Operators acting on lattices are denoted by lowercase Greek letters, of which, in particular, δ , ε , γ , and φ for a dilation, an erosion, an opening, and a closing, respectively. There are some exceptions, for example, **id** for the identity.

Complete Lattices

Partially Ordered Sets

Order relation: Provide the set $\mathcal{L}$ with a binary relation $\leq$ that satisfies the following properties:

$$x \leq x \text{ (reflexivity),}$$

$$x \leq y \text{ and } y \leq x \text{ imply } x = y \text{ (anti-symmetry),}$$

$$x \leq y \text{ and } y \leq z \text{ imply } x \leq z \text{ (transitivity),}$$

for all $x, y, z \in \mathcal{L}$. Then the relation $\leq$ is called a *partial order* and the set $\mathcal{L}$ is *partially ordered* (in short, *p.o.*) by relation $\leq$. The order becomes total when:

$$\forall x, y \in \mathcal{L}, \quad x \leq y \text{ or } y \leq x.$$

Examples (a) The set N of all numbers of the open interval $[0, 1]$ is totally ordered for the usual numerical order. The set of all points of R^d equipped with the relation “ $(x_1, x_2, \dots, x_d) \leq (y_1, y_2, \dots, y_d)$ when $x_i \leq y_i$ for all i ” is partially ordered.

(b) The set $\mathcal{P}(E)$ of all subsets of a set E is partially ordered for the inclusion order $\subseteq$. Let us remark that unlike the set N of the previous example, $\mathcal{P}(E)$ admits a greatest element, E itself, and a least one, namely the empty set $\emptyset$.

Duality: Denote by $(\mathcal{L}, \leq)$ the set $\mathcal{L}$ provided with the order relation $\leq$, and define the relation $\leq^*$ by $x \leq^* y$ if and only if $y \leq x$. This relation $\leq^*$ is an order on $\mathcal{L}$, and $(\mathcal{L}, \leq^*)$ is the *dual* ordered set of $(\mathcal{L}, \leq)$. To each definition, each statement, etc., relative to $(\mathcal{L}, \leq)$, there corresponds a dual notion in $(\mathcal{L}, \geq)$, which is obtained by inverting $\leq$ and $\geq$. This duality principle, which seems to be obvious, plays a fundamental role in mathematical morphology. When an operation has been introduced, it suggests to look also at the dual version, then at the product of both, etc. (for instance: opening and closing).

Complete Lattices

The theory of mathematical morphology rests on the structure of complete lattice. Let $\mathcal{L}$ be a p.o. set, and $\mathcal{K} \subseteq \mathcal{L}$. An element $a \in \mathcal{L}$ is called a *lower bound* of $\mathcal{K}$ if $a \leq x$ for all $x \in \mathcal{K}$, whether a belongs to $\mathcal{K}$ or not. If the family of lower bounds of $\mathcal{K}$ admits a greatest element a_0 , it defines the *infimum* of $\mathcal{K}$. By duality, one introduces the two notions of *upper bound* and *supremum*. The *infimum* (resp., the *supremum*) of a subset $\mathcal{K}$, if it exists, is unique, and is denoted by $\inf \mathcal{K}$ or $\wedge \mathcal{K}$ (resp., $\sup \mathcal{K}$ or $\vee \mathcal{K}$). If $x_i \in \mathcal{L}$ for a (possibly infinite) family I of indexes i , one writes $\wedge_{i \in I} x_i$ for $\wedge \{x_i \mid i \in I\}$, and $\vee_{i \in I} x_i$ for $\vee \{x_i \mid i \in I\}$.

Lattice: A p.o. set $\mathcal{L}$ is a *lattice* when any nonempty finite subset of $\mathcal{L}$ admits an *infimum* and a *supremum*. The lattice is said to be *complete* when this property remains true for all nonempty subsets of $\mathcal{L}$, whether they are finite or not.

By definition, every complete lattice $\mathcal{L}$ has a least element $\mathbf{0} = \wedge \mathcal{L}$ and a greatest element $\mathbf{1} = \vee \mathcal{L}$. These two extreme elements are the *universal bounds* of $\mathcal{L}$. A subset $\mathcal{M}$ of the complete lattice $\mathcal{L}$ constitutes a *complete sublattice* of $\mathcal{L}$ when the *infimum* and the *supremum* of any family in $\mathcal{M}$ belong to $\mathcal{M}$, and when $\mathcal{M}$ contains the universal bounds of $\mathcal{L}$.

Anamorphosis: The term “anamorphosis” appears in painting during the Renaissance, to designate geometrical distortions of figures in the Euclidean plane R^2 that preserve the order of $\mathcal{P}(R^2)$, so that they permit to find back the initial image from its deformation. But this kind of operation appears in many other lattices. So, the map $x \mapsto \log x$, for $x \geq 0$, is an anamorphosis from $\overline{R}_+ \rightarrow \overline{R}$, which extends to numerical functions on E . Formally speaking, it can be defined as follows (Matheron 1996):

Let $\mathcal{L}, \mathcal{M}$ be two complete lattices. A map $\alpha : \mathcal{L} \rightarrow \mathcal{M}$ is an *anamorphosis*, when α is a bijection, and when α and its inverse α^{-1} preserve the order, that is:

$$\forall x, y \in \mathcal{L}, \quad x \leq y \text{ if and only if } \alpha(x) \leq \alpha(y) \quad (1)$$

The most popular morphological operators on gray tone functions are the so-called flat operators, because they commute under anamorphosis.

A map $\mathcal{L} \rightarrow \mathcal{L}$ which coincides with its own inverse ($\alpha^{-1} = \alpha$), like a reflection, is called an *inversion*, or an *involution*, of $\mathcal{L}$.

Remarkable Elements and Families

We now present some remarkable elements or particular subsets that can be found in complete lattices:

Sup-generating family: Let $\mathcal{L}$ be a complete lattice, and $\mathcal{S} \subseteq \mathcal{L}$ a family in $\mathcal{L}$. The class $\mathcal{S}$ is *sup-generating* when every element $a \in \mathcal{L}$ is the *supremum* of the elements of $\mathcal{S}$ smaller than itself:

$$a = \vee \{x \in \mathcal{S} \mid x \leq a\}.$$

Atom: An element $a \neq \mathbf{0}$ of a complete lattice $\mathcal{L}$ is an *atom* when $x \leq a$ implies that $x = \mathbf{0}$ or $x = a$. Any sup-generating family necessarily comprises all the atoms. The lattice $\mathcal{L}$ is *atomistic* when the family of all atoms is sup-generating.

For example, for $\mathcal{L} = \mathcal{P}(E)$, the singletons are atoms, hence the lattice is atomistic.

Co-prime: An element $a \neq \mathbf{0}$ of a complete lattice $\mathcal{L}$ is *co-prime* when $a \leq x \vee y$ implies that $a \leq x$ or $a \leq y$, in a nonexclusive manner. The lattice $\mathcal{L}$ is *co-prime* if the family of its co-primes is sup-generating.

Complement: Let $\mathcal{L}$ be a complete lattice, with universal bounds $\mathbf{0}$ and $\mathbf{1}$. When $x, y \in \mathcal{L}$ are such that $x \wedge y = \mathbf{0}$ and $x \vee y = \mathbf{1}$, they are said to be *complements* of each other. The lattice $\mathcal{L}$ is *complemented* when each of its elements admits at least one complement.

Note that a same element may have several complements. For example, if $\mathcal{L}$ is the lattice of all vector subspaces of R^2 , ordered by inclusion, the complements of a monodimensional subspace are all the other subspaces of dimension 1, whose number is infinite.

Distributivity

Though many properties involve distributivity, this notion is not always simple, because it can be stated at three different levels. Below, we introduce the finite case only. For the two other ones, the reader may refer to Heijmans (1994) and Matheron (1996).

A lattice $\mathcal{L}$ is *distributive* when:

$$\forall x, y, z \in \mathcal{L}, \quad x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z),$$

or equivalently:

$$\forall x, y, z \in \mathcal{L}, \quad x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z).$$

These equalities do not always extend to the case when the collection of those elements inside the first parentheses is infinite.

The notions of distributivity, complementation, atom, and co-prime are closely related (Matheron 1996).

Theorem *In a distributive lattice, the complement of an element, if it exists, is unique and all atoms are co-primes. In a complemented lattice, all co-primes are atoms. Any co-prime complete lattice is infinitely inf-distributive.*

Monotone Convergence and Continuity

Independently of any topology, one can already introduce some convergence and continuity on complete lattices, by relying on their algebraic structure only.

Monotone Limit: Every increasing (resp., decreasing) sequence $x_i (i \in I \subseteq \mathbb{N})$ defines by its *supremum*

(resp. *infimum*) the *sequential monotone limit* $x = \vee_{i \in I} x_i$ (resp. $x = \wedge_{i \in I} x_i$). One then writes $x_i \uparrow x$ (resp. $x_i \downarrow x$). An increasing map $\psi : \mathcal{L} \rightarrow \mathcal{M}$ between the complete lattices $\mathcal{L}$ and $\mathcal{M}$ is $\uparrow$ -continuous when:

$$x_i \uparrow x \text{ in } \mathcal{L} \Rightarrow \psi(x_i) \uparrow \psi(x) \text{ in } \mathcal{M},$$

and $\downarrow$ -continuous when:

$$x_i \downarrow x \text{ in } \mathcal{L} \Rightarrow \psi(x_i) \downarrow \psi(x) \text{ in } \mathcal{M}.$$

The continuity of an Euclidean map is the touchstone of its digitability. In R^2 for example, one can cover any set X by small squares of sides a_i centered at the vertices of a grid, of union X_i . If map ψ is increasing and $\downarrow$ -continuous, then the digital transforms of the X_i tend toward that of the Euclidean set X as $a_i \rightarrow 0$ (Heijmans and Serra 1992; Matheron 1996).

Boolean Lattices

G. Boole modeled the logic of propositions by an algebra with two binary operations, namely disjunction and conjunction, and a unary operation, namely negation, which have the usual property of distributivity. His name was thus given to the corresponding class of lattices:

Boolean Lattice A distributive and complemented lattice, complete or not, is said to be *Boolean*.

In distributive lattices, each element has at most one complement, and in the Boolean case it is unique. This results in the complementation operation on that type of lattice; just as negation in logic, the complementation is an involution:

Theorem Boolean complementation *In a Boolean lattice $\mathcal{L}$, the operation of complementation $x \mapsto x^*$ is an involution of $\mathcal{L}$:*

$$\forall x, y \in \mathcal{L}, \quad x^{**} = x \text{ and } [x \leq y] \Leftrightarrow [x^* \geq y^*].$$

We then have De Morgan's law: for any finite family $x_i (i \in I)$ in $\mathcal{L}$, we have

$$\left(\bigwedge_{i \in I} x_i \right)^* = \bigvee_{i \in I} x_i^* \text{ and } \left(\bigvee_{i \in I} x_i \right)^* = \bigwedge_{i \in I} x_i^*,$$

and when the lattice $\mathcal{L}$ is complete, these two equalities remain valid for the infinite families.

Examples of Lattices

Lattices of Sets

This section is devoted to a few lattices whose elements are subsets of a given arbitrary set E .

Lattices of $\mathcal{P}(E)$ type: The family $\mathcal{P}(E)$ of all subsets of E , ordered by inclusion, constitutes a complete lattice, where the *supremum* and the *infimum* are given by the union and the intersection. This lattice is distributive. Moreover, for all $X \in \mathcal{P}(E)$, the set X^c of those points of E that do not belong to X satisfies the two conditions $X \cap X^c = \emptyset$ and $X \cup X^c = E$; hence X^c turns out to be a complement of X , and it is unique. Therefore $\mathcal{P}(E)$ is Boolean, hence infinitely distributive. The points of E are co-prime atoms, and they form a sup-generating family of $\mathcal{P}(E)$.

One can wonder which ones of all these nice properties suffices to characterize a lattice of $\mathcal{P}(E)$ type. Any finite Boolean lattice is of this type, but it is not always true in the infinite case (see the lattice of the regular closed sets described below). There exist several characterizations of the lattice $\mathcal{P}(E)$, we give here two only.

Theorem *Let $\mathcal{L}$ be a complete lattice. Each of the following two properties is equivalent to the fact that $\mathcal{L}$ is isomorphic to $\mathcal{P}(E)$ for some set E :*

1. $\mathcal{L}$ is co-prime and complemented (Matheron 1996, p. 179).
2. $\mathcal{L}$ is Boolean and atomistic (Heijmans 1994).

Moreover, the set E is uniquely determined (up to an isomorphism): by the co-primes in (1), and by the atoms in (2).

Lattice of convex sets: The set-theoretical lattices are far from being reduced to $\mathcal{P}(E)$ and its sublattices. Here is an example where the order is still the inclusion and the *infimum* the intersection, but where the *supremum* is no longer the union. Remember that a set $X \subseteq R^n$ is *convex* when for all pairs of points $x, y \in X$ the segment $[x, y]$ is contained in X (the definition includes the singletons and the empty set). The class of convex sets is closed under intersection, and partially ordered for inclusion, which induces a complete lattice having the intersection for *infimum*. For the *supremum*, we must consider the intersection of all convex sets that contain $X \subseteq R^2$, that is, its *convex hull* $co(X)$. The *supremum* of the family X_i is then given by the convex hull of their union:

$$\bigvee_{i \in I} X_i = co\left(\bigcup_{i \in I} X_i\right).$$

This lattice is atomistic, but neither complemented, nor distributive.

Lattice of regular closed set: A typical example of Boolean complete lattice is given by the set, ordered by inclusion, of all *regular closed set* of a topological space, namely of

those sets equal to the closure of their interior: $F = \overline{F^\circ}$. The operations of *supremum*, of *infimum*, and of complementation take here the following form:

$$\bigvee_{i \in I} F_i = \overline{\bigcup_{i \in I} F_i} = \left(\bigcup_{i \in I} F_i\right)^\circ, \quad \bigwedge_{i \in I} F_i = \overline{\bigcap_{i \in I} F_i} = \left(\bigcap_{i \in I} F_i\right)^\circ.$$

This complete lattice is Boolean, and we have $\overline{F^c} = (F^\circ)^c$. Regular sets are the only ones which can correctly be approximated by regular grids of points. In geometrical modeling used for image synthesis, objects are represented by regular closed sets, and the Boolean operations on these objects take the above form. For example, in R^3 a closed cube (i.e., that contains its frontier) is a regular closed set, and given two cubes that intersect on a face or on an edge, the *infimum* of the two will be empty, as their intersection has an empty interior.

Lattices of Numerical Functions

Complete chain: One calls a *complete chain* any complete lattice which is totally ordered, as typically the completed Euclidean line $\overline{R} = R \cup \{-\infty, +\infty\}$, or its discrete version $\overline{Z} = Z \cup \{-\infty, +\infty\}$. The sets $\overline{R}_+$ and $\overline{Z}_+$ restrictions of the previous ones to numbers ≥ 0 , are also complete chains. The three chains $\overline{R}$, $\overline{R}_+$, and $\overline{Z}_+$ are isomorphic, but $\overline{Z}$ and $\overline{Z}_+$ are not. In $\overline{R}$ the two operations of numerical *supremum* and *infimum* are continuous for the ordering topology. When the distinction between the various types of complete chains is superfluous, one usually takes the symbol T for representing them in a generic manner.

Lattices $\overline{R}^E$ and $\overline{Z}^E$ of numerical functions: Let E an arbitrary set. Equip the class $\overline{R}^E$ of real functions $F: E \rightarrow \overline{R}$ with the partial order: $F \leq G$ if for all $x \in E$, $F(x) \leq G(x)$. This order induces a complete lattice where the *supremum* and *infimum* are given by:

$$\begin{aligned} G = \bigvee_{i \in I} F_i &\Leftrightarrow \forall x \in E, G(x) = \sup_{i \in I} F_i(x), \\ H = \bigwedge_{i \in I} F_i &\Leftrightarrow \forall x \in E, H(x) = \inf_{i \in I} F_i(x). \end{aligned} \quad (2)$$

This is nothing but a *power* of the lattice $\overline{R}$, that is, a product of lattice $\overline{R}$ by itself $|E|$ times, hence the expression of *power lattice*, and the notation $\overline{R}^E$. The *pulse* functions $u_{x,t}$ defined by:

$$u_{x,t}(y) = t \text{ if } y = x, \quad u_{x,t}(y) = -\infty \text{ if } x \neq y,$$

of parameters $x \in E$ and $t \in R$, are co-primes elements but not atoms, and they constitute a sup-generating family of $\overline{R}^E$; the same notion applies to $\overline{Z}^E$ by taking the pulses $u_{x,t}$ for

$t \in Z$. Both lattices $\overline{R}^E$ and $\overline{Z}^E$ are infinitely distributive, but not complemented.

Note in passing that we must start from the completed line $\overline{R}$, which is a complete chain (or from one of its closed sublattices $\overline{R}_+$, $[0, 1]$, $\overline{Z}$, $\overline{Z}_+$, etc.), and not from the line R , for the latter is not a complete chain. Indeed, equation (2) implies that $G(x)$ might be equal to $+\infty$, or that $H(x)$ might be equal to $-\infty$, even when all $F_i(x)$ are finite.

The lattice $\mathcal{P}(E)$ of sets, ordered by inclusion, is isomorphic to lattice 2^E of all binary functions $E \rightarrow \{0, 1\}$, equipped with the numerical ordering.

Lattices of the Lipschitz functions: The class of all numerical functions is too comprehensive for modeling images of the physical world, and has too many pathological cases unsuitable for discrete approximations. On the other hand, the class of all continuous functions does not form a complete lattice: for example, the functions x^n , $n \geq 0$, taken between 0 and 1 are continuous, whereas $y = \bigwedge_n \geq 0 x^n$ is not. The upper (or lower) semi-continuous functions on R^n do form a lattice, but it is not closed under subtraction, and its *supremum*, namely the topological closure of the usual supremum, is not pointwise. In addition, these functions, which may have no minimum, are inadequate candidates for operations such as the watershed, for example. We must find something else.

When the starting space E is metric of distance d , the functions $F : E \rightarrow \overline{R}$ that are the most convenient for morphological operators are undoubtedly the Lipschitz functions, or their extensions under anamorphoses. The Lipschitz functions of module ω satisfy the inequality:

$$\forall x, y \in E, |F(x) - F(y)| \leq \omega \times d, \quad (3)$$

The class $\mathcal{F}_\omega$ of these functions has three basic properties; for every module ω :

1. The constant functions belong to $\mathcal{F}_\omega$
2. $a \in R$ and $F \in \mathcal{F}_\omega$ imply that $a + F \in \mathcal{F}_\omega$
3. $F \in \mathcal{F}_\omega$ implies that $-F \in \mathcal{F}_\omega$

Conversely (Matheron 1996), every complete sublattice $\mathcal{L}$ of $\overline{R}^E$ satisfying these three properties and $\sup_{F \in \mathcal{L}} |F(y) - F(x)| < +\infty$ for all $x, y \in E$, must be Lipschitz for some metric on E .

When we state the incredible series of properties that the $\mathcal{F}_\omega$ do satisfy, one wonders by which miracle they fit so well with the morphological requirements (Serra 1992; Matheron 1996):

1. If function $F \in \mathcal{F}_\omega$ is finite at one point $x \in E$, then it is finite everywhere.

2. For each module ω , the $\mathcal{F}_\omega$ functions form a complete sublattice of the lattice $\overline{R}^E$ of all numerical functions on $\overline{R}$, hence with pointwise sup and inf.
3. Each $\mathcal{F}_\omega$ is a compact space for both topologies of pointwise convergence, and of uniform convergence, identical here.
4. When the space E is affine, each lattice $\mathcal{F}_\omega$ is closed under any (non-necessarily flat) dilation or erosion which is invariant under translation, and these operations are continuous. For example, if $E = Z^n$ or R^n , and if B stands for the unit ball, then for $F \in \mathcal{F}_\omega$, *Beucher's gradient* $\frac{1}{2}[(F \oplus B) - (F \ominus B)] \in \mathcal{F}_\omega$.
5. If $\{K(x) | x \in E\}$ is a family of variable compact structuring elements, whose Hausdorff distance h satisfies the inequality $h[K(x), K(y)] \leq d(x, y)$, then $\mathcal{F}_\omega$ is closed under flat dilation and erosion according to the $K(x)$, and these operations are continuous.
6. The two above properties extend to *suprema*, to *infima*, and to finite composition products of dilation and erosion.
7. If $g(dy)$ is a measure such that $\int_E |g(dy)| \leq 1$, then the lattice $\mathcal{F}_\omega$ is closed under convolution by g , and this operation is continuous.
8. When E and $\overline{R}$ are sampled by means of regular grids, then the previous operations can be arbitrarily approximated by their digital versions (a consequence of the continuities).
9. Since each $\mathcal{F}_\omega$ is a compact space, one can define probabilities on it, and generalize the theory of the random sets.
10. The finite products of lattices $\mathcal{F}_\omega$ still satisfy all above properties (e.g., multispectral images).

Lattice of Partitions

Now here is a lattice whose objects are no longer sets of points, or numerical functions, but cuttings of space.

Definition Partition *Let E be an arbitrary space. A partition of E is a family $\mathbf{P}$ of subsets of E , called classes, which are:*

1. *Non-empty:* $\emptyset \notin \mathbf{P}$
2. *Mutually disjoint:* $\forall X, Y \in \mathbf{P}, X \neq Y \Rightarrow X \cap Y = \emptyset$
3. *Whose union covers E :* $\bigcup \mathbf{P} = E$

Equivalently, the partition corresponds to a map $D : E \rightarrow \mathcal{P}(E)$ that satisfies the following two conditions:

- for all $x \in E, x \in D(x)$
- for all $x, y \in E, D(x) \cap D(y) \neq \emptyset \Rightarrow D(x) = D(y)$

$D(x)$ is called the *class* of the partition in x .

Partitions intervene in *image segmentation* (see article “*Morphological Filtering*”). To segment a graytone or color image, one partitions its space of definition into zones which are homogeneous in some sense. Therefore, it should be useful to handle these partitions just as we do with sets or with functions.

The set $\mathcal{D}(E)$ of partitions of E is partially ordered by *refinement*: we say that partition D_f is *finer* than partition D_g , or that D_g is *coarser* than D_f , and we write $D_f \leq D_g$, when each class of D_f is included in a class of D_g :

$$D_f \leq D_g \Leftrightarrow \forall x \in E, D_f(x) \subseteq D_g(x).$$

The set $\mathcal{D}(E)$, provided with this order is a complete lattice; the greatest element (the coarsest partition) is the *universal partition* D_1 whose unique class is E , whereas the least element (the finest partition) is the *identity partition* D_0 whose classes are all singletons:

$$\forall x \in E, D_1(x) = E \text{ and } D_0(x) = \{x\}.$$

The class at point x of the *infimum* of a family $\{D_i | i \in I\}$ of partitions is nothing but the intersection of the classes $D_i(x)$:

$$\forall x \in E, \left[\bigwedge_{i \in I} D_i \right](x) = \bigcap_{i \in I} D_i(x).$$

The *supremum* of partitions $D_i (i \in I)$ is less straightforward. Formally, it is the finest partition D such that $D_i(x) \subseteq D(x)$ for all $i \in I$ and $x \in E$, see Fig. 1.

The lattice $\mathcal{D}(E)$ of partitions of E is not distributive. However, it is complemented and atomistic, the atoms being those partitions where exactly one class is a pair $\{x_1, x_2\}$, all other classes being singletons.

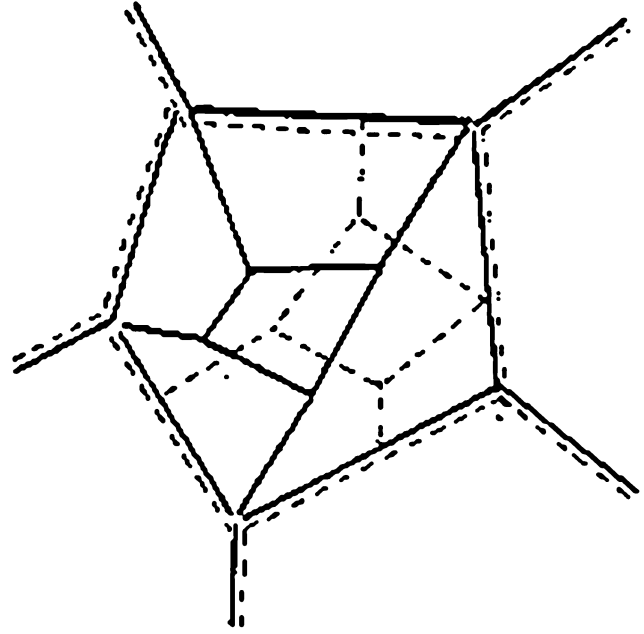
Lattices of Operators

We just modeled objects and cuttings. But we can as well represent as lattices most of the families of morphological operations, which is a major achievement of Mathematical Morphology. In the set-theoretical case, for example, the family $\mathcal{A}$ of all maps from $\mathcal{P}(E)$ into itself is ordered by the following relation between elements α, β of $\mathcal{A}$:

$$\alpha \leq \beta \Leftrightarrow \forall X \in \mathcal{P}(E), \alpha(X) \subseteq \beta(X).$$

This induces the following complete lattice with obvious *supremum*:

$$\alpha = \bigvee_{i \in I} \alpha_i \Leftrightarrow \forall X \in \mathcal{P}(E), \alpha(X) = \bigcup_{i \in I} \alpha_i(X),$$



Mathematical Morphology, Fig. 1 Supremum of two partitions

and *simili modo* for the *infimum*. The universal bounds are the constant operators $X \mapsto E$ and $X \mapsto \emptyset$. The lattice $\mathcal{A}$ is indeed $\mathcal{P}(E)$ at power $\mathcal{P}(E)$; it is therefore Boolean and atomistic.

Duality by complementation: A new duality appears here, different from that induced by the order. As $\mathcal{P}(E)$ is complemented, we can associate with any element $\alpha \in \mathcal{A}$ its *dual* α^* for the complementation by means of the relation

$$\forall X \in \mathcal{P}(E), \alpha^*(X) = [\alpha(X^c)]^c. \quad (4)$$

Operator α^* has all properties dual (in the sense of the order) of those of α ; in particular

$$\alpha_1 \leq \alpha_2 \Leftrightarrow \alpha_1^* \geq \alpha_2^*, \quad \left[\bigvee_{i \in I} \alpha_i \right]^* = \bigwedge_{i \in I} [\alpha_i^*].$$

Note that $(\alpha^*)^* = \alpha$, that is, an operator is the dual of its dual.

We can now generalize and consider the set $\mathcal{O}$ of all operators $\mathcal{L} \rightarrow \mathcal{L}$ for an arbitrary complete lattice $\mathcal{L}$. This set is ordered like $\mathcal{A}$ and has a similar structure of complete lattice. We are thus able to consider operators having the various properties defined below (increasingness, extensivity, closing, etc.) on classes of objects, for example, functions and partitions, as soon as the latter are themselves structured in complete lattices.

Many lattices are not isomorphic to their dual in the sense of the order, and therefore do not admit an inversion, for example: the lattice of convex sets, that of partitions, etc. It means that the duality by complementation cannot be extended to this type of lattice. In practice, in these lattices,

dual operations, for example, opening and closing, will look very dissimilar.

Duality by inversion: In the lattices of numerical functions $E \rightarrow \bar{R}$, for each real constant t , the map $N : \bar{R}^E \rightarrow \bar{R}^E : F \mapsto t - F$ is an inversion. The same applies if we take the interval $[a, b]$ instead of $\bar{R}$, provided that we put $t = a + b$. That allows us to establish a duality similar to equation (4). For any operator $\beta : \bar{R}^E \rightarrow \bar{R}^E$, the dual β^* defined by:

$$\beta^*(F) = N(\beta(N(F))) \quad (5)$$

plays a role similar to α^* in the set-theoretical case.

Increasing operator: An operator $\alpha \in O$ is *increasing* when:

$$\forall x, y \in \mathcal{L}, \quad x \leq y \quad \Rightarrow \quad \alpha(x) \leq \alpha(y).$$

The glasses of a myopic person are a typically increasing. We easily see that the increasing operators form a complete lattice O' , a complete sublattice of O .

Extensive and anti-extensive operator: An operator $\alpha \in O$ is *extensive* (resp., *anti-extensive*) when for all $x \in \mathcal{L}$ we have $x \leq \alpha(x)$ (resp., $x \geq \alpha(x)$). Both families of extensive operators and of anti-extensive ones are closed under *non-empty supremum* and *infimum*, but they do not form complete sublattices of O because they do not admit the same universal bounds. For example, the least extensive operator is the identity $\text{id} : x \mapsto x$ whereas the least element of O is the operator $x \mapsto 0$;

Idempotent operator: An operator $\alpha \in O$ is *idempotent* if $\alpha^2 = \alpha$, that is, when

$$x \in \mathcal{L} \quad \alpha(\alpha(x)) = \alpha(x)$$

Idempotence occurs everywhere in physics and in geosciences, and it also appears in the daily life at all times. To look through a yellow glass (supposed to be bandpass), to empty a bottle, to be born, to die are all idempotent operations. One understands why idempotence, associated with increasingness, is at the root of morphological filtering.

One can also consider the complete lattice of operators $\alpha : \mathcal{L} \rightarrow \mathcal{M}$, where $\mathcal{L}$ and $\mathcal{M}$ are two distinct complete lattices; for example, $\mathcal{L} = T^E$ and $\mathcal{M} = \mathcal{P}(E)$, and α is a binarization. Then some notions keep their meaning (e.g., increasingness), other lose it (e.g., (anti-) extensivity, idempotence). Note that the product $\alpha\beta$ supposes that the starting lattice of α coincides with the arrival lattice of β .

Closings and Openings

Openings and closings are at the basis of morphological filtering of images (see the article “Morphological filtering”).

They are also the easiest operators to characterize. We shall thus describe them in the first place.

Closings in complete lattices had already been studied by algebraist mathematicians since the years 1940, like R. Baer, C. Everett, and O. Ore. These works form the basis of the theory developed subsequently in mathematical morphology.

Moore Families and Closings

In mathematics, one encounters numerous examples of objects closed under one operation. For example, a set X in Euclidean space is convex if it is closed under the operation joining any two points by a segment. If the set is not closed, one defines then its closing as the least closed set containing it, namely its convex hull $C(X)$. Remark that $X \subseteq C(X)$, that for any two sets $X, Y \in R^2$ if $X \subseteq Y$, then $C(X) \subseteq C(Y)$, and finally that $C(X) = C[C(X)]$. The convex hull operation is thus extensive, increasing, and idempotent.

As seen by E. Moore in 1910, the two notions of “closed set” and of “closing” correspond to each other: a closed object is the closing of an object, while the closing of an object is the least closed object containing it. This assumes that among all closed objects containing a given object, there exists one that is the least one. In other words, the lattice structure is perfectly well-suited here.

Definition Let $\mathcal{L}$ be a complete lattice, whose least and greatest elements are 0 and 1 .

1. A closing on $\mathcal{L}$ is an operator φ on $\mathcal{L}$ that is extensive, increasing, and idempotent.
2. The invariance domain of an operator ψ on $\mathcal{L}$ is the set

$$\mathcal{B}_\psi = \{x \in \mathcal{L} \mid \psi(x) = x\}.$$

3. A part $\mathcal{M}$ of $\mathcal{L}$ is a Moore family when $\mathcal{M}$ is closed under the operation of infimum and includes 1 .

In fact, the notions of Moore family and of closing represent the same thing under two different angles:

Theorem Closings and Moore families There is a one-to-one correspondence between Moore families in $\mathcal{L}$ and closings on $\mathcal{L}$:

- To a Moore family $\mathcal{M}$ one associates the closing φ defined by setting for every $x \in \mathcal{L}$: $\varphi(x)$ is the least $y \in \mathcal{M}$ such that $y \geq x$.
- To a closing φ one associates the Moore family defined by its invariance domain $\mathcal{B}_\varphi = \{\varphi(x) \mid x \in \mathcal{L}\}$.

Openings

One can take the dual point of view, by inverting the order and transposing *supremum* and *infimum*. The previous definitions and theorems become thus the following statements:

1. One calls an opening on $\mathcal{L}$ an operator γ that is anti-extensive, increasing, and idempotent.
2. One calls a dual Moore family of $\mathcal{L}$ a part $\mathcal{M}$ closed under the operation of supremum and includes $\mathbf{0}$,

As previously, here is a one-to-one correspondence between dual Moore families in $\mathcal{L}$ and openings on $\mathcal{L}$:

- To a dual Moore family $\mathcal{M}$ one associates the opening γ defined by setting for every $x \in \mathcal{L}$: $\gamma(x)$ is the greatest $y \in \mathcal{M}$ such that $y \leq x$.
- To an opening γ one associates its invariance domain $\mathcal{B}_\gamma = \{\gamma(x) | x \in \mathcal{L}\}$ which is a dual Moore family.

Generation of Closings and Openings

One can construct openings and closings by using their structures.

Theorem Structural theorem of openings Let $\mathcal{L}$ be a complete lattice and $\mathcal{L}'$ be the lattice of the increasing operations on $\mathcal{L}$. The set of all openings on $\mathcal{L}$ is closed under the supremum in $\mathcal{L}'$. For any family $\{\gamma_i\}$ of openings, the supremum $\vee_{i \in I} \gamma_i$ is itself an opening. The least opening is the constant operator $x \mapsto \mathbf{0}$, the greatest one is the identity **id**. (dual statements for the closings).

Remark, however, that neither the product $\gamma_1 \gamma_2$ of two openings nor the infimum $\wedge_{i \in I} \gamma_i$ openings are openings. The structural theorem orients us towards the construction of openings by suprema of a few basic ones. In practice indeed, all openings derive from two ones, namely the opening by adjunction defined by Eq. (6) and the connection opening defined by Eq. (2) of the article “Morphological Filtering.”

Dilation and Erosion

Minkowski Operations

In the same way as closings and openings, dilations and erosions represent a pair of fundamental notions in mathematical morphology. Their correspondence is intimately linked to that of *Galois correspondence*, which had already been studied by mathematicians under the name of *adjunction*.

However, the algebraists have not given attention to the properties of lattices that derive from Euclidean space, such as invariance under translation, convexity, and similarity, and missed completely the Minkowski operations. In fact, it is to the school of integral geometry, which knew nothing about lattices, that one owes the primary adjunction of mathematical morphology. In 1903, Minkowski defined the addition of two Euclidean sets by the relation:

$$X \oplus B = \{b + x, x \in X | b \in B\} = \bigcup_{x \in X} B_x = \bigcup_{b \in B} X_b.$$

where B_x (resp. X_b) indicates the translate of B at point x (resp. X at point x). When one fixes B , the operator $\delta_B : X \mapsto X \oplus B$, is a dilation, in the sense that it commutes with the union, and provides even the general form of dilations invariant under translation in $\mathcal{P}(E)$, where $E = \mathbb{R}^n$ or $\mathbb{Z}^n$. The adjoint erosion:

$$\varepsilon_B : X \mapsto X \ominus B = \bigcap_{b \in \check{B}} X_b,$$

where $\check{B} = \{-b | b \in B\}$ is the reflected set of B , is named the Minkowski subtraction, although he himself did not envisage it. It appeared with Hadwiger in 1957, who nevertheless did not envisage, himself, the opening by adjunction.

The eroded $X \ominus B$ is the locus of those points x that B_x is included in X , and the dilate $X \oplus B$ is the locus of those points x that $\check{B}_x$ hits set X (the geological term of “erosion” obviously has another meaning). In the set case, most of the structuring elements met in practice, like the segment, the square, the hexagon, the disc, the sphere, are symmetrical, that is, $\check{B} = B$ and the small hat can just be omitted. Figure 2 (left and center), illustrates two usual Minkowski operations. Note that the erosion by the disk centered about the origin o is anti-extensive, but not that by the pair of points, that does not contain o .

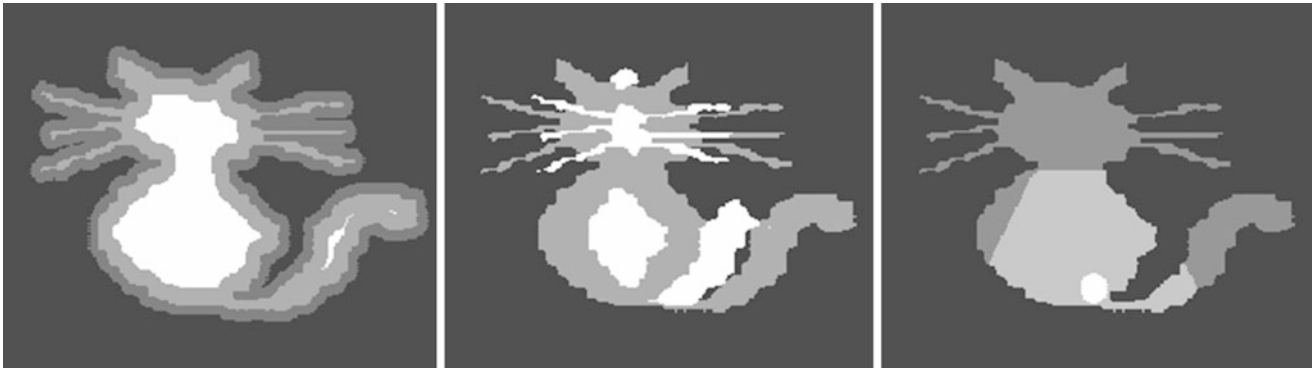
Figure 2 (right) introduces another type of dilation, namely by geodesic disks (Lantuéjoul and Beucher 1981). We present it in the digital case. Consider a digital metric on $\mathbb{Z}^n$, and let $\delta(x)$ stands for the unit ball centered at point x , then the unit geodesic dilation of set Y inside set X is defined by the relation:

$$\delta_X(Y) = \delta(Y) \cap X$$

The dilation of size n of Y inside X is then obtained by iteration:

$$\delta_{X,n}(Y) = \delta(\dots \delta(\delta_X(Y)) \cap X \dots) \cap X \quad n \text{ times}$$

Unlike the Minkowski addition, it is not invariant under translation of Y when the background X is fixed. For n large enough, and if we assumed that set X is regular and bounded,



Mathematical Morphology, Fig. 2 Left: Minkowski addition and subtraction by a disk (the initial cat is the median gray image); center: erosion by a pair of points equidistant from the origin; right: in gray, geodesic dilation of the white spot for the hexagonal metric

the geodesic dilation of point p , considered as a marker, completely invades the connected component of X hit by p .

$$\bigvee_n \delta_{X,n}(X) = \gamma_p(X) \quad (6)$$

The operator γ_p is therefore called the *connection opening at point p* , and is studied in the article “Morphological Filtering.”

It is Matheron (1967) who introduced, in view of granulometries, the opening by adjunction:

$$\gamma_B = \gamma_B \varepsilon_B : X \mapsto (X \ominus B) \oplus B \quad (7)$$

by a structuring element B . The open set $\gamma_B(X)$ is the union of those sets B_x that are included in X . The corresponding Moore family is the invariance domain $\mathcal{B}_\gamma$ of γ_B , which consists of all unions of translates of B . The closing dual for complementation is:

$$\varphi_{\check{B}} = \varepsilon_{\check{B}} \delta_{\check{B}} : X \mapsto (X \oplus \check{B}) \ominus \check{B}.$$

The notations δ_B , ε_B , γ_B , and φ_B mean that one lies in R^n or Z^n and that the operators, with structuring element B , are invariant under translation. For linking openings and granulometries, Matheron bases himself on the following theorem (Matheron, 1975):

Theorem *A family $\{B_\lambda \mid \lambda \geq 0\}$ constitutes the continuous additive semi-group:*

$$B \oplus B_\mu = B_{\lambda+\mu}, \lambda, \mu \geq 0$$

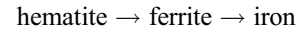
if and only if each B_λ is homothetic with ratio λ of a convex compact B .

In the structure of mathematical morphology, this theorem plays the role of a corner-stone, which distributes the

load in various directions. It leads on the one hand to granulometries, but opens also the door to partial differential equations (Maragos 1996), and finally permits to decompose convex structuring elements for implementing the easily. In particular, any family $\{\gamma_{\lambda B} \mid \lambda \geq 0\}$ of openings by homothetic convex sets turns out to be a granulometry. Figure 3 depicts a granulometry of the black phase by discs, or equivalently increasing closings of the whites.

An Example

In a blast furnace, chemical reductions of the initial pellets of sinters (magnetite or hematite) result in iron, according sequences of the type

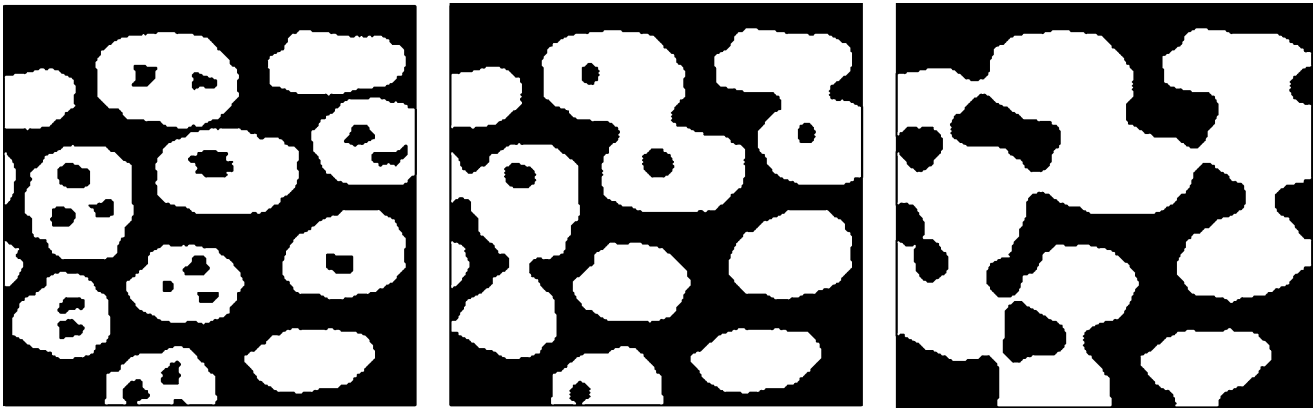


The reductions are obtained by circulation of CO through the pores. However, these overall relations do not tell us how the process occurs (Jeulin 2000). Is the whole hematite pellet transformed into a whole pellet of ferrite, and then into iron?

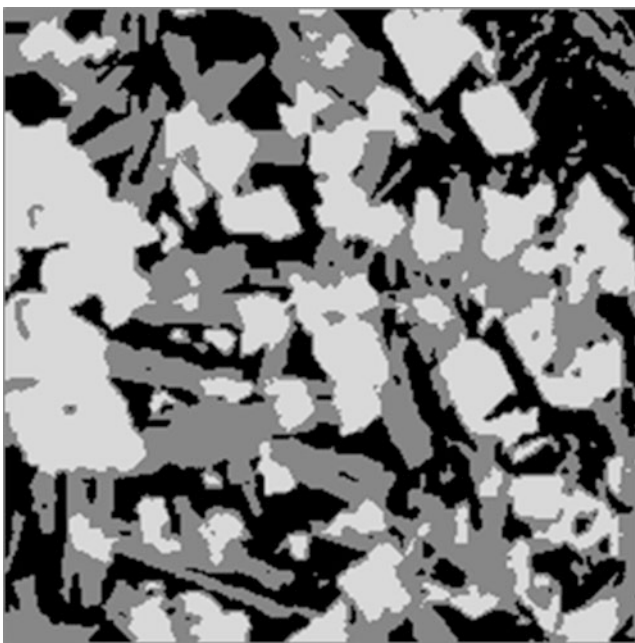
In Fig. 4, the reduction process is caught red-handed. It is a polished section taken from the middle of the blast furnace. The three phases are imbricated. Is the ferrite phase relatively closed to hematite, or not?

A convenient tool for answering the question, here, is the rectangular covariance $C_{ij}(h)$. Take a structuring element made of two points from h apart in the horizontal direction, $C_{ij}(h)$ is the proportion of cases when the left point falls in phase i and the right one in phase j (Fig. 5).

The slope of covariance near the origin is proportional to the contact surface per unit volume. The two ferrite covariances are similar. Unlike, that between hematite and pores starts from more below, which means smaller contact surface. Moreover it presents a hole effect for $h = 50$ this indicates a halo of ferrite around hematite.



Mathematical Morphology, Fig. 3 Initial set, and closings φ_B (center) and $\varphi_{B'}$ (right) with B and B' discs, and $B \subseteq B'$



Mathematical Morphology, Fig. 4 Light gray: hematite; dark gray: ferrite; black: pores

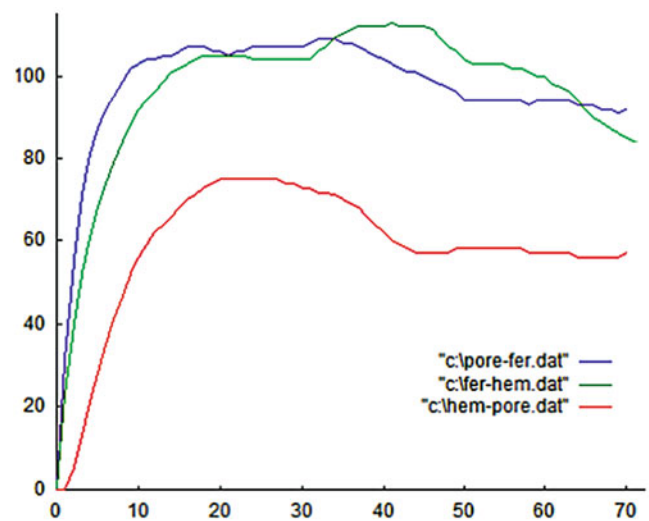
Dilations and Erosions, Adjunctions

The generalization of the Minkowski operations, and of their links, is straightforward:

Definition. Dilation, Erosion, and Djunction Let $\mathcal{L}$ and $\mathcal{M}$ be two complete lattices (identical or distinct).

1. An operator $\delta : \mathcal{L} \rightarrow \mathcal{M}$ is a dilation when it preserves the supremum:

$$\forall \{x_i | i \in I\} \subseteq \mathcal{L}, \quad \varepsilon \left(\bigwedge_{i \in I} x_i \right) = \bigwedge_{i \in I} \varepsilon(x_i);$$



Mathematical Morphology, Fig. 5 The three rectangle covariances of the sinter specimen

in particular, for I empty, $\varepsilon(\mathbf{1}) = \mathbf{1}$.

3. Two operators $\delta : \mathcal{L} \rightarrow \mathcal{M}$ and $\varepsilon : \mathcal{M} \rightarrow \mathcal{L}$ form an adjunction (ε, δ) when:

$$\forall x \in \mathcal{L}, \forall y \in \mathcal{M}, \quad \delta(x) \leq y \Leftrightarrow x \leq \varepsilon(y). \quad (8)$$

Dilation and erosion are dual notions. It is easy to see that both are *increasing operators*. Conversely, every increasing operator preserving $\mathbf{0}$ (respectively, preserving $\mathbf{1}$) is an *infimum* (respectively, a *supremum*) of dilations (respectively, erosions) (Serra 1988).

Theorem Adjunctions. Let $\mathcal{L}$ and $\mathcal{M}$ be two complete lattices. Adjunctions constitute a bijection between dilations $\mathcal{L} \rightarrow \mathcal{M}$ and erosions $\mathcal{M} \rightarrow \mathcal{L}$, that is:

1. Given two operators $\delta : \mathcal{L} \rightarrow \mathcal{M}$ and $\varepsilon : \mathcal{M} \rightarrow \mathcal{L}$ forming an adjunction (ε, δ) , δ is a dilation and ε is an erosion.
2. For every dilation $\delta : \mathcal{L} \rightarrow \mathcal{M}$, (resp. erosion ε) there exists a unique erosion $\varepsilon : \mathcal{M} \rightarrow \mathcal{L}$ (resp. dilation δ) such that (ε, δ) is an adjunction.

Composing the erosion and the dilation by a structuring element, one obtains according to the order of composition the opening γ and the closing φ by that structuring element. That remains true in the general case.

Theorem Opening and closing by adjunction Let $\mathcal{L}$ and $\mathcal{M}$ be two complete lattices, and let a dilation $\delta : \mathcal{L} \rightarrow \mathcal{M}$ and an erosion $\varepsilon : \mathcal{M} \rightarrow \mathcal{L}$ form an adjunction (ε, δ) . Then:

1. $\varepsilon = \varepsilon\delta\varepsilon$ and $\delta = \delta\varepsilon\delta$.
2. $\delta\varepsilon$ is an opening on $\mathcal{M}$ whose invariant elements are the dilates of the initial elements $\mathcal{B}_{\delta\varepsilon} = \delta(\mathcal{L})$.
3. $\varepsilon\delta$ is a closing on $\mathcal{L}$ and $\mathcal{B}_{\varepsilon\delta} = \varepsilon(\mathcal{M})$.

The two lattices $\mathcal{L}$ and $\mathcal{M}$ often are identical, and $\delta, \varepsilon, \delta\varepsilon$, and $\varepsilon\delta$ are just operators on $\mathcal{L}$.

General Set-Theoretical Case

Let us illustrate these notions in the case of lattices of the form $\mathcal{P}(E)$. According to the definition of dilation and erosion, an operator is a dilation if it preserves the union, and an erosion if it preserves the intersection. One can characterize a dilation by its behavior on the points identified to singletons; for every point p , let us write $\delta(p)$ for $\delta(\{p\})$. Then:

Theorem Let E_1 and E_2 be two spaces. A map $\delta : \mathcal{P}(E_1) \rightarrow \mathcal{P}(E_2)$ is a dilation if and only if:

$$\forall X \in \mathcal{P}(E_1), \quad \delta(X) = \bigcup_{x \in X} \delta(x). \quad (9)$$

The adjoint erosion $\varepsilon : \mathcal{P}(E_2) \rightarrow \mathcal{P}(E_1)$ is then given by:

$$\forall Y \in \mathcal{P}(E_2), \quad \varepsilon(Y) = \{x \in E_1 \mid \delta(x) \subseteq Y\}. \quad (10)$$

In the case where $E_1 = E_2 = E = \mathbb{Z}^n$ or $\mathbb{R}^n$, one easily sees by equation (9) that every dilation invariant under translation is of the form $\delta_B : X \mapsto X \oplus B$, with the structuring element $B = \delta(o)$, the dilation of the origin. An erosion is invariant under translation if and only if the adjointed dilation is invariant under translation, hence it will be of the form $\varepsilon_B : X \mapsto X \ominus B$. The dilation δ_B is increasing in B , but ε_B is decreasing in B . Moreover, one has:

$$\bigvee_{i \in I} \delta_{B_i} = \delta_{\bigcup_{i \in I} B_i} \quad \text{and} \quad \bigwedge_{i \in I} \varepsilon_{B_i} = \varepsilon_{\bigcup_{i \in I} B_i}.$$

The rules of compositions of dilations and erosions derive from these relations, which imply:

$$\delta_{B_1} \delta_{B_2} = \delta_{B_2} \delta_{B_1} = \delta_{B_1 \oplus B_2}$$

$$\varepsilon_{B_1} \varepsilon_{B_2} = \varepsilon_{B_2} \varepsilon_{B_1} = \varepsilon_{B_1 \oplus B_2}$$

The opening $\gamma_B = \delta_B \varepsilon_B$ admits for invariant sets the dilated by B , that is, the $A \oplus B$ for $A \in \mathcal{P}(E)$. Similarly, the invariance domain of the closing $\varphi_B = \varepsilon_B \delta_B$ is formed of sets eroded by B , that is, the $A \ominus B$ pour $A \in \mathcal{P}(E)$.

We conclude this section by a more general result: every opening invariant under translation, in $\mathbb{R}^n$ or $\mathbb{Z}^n$, can be written as a supremum of openings of the type γ_B (Matheron 1975), and more generally, every opening in a complete lattice expresses itself under the form of a supremum of openings by adjunction (Heijmans and Ronse 1990).

Case of the Unit Circle

Morphological operations on the unit circle appear in color images processing, where the hue spans the unit circle. We will not develop morphology for color images here, and refer to Angulo (2007). Another field of applications is the processing of directions, more useful in geosciences. We give below a few indications in the 2-D case of the unit disc C . The approach extends to the unit sphere, but the formalism is less simple.

The unit disc C , like the round table of King Arthur knights, has no order of importance, and no dominant position. This signifies we cannot construct a lattice on it, unless the phenomenon under study imposes its origin. However, there are three paths to bypass this interdiction, by focusing on *increments*, a wide class which covers residues, gradients, medians, etc. (Hanbury and Serra 2001).

The points a_i on the unit disc C of center O are indicated by their angle, between 0 and 2π from an arbitrary origin a_0 . Given two points a and a' , we use the notation $a \div a'$ to indicate the value of the acute angle $a\widehat{O}a'$, that is,

$$\begin{aligned} a \div a' &= |a - a'| & \text{if } |a - a'| \leq \pi \\ a \div a' &= 2\pi - |a - a'| & \text{if } |a - a'| \geq \pi \end{aligned}$$

Now consider the opening by adjunction $\gamma_B(f)$ of function f by B , and indicate by $\{B_i, i \in I\}$ the family of structuring elements which contain point x . The residue at point x can be written

$$f(x) - \gamma_B(f)(x) = -\sup_{i \in I} \left\{ \inf_{y \in B_i} [f(y) - f(x)] \right\},$$

in which there are only increments of the function f around point x . This general relation can be transposed to circular values a and gives

$$f(x) - \gamma_B(f)(x) = -\sup_{i \in I} \left\{ \inf_{y \in B_i} [a(x) \div a(y)] \right\}.$$

Gradients and medians on the unit disc are obtained in the same manner.

An Example

When sorting oak boards destined to make furniture, it is necessary to find the knots, which appear as fast changes of direction, and to measure the sizes. The average direction of the veins is calculated over small squares, and the residues of circular openings on these directions indicate the knots. Figure 6 depicts the residues found onto the original image, for the four sizes 3, 5, 7, and 9 of the structuring element.

Case of Numerical Functions

Let $T = \bar{R}, \bar{Z}$ or a closed interval of $\bar{R}$ or $\bar{Z}$ and $E = R^n$ or $E = Z^n$. A numerical function $E \rightarrow T$ is equivalent to its *subgraph*, or *umbra* in the product space $E \times T$ (Sternberg, 1986), that is to the set of all points (x, t) which are below $(x, F(x))$, and dilating the umbra of function F by that of function G still results into an umbra, which in turn defines

the dilated function $\delta_G(F)$. In the translation invariant case the formalism is as follows:

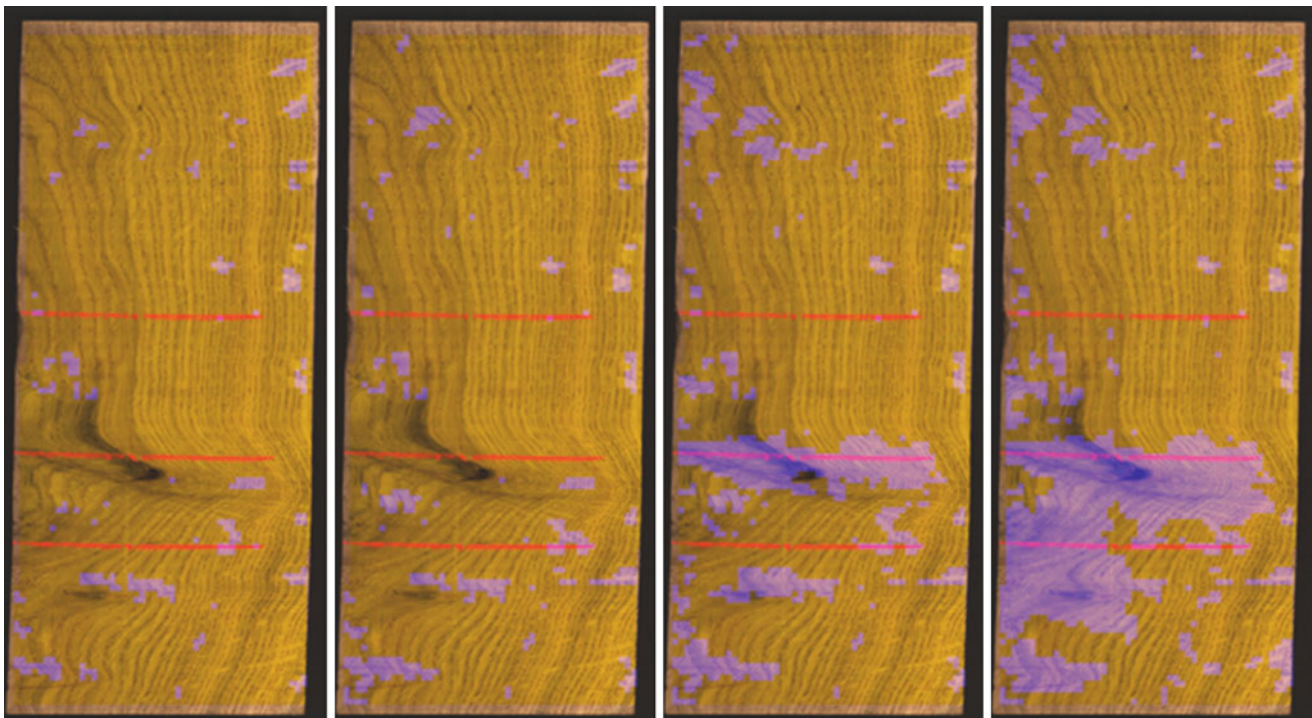
$$\begin{aligned} \delta(F \oplus G)(x) &= \bigvee_{y \in E} [F(x - y) + G(y)], \\ \varepsilon(F \ominus G)(x) &= \bigwedge_{y \in E} [F(x + y) - G(y)] \end{aligned} \quad (11)$$

An Example

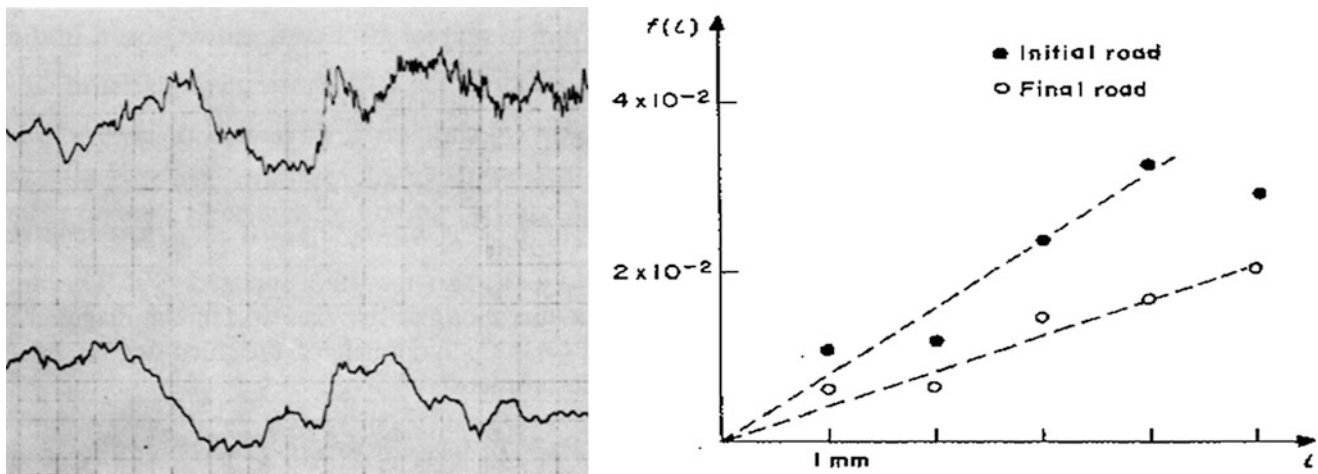
This example concerns the wear of road surface. Under the traffic, the bitumen of a road loses its initial roughness quality, and one wishes to measure this phenomenon (Serra 1984). A ring of 20 m emulates the road, and a massive set of four wheels emulates the truck. Two experiments, of 1.000 and then 10.000 passages of 6.5 t trucks going at 65 km/h, are made. The relief of the road is measured afterwards by a sensor which produces the two plots of Fig. 7 (left).

The set under the plot is eroded by a segment of length l and direction α , and the area reduction is averaged in α . This result in the function $P(l)$ whose second derivative $f(l)$ is the histogram of the intercepts. It admits the following limited expansion near the origin:

$$f(l) = \frac{a}{E(C)} l$$



Mathematical Morphology, Fig. 6 Detection of knots in an oak board as changes in direction



Mathematical Morphology, Fig. 7 Left: changes in the roughness of a road surface; right: beginnings of the histograms of the road intercepts

where a is a constant and C stands for the average of the average curvature, assuming that the road relief admits curvatures. By measuring $f(h)$ in six directions and applying the above formula, we find: $E(C) = 2210^{-2}\text{mm}^{-2}$ before wear and $E(C2) = 10^{-2}\text{mm}^{-2}$ after wear (Fig. 7 (right)). In spite of the curvatures assumption, of the poor number of directions, and the passage to a second derivation, the model still provides a pertinent information. Remark the stereological meaning of the result: by means of 1-D structuring elements one reaches, by rotation averaging, an estimate of the 3-D average curvature.

Flat Structuring Elements

The most popular numerical dilations and erosions are those obtained by level set approach, where each thresholded set is treated individually. This amounts to use “flat” structuring functions, that is, to take for G the half cylinder:

$$x \in B \Rightarrow G_B(x) = 0 \quad x \notin B \Rightarrow G_B(x) = -\infty$$

The relations (Eq. 11) of numerical dilation and erosion take the simpler form:

$$\begin{aligned} \delta(F \oplus G)(x) &= \bigvee_{y \in E} [F(x-y), y \in B], \\ \varepsilon(F \ominus G)(x) &= \bigwedge_{y \in E} [F(x-y), y \in -B] \end{aligned}$$

The numerical geodesic dilation is obtained by replacing the binary dilation by its numerical version in the binary

geodesic formalism. For B large enough, the dilation becomes idempotent. It is called then the reconstruction opening of F inside the mask function M (Fig. 8 right).

The operations based on flat structuring elements commute under anamorphosis of the gray levels, and are the only ones to do it. This remarkable property implies three very useful consequences:

1. It makes the processing independent of the creation of the image. For example, the angiograph below may, or may not, have been obtained by a camera equipped with the so called “ γ correction” which takes the logarithm of the signal. A flat opening γ ensures us that

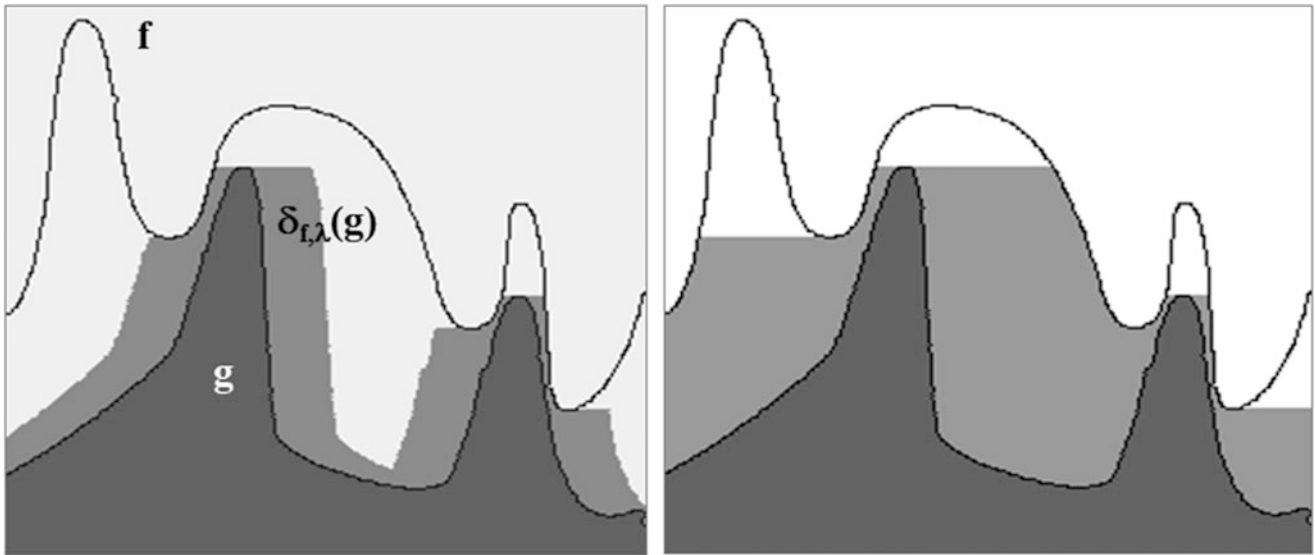
$$\log[\gamma(F)] = \gamma[\log(F)].$$

2. An n-bits image is transformed into an n-bits image, and in particular the indicator functions are preserved.
3. Reducing the dynamic of function F does not damage the processing. For example, in demographic data, F often range from 1 to 10^5 , but then $\sqrt{F}$ ranges from 1 to 512 and can be considered as a usual image of 9 bits.

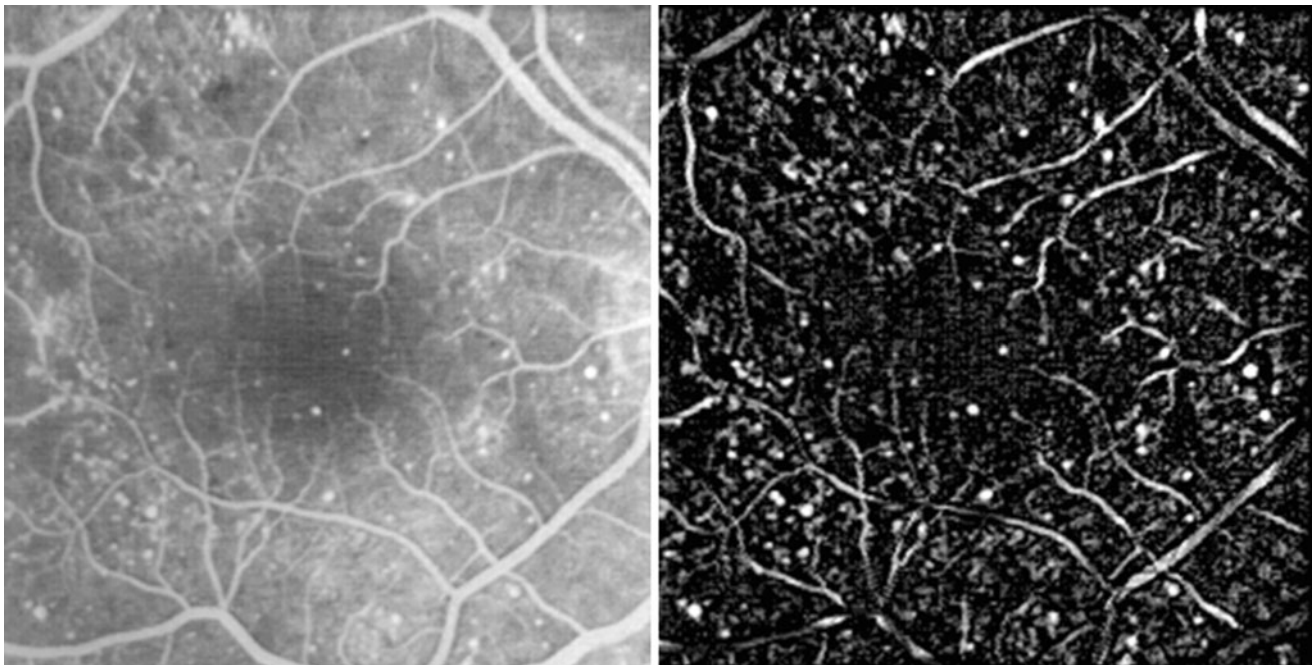
An Example

In ophthalmology, the presence of aneurysms on retina is the sign of diabetes. How to extract them from an angiograph?

The left part of Fig. 9 depicts the angiograph F of the retina of a diabetic patient. The aneurysms consist of small white spots, but they are blurred by several vessels of various thicknesses and by the heterogeneity of the background.



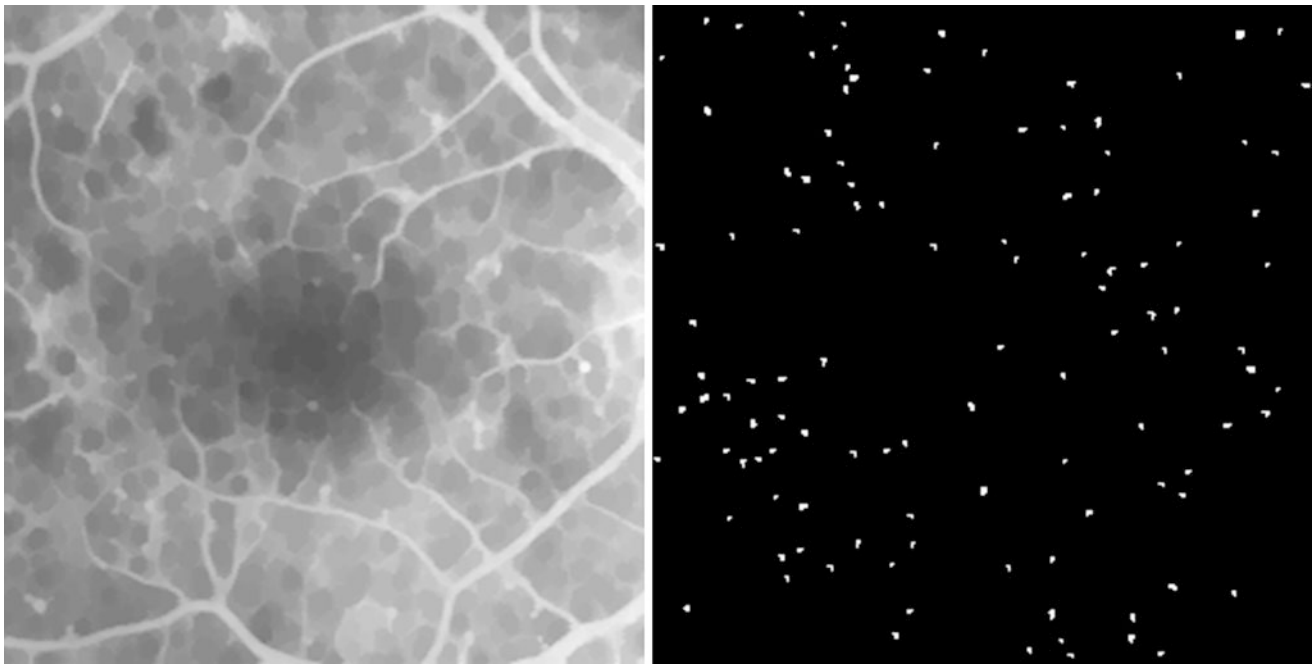
Mathematical Morphology, Fig. 8 Left: Geodesic flat dilation; right: and reconstruction opening



Mathematical Morphology, Fig. 9 Left: Retina image; right: residue of a flat opening by segments

The first idea here is to perform, in all directions, openings by linear structuring elements, small but longer than the diameters of the aneurysms, and to take their union $\gamma(F)$ over all directions. In practice the directional average is performed in

the three directions of a hexagonal grid. The segments are included inside the long narrow tubes that are the vessels, which therefore belong to the opening and should disappear in the residue $F - \gamma(F)$. One can see this residue in Fig. 9



Mathematical Morphology, Fig. 10 Left: reconstruction opening of the angiograph; right: corresponding residue

(right). The former gray background is now uniformly black, which is nice, but some regions of the vessels were not caught, and are mixed with the aneurysms in the residue.

This partial result suggests to replace the adjunction opening by a (still flat) reconstruction opening which invades the vessels.

Figure 10 (left) depicts the reconstruction opening $\gamma_{\text{rec}}(F)$. All vessels have been reconstructed, and indeed the residue $F - \gamma_{\text{rec}}(F)$, visible in Fig. 10 (right), exclusively shows the aneurysms.

Summary

The above theory has suggested several developments, in various directions, that were not presented here. In geosciences, the potential of description of mathematical morphology is often used for generating synthetic images of reliefs, or of rivers networks (Sagar 2014). In physics, mathematical morphology leads to stochastic models for sets and functions, which serve in the relationships between textures and physical properties of solid materials (Jeulin 2000). In mathematics, two other branches are the concern of differential geometry (Maragos 1996) and of the digital version of the method (Kiselman 2020; Najman et al. 2005).

Cross-References

- [Matheron, Georges](#)
- [Morphological Filtering](#)

Bibliography

- Angulo J (2007) Morphological colour operators in totally ordered lattices based on distances. Application to image filtering, enhancement and analysis. *Comput Vis Image Underst* 107(2–3):56–73
- Hanbury A, Serra J (2001) Morphological operators on the unit circle. *IEEE Trans Image Process* 10(12):1842–1850
- Heijmans M (1994) *Morphological image operators*. Academic, Boston
- Heijmans H, Ronse C (1990) The algebraic basis of mathematical morphology: I. Dilations and erosions. *Comput Vis Graph Image Process* 50:245–295
- Heijmans H, Serra J (1992) Convergence, continuity and iteration in mathematical morphology. *J Vis Commun Image Represent* 3:84–102
- Jeulin D (2000) Random texture models for materials structures. *Stat Comput* 10:121–131
- Kiselman O (2020) *Elements of digital geometry, mathematical morphology, and discrete optimisation*. Cambridge University Press, Cambridge
- Lantuéjoul C, Beucher S (1981) On the use of the geodesic metric in image analysis. *J Microsc* 121:39–49
- Maragos P (1996) Differential morphology and image processing. *IEEE Trans Image Process* 5(8):922–937

- Matheron G (1967) *Eléments pour une théorie des milieux poreux*. Masson, Paris
- Matheron G (1975) *Random sets and integral geometry*. Wiley, New York
- Matheron G (1996) *Treillis compacts et treillis coprimaires*. Tech. Rep. N-5/96/G, Ecole des Mines de Paris, Centre de Géostatistique
- Matheron G, Serra J (2002) The birth of mathematical morphology. In: Talbot H, Beare R (eds) *Proceedings of VIth international symposium on mathematical morphology*. Commonwealth Scientific and Industrial Research Organisation, Sydney, pp 1–16
- Najman L, Talbot H (2010) *Mathematical morphology: from theory to applications*. Wiley, Hoboken
- Najman L, Couprie M, Bertrand G (2005) Watersheds, mosaics and the emergence paradigm. *Discrete Appl Math* 147(2–3):301–324, special issue on DGCI
- Ronse C, Serra J (2010) Algebraic foundations of morphology. In: Najman L, Talbot H (eds) *Mathematical morphology*. Wiley, Hoboken, pp 35–80
- Sagar BSD (2014) Cartograms via mathematical morphology. *Inf Vis* 13(1):42–58
- Serra J (1982) *Image analysis and mathematical morphology*. Academic, London
- Serra J (1984) Descriptors of flatness and roughness. *J Microsc* 134(3):227–243
- Serra J (1992) Equicontinuous functions, a model for mathematical morphology. In: *Non-linear algebra and morphological image processing, proceedings*, vol 1769. SPIE, San Diego, pp 252–263
- Sternberg S (1986) Grayscale morphology. *Comput Graph Image Process* 35:333–355

Matheron, Georges

Jean Serra
Centre de Morphologie Mathématique, Ecole des Mines,
Paristech, Paris, France

Biography

G. Matheron, who was born in Paris in December 1930, entered the prestigious Ecole Polytechnique in 1949 and the Ecole des Mines de Paris two years later. In 1952, he landed in Algeria with wife and child for working at the Algerian Mining Survey. He took over its scientific management in 1956, then its general management (1958). Starting from the works of Krige, from South Africa, and de Wijs, from the Netherlands, he created a theory for estimating mining resources he named Geostatistics. He designed the basic mathematical notion of a variogram, which is both the key for estimation problems and excellent tool for the describing spatial variability within mineral deposits.

He invented random functions with stationary increments, while being skeptical about the underlying

probabilistic framework. Every deposit is a unique phenomenon. Studied in itself, it does not enter the scope of probabilities. It was within that semi-random framework that G. Matheron wrote his two volumes “*Trait de Gostatistique Applique*” in 1962, which was based upon the Algerian experience. He defended his PhD thesis in 1963, in which the deterministic and random parts of his approach were developed successively. But G. Matheron waited 20 years before formulating, in *Estimating and Choosing* (Matheron 1989), which choice to make of either approach according to the context.

In 1968, the Ecole des Mines de Paris gave G. Matheron the opportunity to create the Centre de Morphologie Mathématique, in Fontainebleau. From that time onward, geostatistics gained international recognition, and requests for mining estimations arrived from the five continents. In 1971 he wrote the notes for a course on the theory of regionalized variables (Matheron 1971) which subsequently became the “bible” for ore reserve estimation in mining industry, although it was not published as a book until 2019 (Pawlowsky-Glahn and Serra 2019)! He was also asked to evaluate oil reservoirs, to map atmospheric pressures, submarine hydrography, etc. G. Matheron responded to these requests by inventing, in 1969, the theory of universal kriging which does not require the stationarity constraint, and also inventing nonlinear geostatistics in 1973 for predicting statistical distributions of ore grades in mining panels (Fig. 1).

The mining problem of the suitability of grade for grinding led him and Jean Serra to extend the concept of variogram to that of the hit-or-miss transformation, then to that of set opening. They called this new branch of science mathematical morphology, which became progressively independent of geostatistics. G. Matheron extracted from this newly developed morphological applications some general approaches, which could be conceptualized, such as granulometry, increasing operators, Poisson hyper-planes, and Boolean sets, and gathered them into the book entitled “*Random Sets and Integral Geometry*,” in 1975 (Matheron 1975).

In the 1980s, G. Matheron based Mathematical Morphology on the broader level of the complete lattices, thus providing a common approach to sets, functions, and partitions, and constructed the theory of the increasing and idempotent operators that he called morphological filtering (Matheron 1988). He retired in 1995 and passed away five years later.

One can find the 11,000 pages of scientific notes written by G. Matheron in the index by name of [HTTP://eg.ensmp.fr/bibliotheque](http://eg.ensmp.fr/bibliotheque)



1958 Alger



1985 Fontainebleau

Matheron, Georges, Fig. 1 Two photographs of Georges Matheron (courtesy of Mrs Françoise Matheron)

Bibliography

- Matheron G (1971) The theory of regionalised variables and its applications. Ecole des Mines de Paris, Paris. 211p
- Matheron G (1975) Random sets and integral geometry. Wiley, New York
- Matheron G (1989) Estimating and choosing – an essay on probability in practice. Springer, Berlin/New York
- Pawlowsky-Glahn V, Serra J (eds) (2019) Matheron's theory of regionalized variables. Oxford University Press, Oxford

Matrix Algebra

Deeksha Aggarwal and Uttam Kumar
Spatial Computing Laboratory, Center for Data Sciences,
International Institute of Information Technology Bangalore
(IIITB), Bangalore, Karnataka, India

Definition

Matrices are one of the most powerful tools in mathematics and geosciences with applications in Earth system science. The evolution of the concept of matrices is the result of an attempt to obtain compact and simple methods of solving system of linear equations. A matrix is an ordered rectangular array of numbers or functions. The numbers or functions are called the elements or the entries of the matrix. We denote matrices by capital letters. The following are some examples of matrices:

$$A = \begin{bmatrix} 2 & 12 \\ 4 & 5 \\ 3 & 1 \end{bmatrix}, B = \begin{bmatrix} 1 & 2 \\ 14 & 15 \end{bmatrix}, C = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 10 \end{bmatrix}$$

In the above examples, the horizontal lines of elements are said to constitute rows of the matrix and the vertical lines of elements are said to constitute columns of the matrix. Thus A has 3 rows and 2 columns, B has 2 rows and 2 columns while C has 2 rows and 3 columns.

Matrix algebra as a subset of linear algebra focuses primarily on the basic concepts and solution techniques. Numerous problems that are addressed in engineering sciences and mathematics are solved by properly setting up a system of linear equations.

Introduction

Order of a Matrix

A matrix containing n rows and m columns is called a matrix having an order of $n \times m$ (read as n by m). The total number of elements in a $n \times m$ matrix is nm . In general, a matrix containing n rows and m columns is represented as

$$A_{n,m} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,m} \end{bmatrix}_{n \times m}$$

Types of Matrices

Below are the different types of matrices:

Column Matrix: When a matrix has only one column, it is said to be a column matrix. For example:

$$A_{n,1} = \begin{bmatrix} a_{1,1} \\ a_{2,1} \\ \vdots \\ a_{n,1} \end{bmatrix}_{n \times 1}$$

Row Matrix: When a matrix has only one row, it is said to be a row matrix. For example:

$$A_{1,m} = [a_{1,1} \ a_{1,2} \ \dots \ a_{1,m}]_{1 \times m}$$

Square Matrix: A matrix having number of rows equal to the number of columns, is said to be a square matrix ($n = m$). The order of such a matrix is n . For example matrix A (shown below) is a square matrix of order $n \times n$:

$$A_{n,n} = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{bmatrix}_{n \times n}$$

Diagonal Matrix: A diagonal matrix is a square matrix having all its non-diagonal elements as zero. For example:

$$A_{n,n} = \begin{bmatrix} a_{1,1} & 0 & \dots & 0 \\ 0 & a_{2,2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{n,n} \end{bmatrix}_{n \times n}$$

Scalar Matrix: A diagonal matrix is said to be a scalar matrix when all its diagonal elements are equal. For example:

$$A_{n,n} = \begin{bmatrix} a & 0 & \dots & 0 \\ 0 & a & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a \end{bmatrix}_{n \times n}$$

Identity Matrix: An identity matrix is a square and diagonal matrix having all its diagonal elements as one. It is generally denoted by I_n . Scalar matrix is an identity matrix when $a = 1$. For example:

$$I_n = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}_{n \times n}$$

Power of a Matrix: For a matrix A of order $n \times n$, we write A^k for the product of k (scalar) copies of A times the identity as given below:

$$A^k = A \dots A I \quad (1)$$

This includes the case when $k = 0$ and $A^0 = I$.

Rank of a Matrix: The rank of a matrix A , denoted by $\text{rank} A$, is the dimension of column space of A .

Null/Zero Matrix: A null or zero matrix is a matrix having all its elements as zero. It is generally denoted by O . Scalar matrix is a null matrix when $a = 0$. For example:

$$O = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}_{n \times n}$$

Matrix Addition and Subtraction

Consider two matrices A and B of the same order $n \times m$ denoted as:

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,m} \\ a_{2,1} & a_{2,2} & \dots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,m} \end{bmatrix}_{n \times m}$$

and

$$B = \begin{bmatrix} b_{1,1} & b_{1,2} & \dots & b_{1,m} \\ b_{2,1} & b_{2,2} & \dots & b_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n,1} & b_{n,2} & \dots & b_{n,m} \end{bmatrix}_{n \times m}$$

The sum of $A + B$ is obtained by adding the corresponding elements of the given matrices. Moreover, the two matrices have to be of the same order and the resultant matrix C will also be of the same order $n \times m$ as shown below:

$$C = A + B$$

$$C = \begin{bmatrix} a_{1,1} + b_{1,1} & a_{1,2} + b_{1,2} & \cdots & a_{1,m} + b_{1,m} \\ a_{2,1} + b_{2,1} & a_{2,2} + b_{2,2} & \cdots & a_{2,m} + b_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} + b_{n,1} & a_{n,2} + b_{n,2} & \cdots & a_{n,m} + b_{n,m} \end{bmatrix}_{n \times m}$$

The difference of $A - B$ is obtained by subtracting the corresponding elements of the given matrices, the two matrices have to be of the same order and the resultant matrix C will also be of the same order $n \times m$ as shown below:

$$C = A - B$$

$$C = \begin{bmatrix} a_{1,1} - b_{1,1} & a_{1,2} - b_{1,2} & \cdots & a_{1,m} - b_{1,m} \\ a_{2,1} - b_{2,1} & a_{2,2} - b_{2,2} & \cdots & a_{2,m} - b_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} - b_{n,1} & a_{n,2} - b_{n,2} & \cdots & a_{n,m} - b_{n,m} \end{bmatrix}_{n \times m}$$

The addition of matrices satisfy the following properties:

- (i) **Commutative Law:** If two matrices A and B are of same order, then $A + B = B + A$ in case of addition of two matrices.
- (ii) **Associative Law:** If three matrices A , B , and C are of same order, then $(A + B) + C = A + (B + C)$.
- (iii) **Additive Identity:** If two matrices A and O are of same order among which O is a null matrix, then $A + O = O + A = A$ in case of addition of two matrices.

Matrix Multiplication

The multiplication or product of two matrices A of order $n \times m$ and B of order $s \times q$ can be obtained only if the number of columns in A are equal to the number of rows in B , i.e., when $m = s$. The resultant matrix C obtained by multiplying A and B will be of the order $n \times q$. The elements in C will be obtained by element wise multiplication of the rows of A and columns of B and taking sum of all these products as shown below: Let $A = [a_{i,j}]_{n \times m}$ and $B = [b_{j,k}]_{m \times q}$ ($m = s$) where $i \in [0, n]$, $j \in [0, m]$, $k \in [0, q]$, then the i_{th} row of A is $[a_{i,1}, a_{i,2}, \dots, a_{i,m}]$ and the k^{th} column of B is

$$\begin{bmatrix} b_{1,k} \\ b_{2,k} \\ \vdots \\ b_{m,k} \end{bmatrix}, \text{ then the element } c_{i,k} \text{ in the matrix } C \text{ will be:}$$

$$c_{i,k} = a_{i,1}b_{1,k} + a_{i,2}b_{2,k} + \cdots + a_{i,m}b_{m,k}$$

$$c_{i,k} = \sum_{j=1}^n a_{i,j}b_{j,k}$$

$$C = [c_{i,k}]_{n \times q}$$

Let A be $m \times n$ and let B and C have sizes for which the indicated sums and products are defined. The matrix multiplication satisfy the following properties:

- (i) $A(BC) = (AB)C + kB$
- (ii) $A(B + C) = AB + AC$
- (iii) $(B + C)A = BA + CA$
- (iv) $k(A + B) = kA + kB$
- (v) $k(AB) = (kA)B = A(kB)$, for any scalar k .

The proof of the properties are beyond the scope of this chapter.

Warnings:

1. In general, $AB \neq BA$
2. If $AB = AC$, then it is not true in general that $B = C$.
3. If a product AB is a null matrix, one cannot conclude in general that either $A = 0$ or $B = 0$.

Matrix Scalar Multiplication

Multiplication of a matrix by a scalar k is obtained by multiplying each element of the matrix with k as shown below:

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,m} \end{bmatrix}_{n \times m}$$

then

$$kA = \begin{bmatrix} ka_{1,1} & ka_{1,2} & \cdots & ka_{1,m} \\ ka_{2,1} & ka_{2,2} & \cdots & ka_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ ka_{n,1} & ka_{n,2} & \cdots & ka_{n,m} \end{bmatrix}_{n \times m}$$

The scalar multiplication of a matrix satisfy the following properties:

- (i) $k(A + B) = kA + kB$
- (ii) $kA = Ak$
- (iii) $OA = 0$, where O is a null matrix.
- (iv) $(k + l)A = kA + lA$, where k and l are scalars.
- (v) $k(A + B) = kA + kB$
- (vi) $k(sA) = (ks)A$

Operations on Matrices

This section discusses some of the operations that can be performed on a matrix to obtain another matrix as explained below:

i. **Transpose of a Matrix:** For a matrix A of order $n \times m$, transpose of A is obtained by interchanging rows of A with columns of A . It is denoted by A^T and will have the order as $m \times n$. For example: if

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,m} \end{bmatrix}_{n \times m}$$

then

$$A^T = \begin{bmatrix} a_{1,1} & a_{2,1} & \cdots & a_{n,1} \\ a_{1,2} & a_{2,2} & \cdots & a_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1,m} & a_{2,m} & \cdots & a_{n,m} \end{bmatrix}_{m \times n}$$

The transpose of a matrix satisfy the following properties:

- (i) $(A + B)^T = A^T + B^T$ and $(A - B)^T = A^T - B^T$
- (ii) $(kA)^T = kA^T$, where k is a scalar.
- (iii) $(AB)^T = B^T A^T$, where A and B are matrices.
- (iv) $(A^{-1})^T = (A^T)^{-1}$
- (v) $(A^T)^T = A$

Note: The transpose of a product of matrices equals the product of their transposes in the reverse order.

ii. **Trace of a Matrix:** For a square matrix A of order $n \times n$, trace of A is the sum obtained by adding its diagonal elements. It is denoted by $tr(A)$. Trace of a matrix exist only for square matrices.

$$tr(A) = a_{1,1} + a_{2,2} + \cdots + a_{n,n}$$

The trace of a matrix satisfy the following properties:

- (i) $tr(A + B) = tr(A) + tr(B)$ and $tr(A - B) = tr(A) - tr(B)$
- (ii) $tr(kA) = ktr(A)$, where k is a scalar.
- (iii) $tr(AB) = tr(BA)$
- (iv) $tr(A^T) = tr(A)$

iii. Determinant of a Matrix:

We have studied about matrices and arithmetic of matrices in the previous sections. In this section we learn that a system of algebraic equations can be expressed in the form of matrices as shown below:

$$\begin{aligned} ax + by &= c_1 \\ cx + dy &= c_2 \end{aligned}$$

can be expressed as:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$$

To know whether the system of linear equations has a unique solution or not, we compute a number which defines the uniqueness of a solution and is known as the determinant (det) formulated as $det(A)$ for a matrix A as:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$det(D) = ad - bc$$

The determinant of a matrix satisfy the following properties:

- (i) $det(A)$ remains unchanged even if its rows and columns are interchanged, i.e., $det(A) = det(A^T)$.
- (ii) The sign of $det(A)$ changes if any two rows or columns are interchanged. Interchanging two rows or two columns are denoted by $R_i \leftrightarrow R_j$ and $C_i \leftrightarrow C_j$ respectively. Hence, $det(A) = -det(A)$ if $R_i \leftrightarrow R_j$ or $C_i \leftrightarrow C_j$.
- (iii) The value of $det(A)$ is zero if any of the two rows or columns are identical.
- (iv) For a matrix A , if any row or column gets multiplied by a scalar value k then $det(A) = kdet(A)$.
- (v) If a matrix A can be expressed as a sum of two matrices (say B and C) then $det(A) = det(B) + det(C)$.

iv. Inverse of a Matrix

For a square matrix A of order n , A is said to be invertible if there exist another square matrix B of the same order n , such that below equation satisfies:

$$AB = BA = I$$

Then, B is called the inverse of A and is denoted by A^{-1} (read as A inverse). Moreover, if B is the inverse of A , then A is also the inverse of B . A matrix that is not invertible is sometimes called a singular matrix, and an invertible matrix is called a nonsingular matrix.

The inverse of a matrix follow the below properties:

- (i) Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, if $ad - bc \neq 0$, then A is invertible and $A^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$. If $ad - bc = 0$, then A is not invertible. That is, a square matrix is only invertible if and only if its $det(A) \neq 0$.
- (ii) If matrix A of size $n \times n$ is invertible, then for each b in $\mathfrak{R}^n$, the equation $Ax = b$ has the unique solution $x = A^{-1}b$.

- (iii) If A is an invertible matrix, then A^{-1} is also invertible and $(A^{-1})^{-1} = A$.
- (iv) If A and B are $n \times n$ invertible matrices, then so is AB , then the inverse of AB is the product of A and B in the reverse order as: $(AB)^{-1} = B^{-1}A^{-1}$.
- (v) If A is an invertible matrix, then so is A^T , then the inverse of A^T is the transpose of A^{-1} as: $(A^T)^{-1} = (A^{-1})^T$.

Note: The product of $n \times n$ invertible matrices is invertible, and the inverse is the product of their inverses in the reverse order.

Eigenvalues and Eigenvectors

In the previous section we discussed about scalar and matrix multiplication with matrix A . In this section, we will discuss the results and outcome of multiplying a vector $\vec{p}$ with the matrix A . To define formally, let A be a matrix of order $n \times n$, $\vec{p}$ be a non-zero column vector of order $n \times 1$ and λ be a scalar. Then $\vec{p}$ is an eigenvector of A and λ is an eigenvalue of A if the below condition is satisfied:

$$A\vec{p} = \lambda\vec{p}$$

Matrix A should be square and $\vec{p}$ should be non-zero. In order to find the eigenvectors or eigenvalues given A , we need to find a non-zero vector $\vec{p}$ and a scalar λ such that $A\vec{p} = \lambda\vec{p}$, which can be further solved as follows:

$$\begin{aligned} A\vec{p} &= \lambda\vec{p} \\ A\vec{p} - \lambda\vec{p} &= \vec{0} \\ (A - \lambda I)\vec{p} &= \vec{0} \end{aligned}$$

The eigenvalues and eigenvectors of a matrix satisfy the following properties for an invertible matrix A of order n :

- (i) For a triangular matrix A , the diagonal values of A are the eigenvalues of A .
- (ii) If λ is an eigenvalue of A with eigenvector $\vec{p}$, then $\frac{1}{\lambda}$ is an eigenvalue of A^{-1} with eigenvector $\vec{p}$.
- (iii) Eigenvalue of A and A^T is same as λ .
- (iv) Sum of eigenvalues of A is equal to the trace of A .
- (v) The product of eigenvalue of A is equal to the determinant of A .

Case Study

In this section, examples on matrix operations and arithmetic are discussed.

Problem 1: If a matrix has 9 elements then what are the possible order of the matrix?

Matrix Algebra, Table 1 Number of chairs and tables made in different factories

Factory	Number of chairs	Number of tables
I	23	12
II	34	15
III	30	18

Solution 1: The order of the matrix is given by the product of the number of rows n and number of columns m . Therefore, we need to find the real values of m and n such that $mn = 9$. Thus, the possible order of the matrix can be: 1×9 , 9×1 , 3×3 .

Problem 2: Consider Table 1 representing the number of tables and chair made in factories I, II and III. Represent this information in terms of a 3×2 matrix. What does the element in the second row and third column represent?

Solution 2: From the Table, the number of rows are three and number of columns are two. Hence, the order of the matrix is 3×2 and the matrix can be written as shown below:

$$A = \begin{bmatrix} 23 & 12 \\ 34 & 15 \\ 30 & 18 \end{bmatrix}_{3 \times 2}$$

The element in the second row and third column represent that 15 tables are made in the factory II.

Problem 3: Let A , B , and C be the three matrices given below:

$$A = \begin{bmatrix} 2 & 12 \\ 4 & 5 \\ 3 & 1 \end{bmatrix}_{3 \times 2} \quad B = \begin{bmatrix} 1 & 2 \\ 14 & 15 \\ 31 & 1 \end{bmatrix}_{3 \times 2} \quad C = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 10 \end{bmatrix}_{2 \times 3}$$

Find: (i) $A + B$ (ii) $A - 2B$ (iii) $3(A + B)$ (iv) AC

Solution 3: (i) Let $D = A + B$

$$\begin{aligned} D &= \begin{bmatrix} 2 & 12 \\ 4 & 5 \\ 3 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 2 \\ 14 & 15 \\ 31 & 1 \end{bmatrix} \\ D &= \begin{bmatrix} 2+1 & 12+2 \\ 4+14 & 5+15 \\ 3+31 & 1+1 \end{bmatrix} = \begin{bmatrix} 3 & 14 \\ 18 & 20 \\ 34 & 2 \end{bmatrix} \end{aligned}$$

(ii) Let $E = A - 2B$

$$E = \begin{bmatrix} 2 & 12 \\ 4 & 5 \\ 3 & 1 \end{bmatrix} - 2 \begin{bmatrix} 1 & 2 \\ 14 & 15 \\ 31 & 1 \end{bmatrix}, E = \begin{bmatrix} 2-2 & 12-4 \\ 4-28 & 5-30 \\ 3-62 & 1-2 \end{bmatrix}$$

$$E = \begin{bmatrix} 0 & 8 \\ -24 & -25 \\ -59 & -1 \end{bmatrix}$$

(iii) From (i) $D = A + B$ hence, let $F = 3D$

$$F = 3 \begin{bmatrix} 3 & 14 \\ 18 & 20 \\ 34 & 2 \end{bmatrix}, F = \begin{bmatrix} 3 \times 3 & 14 \times 3 \\ 18 \times 3 & 20 \times 3 \\ 34 \times 3 & 2 \times 3 \end{bmatrix}$$

$$F = \begin{bmatrix} 9 & 42 \\ 54 & 60 \\ 102 & 6 \end{bmatrix}$$

(iv) Let $G = AC$

$$G = \begin{bmatrix} 2 & 12 \\ 4 & 5 \\ 3 & 1 \end{bmatrix}_{3 \times 2} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 10 \end{bmatrix}_{2 \times 3}$$

$$G = \begin{bmatrix} 2 \times 1 + 12 \times 4 & 2 \times 2 + 12 \times 5 & 2 \times 3 + 12 \times 10 \\ 4 \times 1 + 5 \times 4 & 4 \times 2 + 5 \times 5 & 4 \times 3 + 5 \times 10 \\ 3 \times 1 + 1 \times 4 & 3 \times 2 + 1 \times 5 & 3 \times 3 + 1 \times 10 \end{bmatrix}_{3 \times 3}$$

$$G = \begin{bmatrix} 50 & 64 & 126 \\ 24 & 33 & 62 \\ 7 & 11 & 19 \end{bmatrix}_{3 \times 3}$$

Problem 4: Let A be a matrix as given below:

$$A = \begin{bmatrix} 2 & 12 & 1 \\ 4 & 5 & 1 \\ 3 & 1 & 1 \end{bmatrix}_{3 \times 3}$$

Find: (i) A^T (ii) $\text{tr}(A)$ (iii) $\det(A)$ (iv) A^{-1}

Solution 4:

(i) A^T can be obtained by interchanging rows and columns as shown below:

$$A^T = \begin{bmatrix} 2 & 4 & 3 \\ 12 & 5 & 1 \\ 1 & 1 & 1 \end{bmatrix}_{3 \times 3}$$

(ii) The trace of a matrix can be obtained by adding the diagonal elements as shown below:

$$\text{tr}(A) = 2 + 5 + 1$$

$$\text{tr}(A) = 8$$

(iii) Determinant of matrix A is:

$$\det(A) = 2 \begin{vmatrix} 5 & 1 \\ 1 & 1 \end{vmatrix} - 12 \begin{vmatrix} 4 & 1 \\ 3 & 1 \end{vmatrix} + 1 \begin{vmatrix} 4 & 5 \\ 3 & 1 \end{vmatrix}$$

$$\det(A) = 2((5 \times 1) - (1 \times 1)) - 12((4 \times 1) - (1 \times 3))$$

$$+ 1((4 \times 1) - (5 \times 3)) \det(A) = -15$$

Problem 5: Find rank of the matrix $A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 0 & 5 \end{bmatrix}$

Solution 5: Reduce A into its echelon form:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & -3 & -2 \\ 0 & 0 & 0 \end{bmatrix}$$

The above matrix is in row echelon form. Number of non-zero rows = 2. Hence the rank of matrix $A = 2$.

Problem 6: Find eigenvalues of the matrix $A = \begin{bmatrix} -3 & 15 \\ 3 & 9 \end{bmatrix}$ and find eigenvector for each eigenvalue.

Solution 6:

$$\det(A - \lambda I) = \begin{vmatrix} -3 & 15 \\ 3 & 9 \end{vmatrix} - \lambda \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix}$$

$$\det(A - \lambda I) = \begin{vmatrix} -3 - \lambda & 15 \\ 3 & 9 - \lambda \end{vmatrix}$$

$$\det(A - \lambda I) = ((-3 - \lambda) \times 9) - (15 \times 3)$$

$$= (-3 - \lambda)(9) - 45$$

$$= -27 - 9\lambda - 45$$

$$= -72 - 9\lambda$$

$$\det(A) = -15$$

Conclusion

Algebraic operations with matrices help in analyzing and solving equations in geocomputing. In this chapter we saw some basic operations for handling two or more matrices. Furthermore, the definitions and properties in this chapter also provide basic tools for handling many applications of linear algebra involving two or more matrices.

Cross-References

- [Earth System Science](#)
- [Eigenvalues and Eigenvectors](#)
- [Geocomputing](#)
- [Multivariate Data Analysis in Geosciences, Tools](#)

Bibliography

- Cheney W, Kincaid D (2009) Linear algebra: theory and applications. Aust Math Soc 110:544–550
- Lay DC, McDonald J. Addison-Wesley, 2012, Linear algebra and its applications, 0321388836, 9780321388834.
- Sharma R, Sharma A (2017) Curriculum of Mathematic implemented by NCERT under the plan NCF-2005 at secondary level: an analytical study. Department of Education University of Calcutta, 2277, pp 18–25

Maximum Entropy Method

Dionissios T. Hristopoulos¹ and Emmanouil A. Varouchakis²

¹School of Electrical and Computer Engineering, Technical University of Crete, Chania, Crete, Greece

²School of Mineral Resources Engineering, Technical University of Crete, Chania, Greece

Abbreviations

MEM	maximum entropy method
MaxEnt	maximum entropy
BME	Bayesian Maximum entropy

Definition

The principle of maximum entropy states that the most suitable probability model for a given system maximizes the Shannon entropy subject to the constraints imposed by the data and – if available – other prior knowledge of the system. The maximum entropy distribution is the most general probability distribution function conditionally on the constraints. In the geosciences, the principle of maximum entropy is

mainly used in two ways: (1) in the maximum entropy method (MEM) for the parametric estimation of the power spectrum and (2) for constructing joint probability models suitable for spatial and spatiotemporal datasets.

Overview

The concept of *entropy* was introduced in thermodynamics by the German physicist Rudolf Clausius in the nineteenth century. Clausius used entropy to measure the thermal energy of a machine per unit temperature which cannot be used to generate useful work. The Austrian physicist Ludwig Boltzmann used entropy in statistical mechanics to quantify the *randomness (disorder)* of a system. The statistical mechanics definition of entropy reflects the number of microscopic configurations which are accessible by the system. In the twentieth century, the concept of entropy was used by the American mathematician Claude Shannon (1948) to measure the average information contained in signals. Shannon's influential paper founded the field of *information theory*. Consequently, the terms *Shannon entropy* and *information entropy* are used to distinguish between the entropy content of signals and the mechanistic notion of entropy used in thermodynamics and statistical mechanics.

The connection between information theory and statistical mechanics was investigated in two seminal papers by the American physicist Edwin T. Jaynes (1957a, b). He showed that the formulation of statistical mechanics can be derived from the principle of maximum entropy without the need for additional assumptions. The *principle of maximum entropy* is instrumental in establishing this connection; it dictates that given partial knowledge of the system, the least biased estimate for the probability distribution maximizes Shannon entropy under the specified constraints. According to Jaynes, “Entropy maximization is not an application of a law of physics, but merely a method of reasoning that ensures that no arbitrary assumptions are made.” The work of Jaynes opened the door for the application of MEM to various fields of science and engineering that involved ill-posed problems characterized by incomplete information. Notable application areas include spectral analysis (Burg 1972), image restoration (Skilling and Bryan 1984), geostatistics (Christakos 1990), quantum mechanics, condensed matter physics, tomography, crystallography, chemical spectroscopy, and astronomy, among others (Skilling 2013).

Methodology

According to *Laplace's principle of indifference*, if prior knowledge regarding possible outcomes of an experiment is unavailable, the uniform probability distribution is the most impartial choice. MEM employs this principle by

incorporating data-imposed constraints in the model inference process. The MEM probability model depends on the constraints used: The MEM model for a nonnegative random variable with known mean is the exponential distribution. If the constraints include the mean and the variance, the Gaussian (normal) distribution is obtained. In the case of multivariate and spatially distributed processes, the MEM model constrained on the mean and the covariance function is the joint Gaussian distribution.

Notation

In the following, it is assumed that $X = (X_1, \dots, X_n)^T$ (T denotes the transpose) is an n -dimensional random vector defined in a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, where Ω is the sample space, $\mathcal{F}$ is the sigma-algebra of events, and $\mathcal{P}$ is the probability measure. The realizations $\mathbf{x} \in \mathbb{R}^n$ of the random vector $\mathbf{X}$ can take either discrete or continuous values.

The expectation of the function $g(\mathbf{X})$ over the ensemble of states $\mathbf{x}$ is denoted by $\mathbb{E}[g(\mathbf{X})]$. The expectation involves the *joint probability mass function* (PMF) $p(\mathbf{x})$ if the random variables X_i take discrete values or the *joint probability density function* (PDF) $f(\mathbf{x})$ if the X_i are continuous. We use the *trace operator*, Tr , as a unifying symbol to denote summation (if the random variables X_i are discrete) or integration (if the X_i are continuous) over all probable states $\mathbf{x}$.

General Formulation

Assuming that the PMF $p(\mathbf{x})$ (in the discrete case) or the PDF $f(\mathbf{x})$ (in the continuous case) is known, Shannon's entropy can be expressed as

$$S = - \sum_{\mathbf{x} \in \Omega} p(\mathbf{x}) \ln p(\mathbf{x}), \quad \text{discrete}, \quad (1)$$

$$S = - \int_{\mathbb{R}} dx_1 \dots \int_{\mathbb{R}} dx_n f(\mathbf{x}) \ln f(\mathbf{x}), \quad \text{continuous}. \quad (2)$$

The entropy for both the discrete and continuous cases can be expressed as $S = -\mathbb{E}[\ln f(\mathbf{x})]$, where $f(\cdot)$ here stands for the PMF in the discrete case and the PDF in the continuous case. For the sake of brevity, in the following, we use $f(\cdot)$ to denote the PDF or the PMF and refer to it as “probability distribution.” We also use the term “summation” to denote either summation (for discrete variables) or integration (for continuous variables) over all possible states $\mathbf{x}$.

Let $\{g_m(\mathbf{x})\}_{m=1}^M$ represent a set of M sampling functions, the averages of which can be determined from the data

(observations). We will denote sample averages by means of $\overline{g_m(\mathbf{x})}$. Such sampling averages can include the mean, variance, covariance, higher-order moments, or more complicated functions of the sample values. Then, according to the *principle of maximum entropy*, the probability distribution should respect the constraints (the symbol $\equiv$ denotes equivalence)

$$\mathbb{E}[g_m(\mathbf{x})] \equiv \text{Tr}_{\mathbf{x}}[g_m(\mathbf{x})f(\mathbf{x})] = \overline{g_m(\mathbf{x})}, \quad m = 1, \dots, M, \quad (3)$$

where $\text{Tr}_{\mathbf{x}}[\cdot]$ denotes the “summation” over all possible states $\mathbf{x}$. The above equations are supplemented by the *normalization constraint* $\text{Tr}_{\mathbf{x}}f(\mathbf{x}) = 1$ which ensures the proper normalization of the probability distribution.

The maximum entropy (henceforward, *MaxEnt*) distribution maximizes the entropy under the above constraints. This implies a constrained optimization problem defined by means of the following Lagrange functional

$$\begin{aligned} \mathcal{L}[f] = & -S + \sum_{m=1}^M \lambda_m [\overline{g_m(\mathbf{x})} - \text{Tr}_{\mathbf{x}}g_m(\mathbf{x})f(\mathbf{x})] \\ & + \lambda_0 [1 - \text{Tr}_{\mathbf{x}}f(\mathbf{x})]. \end{aligned} \quad (4)$$

The minimization of $\mathcal{L}[f]$ can be performed using the calculus of variations to find the stationary point of the Lagrange functional. At the stationary point, the functional derivative $\delta\mathcal{L}/\delta f$ vanishes, i.e.,

$$0 = \frac{\delta\mathcal{L}}{\delta f} = \ln f(\mathbf{x}) + 1 - \sum_{m=1}^M \lambda_m g_m(\mathbf{x}) - \lambda_0.$$

The above equation leads to the *MaxEnt probability distribution*

$$f(\mathbf{x}) = \frac{1}{Z} \exp \left[\sum_{m=1}^M \lambda_m g_m(\mathbf{x}) \right], \quad \ln Z = 1 - \lambda_0, \quad (5)$$

where the constant Z is the so-called partition function which normalizes the MaxEnt distribution. Since $f(\mathbf{x})$ is normalized by construction, it follows from Eq. (5) that

$$Z = \text{Tr}_{\mathbf{x}} \exp \left[\sum_{m=1}^M \lambda_m g_m(\mathbf{x}) \right]. \quad (6)$$

The implication of Eqs. (5) and (6) is that λ_0 depends on the Lagrange multipliers $\{\lambda_m\}_{m=1}^M$. These need to be determined by solving the following system of M nonlinear constraint equations

$$\overline{g_m(\mathbf{x})} = \text{Tr}_{\mathbf{x}} g_m(\mathbf{x}) f(\mathbf{x}) = \frac{\partial Z}{\partial \lambda_m}, \quad m = 1, \dots, M. \quad (7)$$

In spatial and spatiotemporal problems, the constraints involve joint moments of the field (Christakos 1990, 2000; Hristopulos 2020). The constraints are expressed in terms of real-space coordinates. In the case of time series analysis, it is customary to express the MEM solution in the spectral domain (see next section).

Spectral Analysis

MEM has found considerable success in geophysics as a method for estimating the power spectrum of stationary random processes (Burg 1972; Ulrych and Bishop 1975). In spectral analysis, MEM is also known as *all poles* and *autoregressive (AR) method*.

For a time series with a constant time step δt , the MEM power spectral density at frequency f is given by

$$P(f) = \frac{c_0}{|1 + \sum_{m=1}^M c_m \exp(2\pi i f m \delta t)|^2}, \quad -f_N \leq f \leq f_N, \quad (8)$$

where $f_N = 1/2\delta t$ is the Nyquist frequency (i.e., the maximum frequency which can be resolved with the time step δt) and $\{c_m\}_{m=0}^M$ are coefficients which need to be estimated from the data.

The term “all-poles” method becomes obvious based on Eq. (8), since $P(f)$ has poles in the complex plane. The poles coincide with the zeros in the denominator of the fraction that appears in the right-hand side of Eq. (8). The connection with AR time series models of order m lies in the fact that the latter share the spectral density given by Eq. (8). The order of the AR model which is constructed based on MaxEnt is equal to the maximum lag, $m\delta t$, for which the auto-covariance function can be reliably estimated based on the available data.

Applications

Maximum entropy has been extensively used in many disciplines of geoscience. Notable fields of application include geophysics, seismology and hydrology, estimation of rainfall variability and evapotranspiration, assessment of landslide susceptibility, classification of remote sensing imagery, prediction of categorical variables, investigations of mineral potential prospectivity, and applications in land use and climate models.

The principle of maximum entropy is used in Bayesian inference to obtain prior distributions (Skilling 2013). This

connection has been exploited in the Bayesian MaxEnt (BME) framework for spatial and spatiotemporal model construction and estimation (Christakos 1990, 2000). BME allows incorporating prior physical knowledge of the spatial or spatiotemporal process in the probability model. In spatial estimation problems, the application of the method of maximum entropy assumes that the mean, covariance, and possibly higher-order moments provide the spatial constraints for random field models. A different approach proposes local constraints that involve geometric properties, such as the square of the gradient and the linearized curvature of the random field (Hristopulos 2020). This approach leads to spatial models with sparse precision matrix structure which is computationally beneficial for the estimation and prediction of large datasets. In recent years, various generalized entropy functions have been proposed (e.g., Rényi, Tsallis, Kaniadakis entropies) for applications that involve long-memory, non-ergodic, and non-Gaussian processes. In principle, it is possible to build generalized MEM distributions using extended notions of entropy (Hristopulos 2020). The mathematical tractability and the application of such generalized maximum entropy principles in geoscience are open research topics.

Summary or Conclusions

The notion of entropy is central in statistical mechanics and information theory. Entropy provides a measure of the information incorporated in a specific signal or dataset. The principle of maximum entropy can be used to derive probability distributions which possess the largest possible uncertainty given the constraints imposed by the data. Equivalently, maximum entropy minimizes the amount of prior information which is integrated into the probability distribution. Thus, the maximum entropy probability distributions are unbiased with respect to what is not known. The most prominent applications of maximum entropy in geoscience include the estimation of spectral density for stationary random processes and the construction of joint probability models for spatial and spatiotemporal datasets.

Cross-References

- [Fast Fourier Transform](#)
- [High-Order Spatial Stochastic Models](#)
- [Power Spectral Density](#)
- [Spatial Analysis](#)
- [Spatial Statistics](#)
- [Spectral Analysis](#)
- [Stochastic Geometry in the Geosciences](#)
- [Time Series Analysis](#)

Bibliography

- Burg JP (1972) The relationship between maximum entropy spectra and maximum likelihood spectra. *Geophysics* 37(2):375–376
- Christakos G (1990) A Bayesian/maximum-entropy view to the spatial estimation problem. *Math Geol* 22(7):763–777
- Christakos G (2000) Modern spatiotemporal geostatistics, International Association for Mathematical Geology Studies in mathematical geology, vol 6. Oxford University Press, Oxford
- Hristopulos DT (2020) Random fields for spatial data modeling: a primer for scientists and engineers. Springer Netherlands, Dordrecht
- Jaynes ET (1957a) Information theory and statistical mechanics. I. *Phys Rev* 106(4):620–630
- Jaynes ET (1957b) Information theory and statistical mechanics. II. *Phys Rev* 108(2):171–190
- Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27(3):379–423
- Skilling J (2013) Maximum entropy and Bayesian methods: Cambridge, England, 1988, vol 36. Springer Science & Business Media, Cham
- Skilling J, Bryan R (1984) Maximum entropy image reconstruction—general algorithm. *Mon Not R Astron Soc* 211:111–124
- Ulfrych TJ, Bishop TN (1975) Maximum entropy spectral analysis and autoregressive decomposition. *Rev Geophys* 13(1):183–200

Maximum Entropy Spectral Analysis

Eulogio Pardo-Igúzquiza¹ and Francisco J. Rodríguez-Tovar²

¹Instituto Geológico y Minero de España (IGME), Madrid, Spain

²Universidad de Granada, Granada, Spain

Synonyms

Autoregressive spectral estimation, all-poles spectral estimation

Definition

Maximum entropy spectral analysis (MESA) is the statistical estimation of the power spectrum of a stationary time series using the maximum entropy (ME) method. The resulting ME spectral estimator is parametric and equivalent to an autoregressive (AR) spectral estimator.

General Comments

In geosciences, there is a plethora of time series of interest in applied geosciences. A sequence of varve thicknesses, a seismogram or the isotopic content in a stalagmite are but three common examples in cyclostratigraphy, geophysics and paleoclimatology, respectively. The time series is modeled as a stochastic process (or random function) where a topic

of frequent interest is the estimation of its power spectrum (or spectral density) which is a function that describes the distribution of the variance (power or energy) of the time series across frequencies. In general, the power spectrum of a time series is unknown in advance and must be estimated from the experimental data. The main task of spectral analysis is thus the estimation of the power spectrum for the available data. There are many methods for the estimation of the power spectrum, like the periodogram, the Blackman–Tukey approach, multitaper estimators, singular value decomposition analysis, etc. MESA is another important spectral analysis method that must be added to the list, being a popular choice in geosciences (Ables 1974), and particularly appealing because of its high resolution and its good performance with short time series. The method was first proposed by Burg (1967).

MESA Method

The maximum entropy spectral estimator (Burg 1967; Papoulis 1984) is the spectral density $\hat{S}(\omega)$ that maximizes the entropy (E):

$$E = \int_{-\pi}^{\pi} \ln\{S(\omega)\} d\omega, \quad (1)$$

over the class of all densities $S(\omega)$ that satisfy the constraints given by making (theoretical) covariances equal to sample covariances (i.e. covariances estimated from the data):

$$\int_{-\pi}^{\pi} e^{i\omega h} S(\omega) d\omega = \hat{C}(h), \quad (2)$$

for $h \in [-q, q + 1, \dots, -1, 0, 1, \dots, q]$.

Where $\hat{C}(h) = \hat{C}(-h)$ is the estimated covariance for lag h , i is the imaginary unit and $\omega = 2\pi f$ is the angular frequency (radians per sampling interval) while f is the frequency in cycles per sampling interval.

Next, it may be shown that the ME spectral estimator has the form (Brockwell and Davis 1991):

$$\hat{S}(\omega) = \frac{\sigma_q^2}{\left| 1 + \sum_{k=1}^q a_k e^{-ik\omega} \right|^2}, \quad (3)$$

which is equal to the AR spectral estimator of order q . Strictly speaking the AR estimator provided in Eq. (3) is the ME estimator when the random function (stochastic process) is Gaussian (Papoulis 1984).

In the previous equations: σ_q^2 is the variance of a zero-mean white noise stochastic process that depends on the order q of

the AR process, and $\{a_k; k = 1, \dots, q\}$ are the q autoregressive coefficients that are also dependent on the chosen order q . These $q + 1$ parameters must be estimated from the experimental data in order to apply the ME spectral estimator of Eq. (3).

A possibility for obtaining the estimation of the previous $q + 1$ parameters is to use the Yule-Walker equations (Papoulis 1984). However, in order to avoid the inversion of large covariance matrixes, it is possible to use the Durbin-Levinson algorithm (Brockwell and Davis 1991), which gives the estimates of the AR coefficients and noise variance recursively. Another popular recursive algorithm is given by Burg (1967) which uses estimates of forward and backward estimation errors and their mean square values (Papoulis 1984).

Advantages of MESA

High frequency resolution. MESA is capable of resolving discrete frequencies close together even in short time series.

Consistent power spectrum estimates. In this sense, MESA produces neither negative estimates nor those that do not coincide with their sample covariance values.

There is no leakage of power between nearby frequencies. MESA does not experience the side lobe problem that is

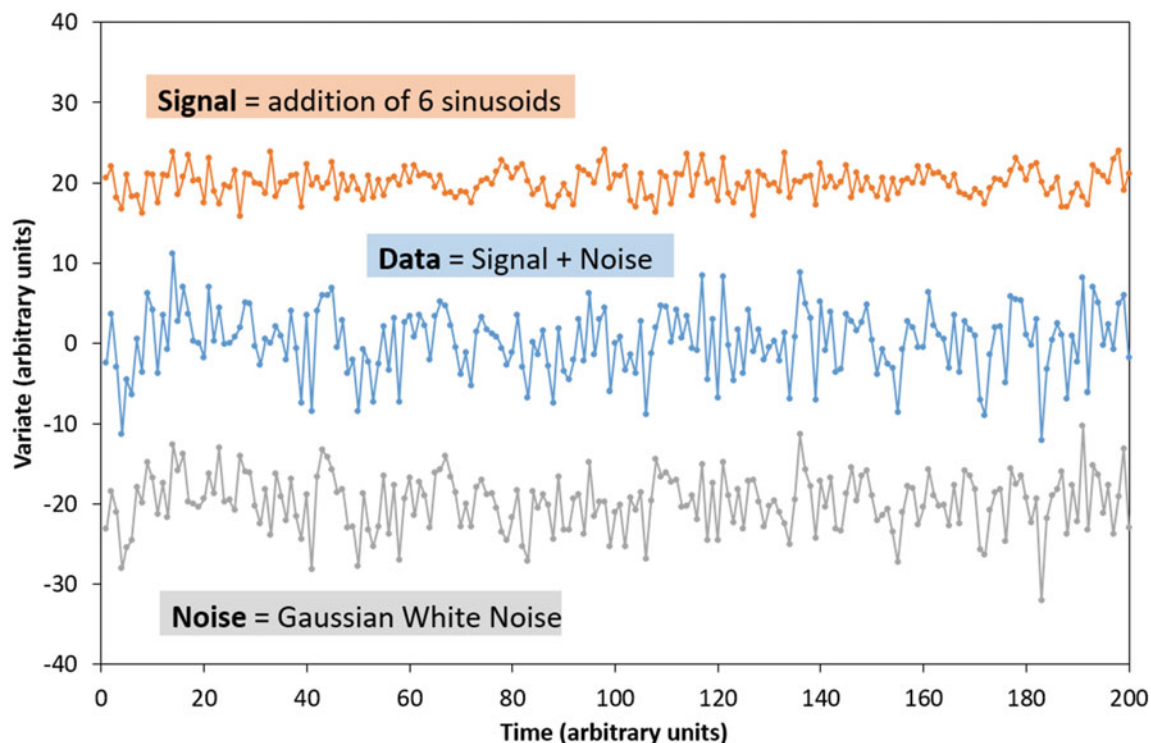
common with other spectral estimators and that is caused by the use of window functions.

Efficient spectrogram analysis. MESA can be efficiently used for estimating the spectrogram (Pardo-Igúzquiza and Rodríguez-Tovar 2006) because it achieves good results with short time series.

Difficulties in the Application of MESA

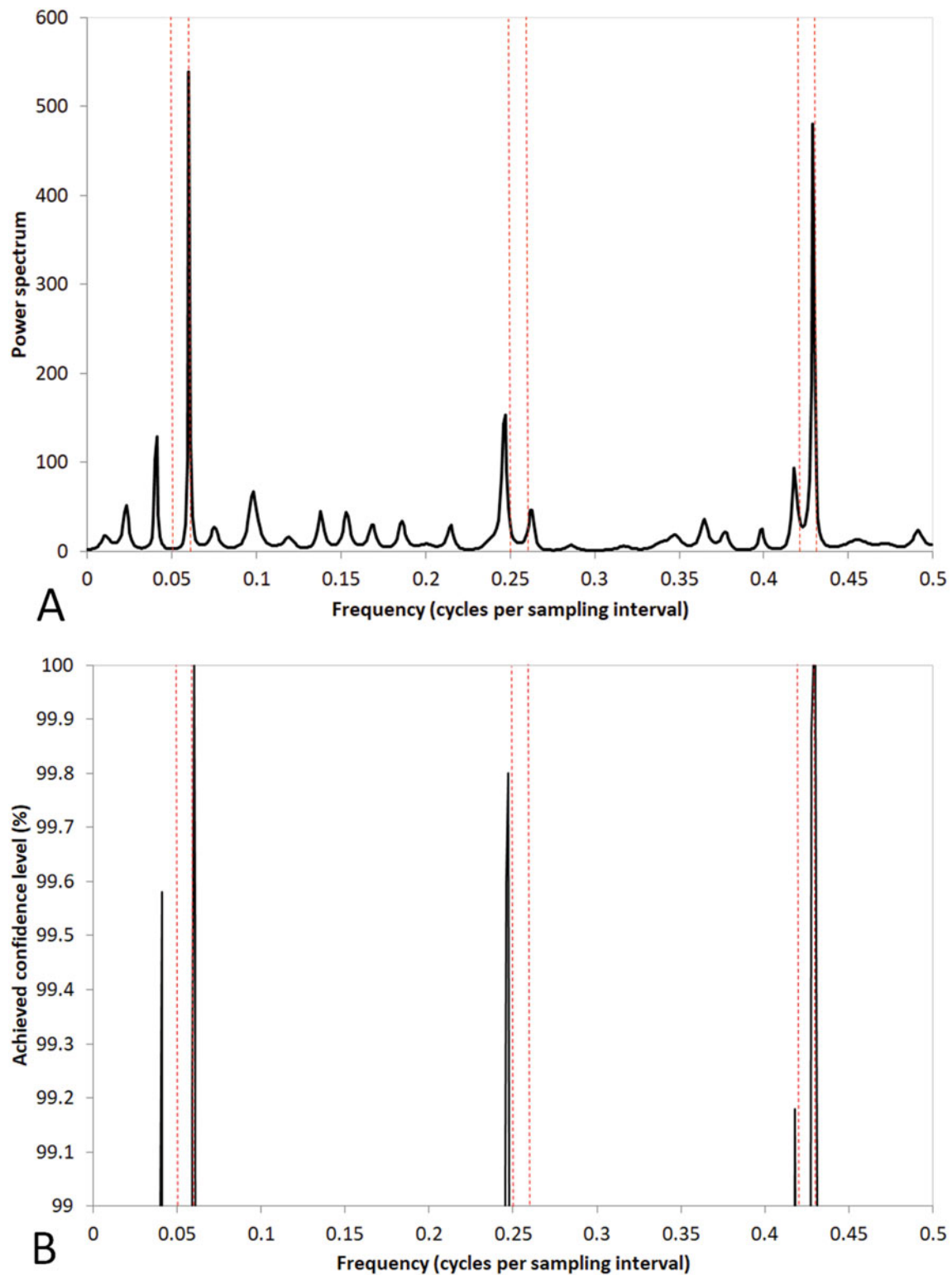
An important decision is the choice of an optimum autoregressive order q for the MESA estimator given in Eq. (3). Many alternatives have been proposed. Some methods are based on criteria like the Akaike information criterion or the final error prediction (Ulrych and Bishop 1975). The order is often chosen empirically with a value between $N/5$ and $N/2$, N being the number of experimental data points. $2N/\ln(2N)$ is another popular empirical choice. If the chosen order is small, the variance of the estimates is small, but the bias is large, while if the chosen order is large, the bias is small, but the variance increases.

An optimal choice of the order q results in problems of spectrum line splitting, frequency shifting, spurious peaks and biased power estimation. However, these problems are



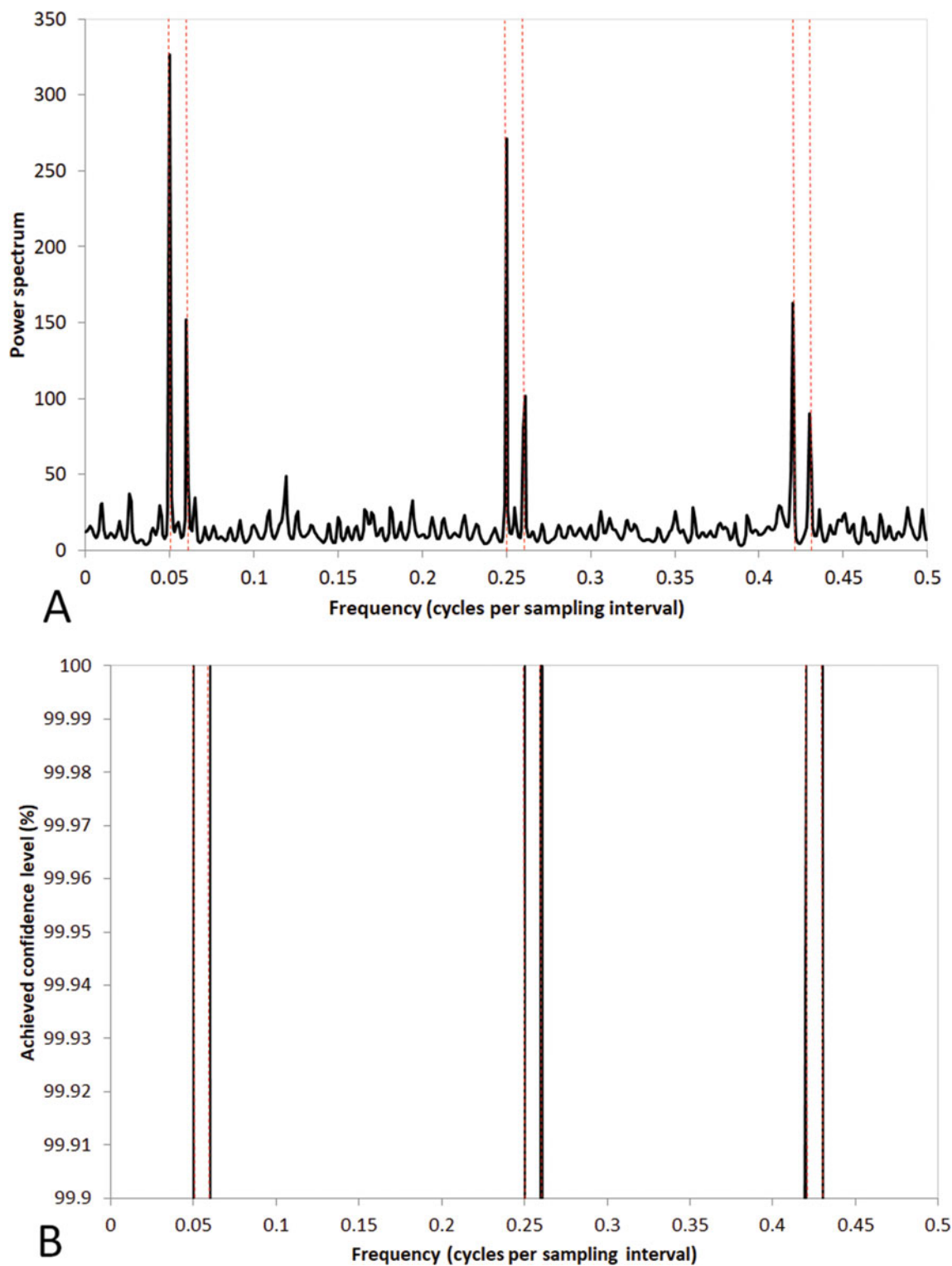
Maximum Entropy Spectral Analysis, Fig. 1 Simulated data time series (Data in the figure) equal to the addition of signal (Signal in the figure) and noise (Noise in the figure). The similarity of signal and noise and how the spectral content is hidden in the data and not at all evident

are both remarkable. The number of data points is 200, which can be considered a short time series. The signal has been shifted upwards by adding a value of 15 and the noise has been shifted downwards by a value of -15 to avoid the overlap and provide a clearer representation



Maximum Entropy Spectral Analysis, Fig. 2 (a): MESA estimated power spectrum of the observed data in Fig. 1 (200 experimental data points). (b): Achieved confidence level of the estimated power spectrum

given in A. The vertical dashed red lines give the location of the theoretical frequencies of the sinusoids of the signal



Maximum Entropy Spectral Analysis, Fig. 3 (a): MESA estimated power spectrum of the observed data with a simulation scheme identical to the one used in Fig. 1 but extending the length of the different time series from 200 to 2000 data points. (b): Achieved confidence level of

the estimated power spectrum given in A. The vertical dashed red lines show the location of the theoretical frequencies of the sinusoids of the signal

common to all spectral methods if the data are of low quality (few data points, irregular sampling, high level of noise, etc.).

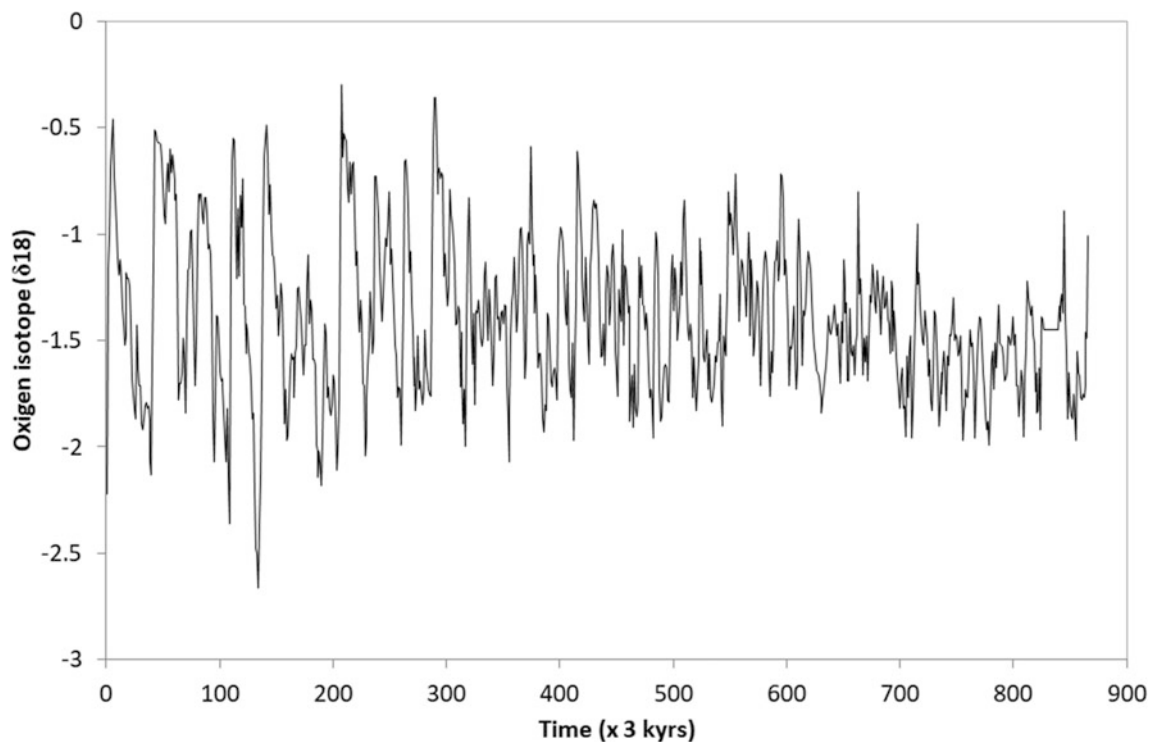
Another difficulty is the statistical evaluation of the reliability or uncertainty of the estimated power spectrum. This evaluation is complex and must be based on strong assumptions regarding the model that should be followed by the experimental data (Baggeroer 1976). This problem has been solved by using the computer intensive permutation test method (Pardo-Igúzquiza and Rodríguez-Tovar 2005). The permutation test is easy to apply, non-parametric but based on the computer intensive method. The original time series is ordered at random (a permutation sequence) many times (for example 5000 times) and for each disordered sequence the ME power spectrum is estimated. The statistical significance of the power spectrum estimated from the original time series may then be assessed by calculating how many times it is higher than the power spectrum at the same frequency for all permutation sequences.

Examples of Application

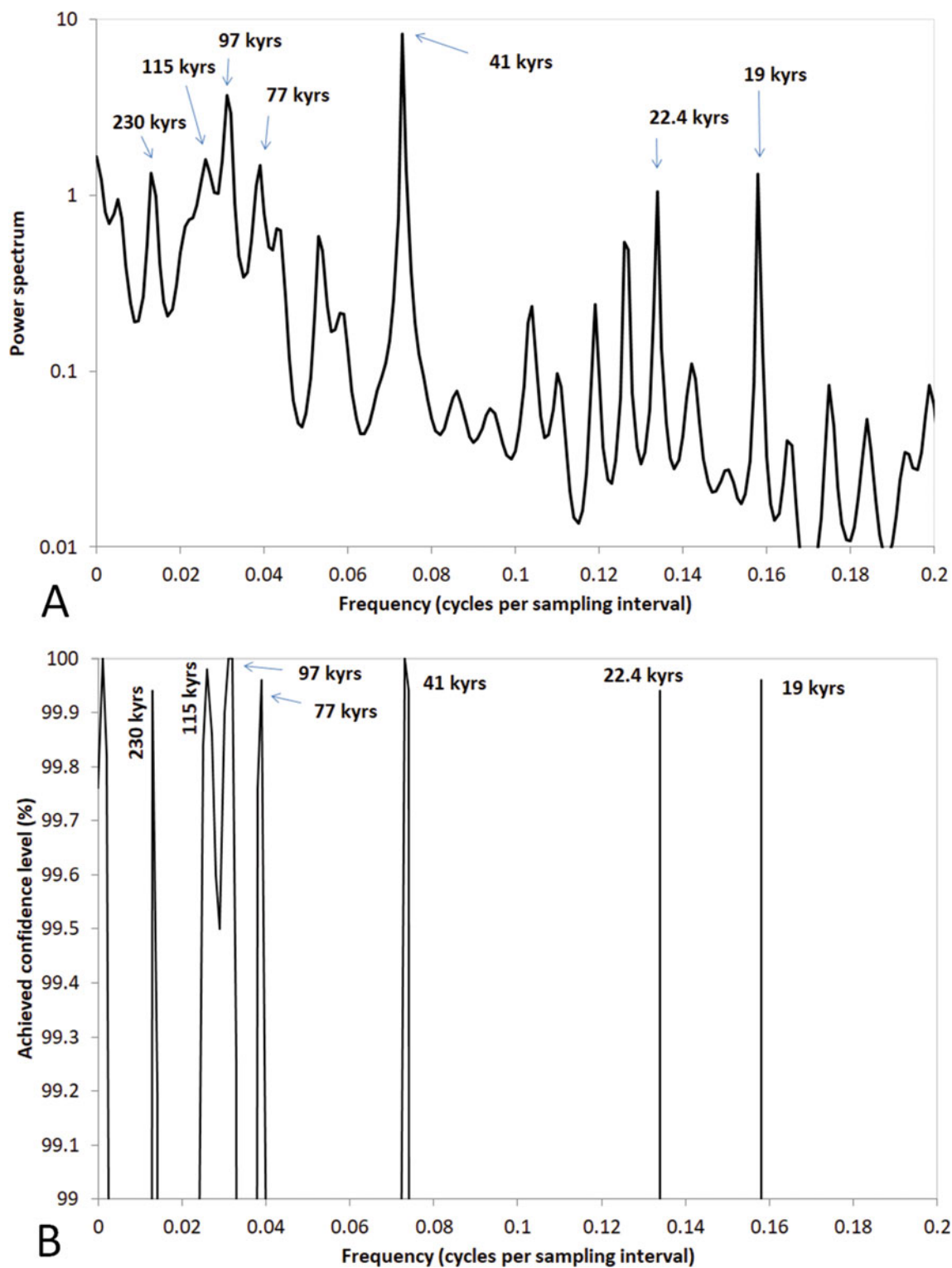
The first example is a short time series (200 data points) that has been simulated following the model: observed data equal to signal plus noise. The signal consists of the superposition of six sinusoids of equal amplitude with frequencies forming three pairs: one in the low (0.05 and 0.06), another in the

middle (0.25 and 0.26) and a third pair in the high frequencies (0.42 and 0.43), where all these are given in cycles per sampling interval that is considered unity. The amplitude of each sinusoid gives a variance of the signal equal to 3. The white noise is a sequence of Gaussian numbers drawn from a Gaussian distribution with mean zero and variance 12. Thus, the signal-to-noise ratio (SNR) of variances is $3/12 = 0.25$. The observed data, signal and noise are represented in Fig. 1. The high similitude of observed data and noise is evident because of the low SNR and it can be correctly said that the signal is hidden in the noise, thus it is a very challenging case. The estimated power spectrum (with $p = 67$) of the time series of observed data is shown in Fig. 2A and the achieved confidence level of the permutation test is shown in Fig. 2B. Figure 2A demonstrates how the three couples of sinusoidal components have been identified although with a frequency shift in the lowest frequency component. Figure 2B demonstrates how all the components have been identified with high statistical significance ($> 99\%$) except the fourth component, i.e. the high frequency component in the middle frequencies. When the number of data points of those observed (signal + noise) is increased to 2000, Figs. 3A (with $p = 150$) and 3B show how these difficulties disappear.

The second example is a sequence of real data (oxygen isotope of $\delta^{18}\text{O}$) measured along a drill core sample of ocean sediment (Raymo et al. 1992). The time series has been represented in Fig. 4 and it has 866 data points with a



Maximum Entropy Spectral Analysis, Fig. 4 Time series of 866 data points of oxygen isotope $\delta^{18}\text{O}$ (Raymo et al. 1992)



Maximum Entropy Spectral Analysis, Fig. 5 (a): MESA estimated power spectrum of the time series shown in Fig. 4. (b): Achieved confidence level of the estimated power spectrum given in A

sampling interval of 3 kyrs. The power spectrum estimated by MESA with an autoregressive order $p = 150$ is shown in Fig. 5A and the significance level achieved by the permutation test is shown in Fig. 5B. Figure 5B demonstrates how the most statistically significant frequencies correspond to periodicities of 19, 22.4, 41, 77, 97, 115 and 230 kyrs. These estimated cycles are in line with Milankovitch periodicities of precession (19 and 22.4 kyrs), obliquity (41 kyrs) and short eccentricity (97 and 115 kyrs). The 77 kyrs can be explained as a combination tone of the main periodicities and the 230 kyrs a harmonic of eccentricity.

Summary

Maximum entropy spectral analysis is a high-resolution spectral estimator that has been widely used in geosciences. It is based on choosing the spectrum that corresponds to the most random (i.e. unpredictable) time series whose covariance function coincides with the known values (Burg 1975). It comes out that the solution to choosing the maximum entropy spectrum is equal to the autoregressive estimator of the power spectrum. The main problem is to select an optimal value for the autoregressive order q . There are several proposals for choosing a reasonable order of the estimator and the statistical significance of the estimated spectrum can be easily assessed by a computer intensive method like the permutation test.

Cross-References

- [Autocorrelation](#)
- [Bayesian Maximum Entropy](#)
- [Entropy](#)
- [Maximum Entropy Method](#)
- [Moving Average](#)
- [Power Spectral Density](#)
- [Signal Analysis](#)
- [Signal Processing in Geosciences](#)
- [Spectral Analysis](#)
- [Time Series Analysis](#)
- [Time Series Analysis in the Geosciences](#)

Bibliography

- Ables JG (1974) Maximum entropy spectral analysis. *Astron Astrophys Suppl* 15:383–393
- Baggeroer AB (1976) Confidence intervals for regression (MEM) spectral estimates. *IEEE Trans Inf Theory* 22(5):534–545
- Brockwell PJ, Davis RA (1991) *Time series: theory and methods*, 2nd edn. Springer, New York, p 577pp

- Burg JP (1967) Maximum entropy spectral analysis. 37th Annual International Meeting of the Society for the Exploration of Geophysics. Oklahoma City, OK, pp 34–41
- Burg JP (1975) Maximum entropy spectral analysis. PhD Dissertation, Stanford University, 127 p
- Papoulis A (1984) *Probability, random variables and stochastic processes*: McGraw-Hill Intern. Editions, Singapore, 576 p
- Pardo-Igúzquiza E, Rodríguez-Tovar FJ (2005) MAXENPER: a program for maximum entropy spectral estimation with assessment of statistical significance by the permutation test. *Comput Geosci* 31(5): 555–567
- Pardo-Igúzquiza E, Rodríguez-Tovar FJ (2006) Maximum entropy spectral analysis of climatic time series revisited: assessing the statistical significance of estimated spectral peaks. *J Geophys Res Atmos* 111(D10202):1–8
- Raymo ME, Hodell D, Jansen E (1992) Response of deep ocean circulation to the initiation of northern hemisphere glaciation (3–2 myr). *Paleoceanography* 7:645–672
- Ulrych TJ, Bishop TN (1975) Maximum entropy spectral analysis and autoregressive decomposition. *Rev Geophys Space Phys* 13(1): 183–200

Maximum Likelihood

Jaya Sreevalsan-Nair

Graphics-Visualization-Computing Lab, International Institute of Information Technology Bangalore, Electronic City, Bangalore, Karnataka, India

Definition

Estimating a probability distribution to fit the available observed data is an important process required by statisticians and data analysts. Such estimation is required for prediction/forecasting, missing data analysis, learning functions with uncertain outcomes, etc. where the estimated probability distribution model of the observed data provides the representative behavior. Once a set of random variables has been identified to model specific observation data, a joint probability distribution is to be defined on the set. The joint probability density function of n independent and identically distributed (i.i.d.) observations data $x_1, x_2, \dots, x_n$ from this process, where the probability function is conditioned on a set of parameters θ , is given by likelihood function $f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i|\theta) = L(\theta|\mathbf{x})$. This function is representative of all information of the sample population from which the observations are made, and this information is what gets used for estimation in the form of an estimator.

The widely used estimators are classified based on the distribution type, as parametric, semi-parametric, and non-parametric estimation methods (Greene 2012). Maximum

likelihood estimation (MLE) is a classical parametric estimation method, generalized method of moments (GMM) is a semi-parametric estimation method, and kernel density estimation is a non-parametric estimation method. Parametric estimators provide the advantage of a consistent approach owing to which it can be applied to a variety of estimation problems. However, it suffers from the limitation of the robustness of underlying assumptions used, when imposing normal, logistic, or any other widely used distributions to fit the given data. The semi- and nonparametric distributions alleviate the restriction of the model but then often suffer from the issue of providing a range of probabilities as outcomes.

As the name suggests, MLE maximizes the likelihood function (Fisher 1912), thus, providing the argument of the likelihood function, θ , with the most probable (likely) occurrence. This is the most intuitive method, as it stems from the joint probability density or mass function of continuous and discrete random variables, respectively. The principle of maximum likelihood states that the most optimal choice of the estimator $\hat{\theta}$ from a set of estimators θ is the one that makes the observed data $\mathbf{x}$ most probable. Thus, this principle provides a choice of an asymptotically efficient estimator for a parameter or a set of parameters.

To improve the practical implementation of determining the maximum value of $L(\theta | \mathbf{x})$, its logarithm, referred to as log-likelihood is widely used as the likelihood function in MLE. Log-likelihood function is given as $\ln L(\theta | \mathbf{x}) =$

$\ln \left(\prod_{i=1}^n f(x_i | \theta) \right) = \sum_{i=1}^n \ln f(x_i | \theta)$. The use of log-likelihood functions is more popular due to three key reasons. Firstly, it gives an accurate outcome, as the logarithm of the function attains the maximum when the function is at its maximum value. Secondly, in terms of convenience of using log-likelihood functions instead of the likelihood functions themselves, for most distributions, and especially the widely used exponential families of distributions, the former give simpler terms for extremum computation than the latter. Thirdly, using additive models is more convenient in most mathematical operations, and thus, the summation in log-likelihood functions is more convenient than the equivalent product in the likelihood function. Thus, the MLE widely considers the logarithm function as the likelihood function.

Overview

A maximum likelihood estimator is a type of extremum estimator, that optimizes a criterion function, which is either $L(\theta | \mathbf{x})$ or $(\ln L(\theta | \mathbf{x}))$ in the case of MLE, and the outcome is the estimator, which is the argument θ at its extremum value.

Least-squares and GMM are other extremum estimators. MLE and least-squares estimator fall under the subclass of M-estimators, where the criteria function is a sum of terms (Greene 2012). Thus, for maximum likelihood estimation of given independent and identically distributed (i.i.d) observation data $x_1, x_2, \dots, x_n$, and selected parameters θ , its estimated value is given as: $\hat{\theta} = \arg \max \left[\frac{1}{n} \sum_{i=1}^n \ln f(x_i | \theta) \right]$.

In practice, estimation properties are evaluated to determine the optimal model within a class of estimators. Maximum likelihood methods have desirable properties, such as the outcomes tend to have minimum variance, thus the smallest confidence interval, and they have approximately normal distributions and sample variances to be used for confidence bounds determination and hypothesis testing. Being an M estimator, MLE has desirable properties of consistency and asymptotic normality. Suppose θ_0 is the unknown parameter of the selected model for which θ is the approximated one, *consistency* of the estimator implies $\hat{\theta} \rightarrow \theta_0$, as $n \rightarrow \infty$. *Asymptotic normality* of $\hat{\theta}$ implies that in addition to the estimator definitely converging to the unknown parameter, it does so at the rate of $\frac{1}{\sqrt{n}}$. This implies faster convergence with larger observations/sample set, given as $\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow^d N(0, \sigma_{\theta_0}^2)$, where $\sigma_{\theta_0}^2$ is called as the asymptotic variance of the estimate $\hat{\theta}$ (Panchenko 2006). At the same time, MLE suffers from the limitations of specificity of the likelihood equations, nontrivial numerical estimation, heavy bias in the case of small samples, and sensitivity to initial conditions (Heckert and Filliben 2003).

MLE is attributed to R. A. Fisher owing to the official documented appearance of the term “maximum likelihood” in literature (Fisher 1912). However, there is evidence of the usage of MLE prior to 1912, albeit with the use of different terminology, e.g., “most probable.” Early variants of MLE include estimating using the maximum value of the posterior distribution, referred to as inverse probability, and minimum of variance of “probable errors” (Edgeworth 1908). The seminal works by both Edgeworth and Fisher have independently established the efficiency of maximum likelihood and the minimization of the asymptotic variance of M-estimators (Pratt 1976). While Edgeworth and predecessors, including Gauss, Laplace, and Hagen, used the principle of inverse probability using posterior mode as an estimator, Fisher replaced the posterior mode with the MLE, thus achieving invariance (Hald 1999). Today, the widespread use of MLE may be equally attributed to both Edgeworth and Fisher for providing proof of the optimality of $\hat{\theta}$. Apart from the use of the maximum likelihood principle for estimation, it can also be effectively used for testing a model using the log-likelihood ratio statistic (Akaike 1973). Thus, these two implications of the principle can be combined to a single problem of statistical decision.

Applications

MLE has been used in geoscientific applications for a long time, which is evident from the fact that geodesists and civil engineers have been early adopters of the predecessors to MLE (Hald 1999). Today, MLE in itself is used in a variety of geoscientific applications, of which a few are listed here. It is one of the methods used in predicting the occurrence of mineral deposits at a regional scale, electrofacies classification in reservoirs for petroleum engineering, parameter estimation for reservoir modeling using multimodal data (geophysical logs, remote sensing images, etc.), and quantifying epistemic uncertainty in geostatistical applications pertaining to natural resources and environment (Sagar et al. 2018).

The principle of maximum likelihood is also extended to semantic classification, applied to satellite images (Bolstad and Lillesand 1991). In maximum likelihood classification, the different classes, which are present in the bands of multispectral satellite images, are modeled using different normal distributions, and the labeling is done for each pixel-based on the probability it belongs to each class.

Future Scope

The scope of applications of utilizing maximum likelihood is continuously growing larger, owing to its popularity over several decades. Increasingly, the maximum likelihood-based methods, such as MLE and classification, use combinations of methods, e.g., polynomial methods, to improve the efficiency of the data analysis methods used in the application. Overall, the maximum likelihood estimation is an efficient statistical modeling and inference method.

Bibliography

- Akaike H (1973) Information theory and an extension of the maximum likelihood principle. *Proceeding of the second international symposium on information theory*, pp 267–281
- Bolstad P, Lillesand T (1991) Rapid maximum likelihood classification. *Photogramm Eng Remote Sens* 57(1):67–74
- Edgeworth FY (1908) On the probable errors of frequency-constants. *J R Stat Soc* 71(2):381–397
- Fisher RA (1912) On an absolute criterion for fitting frequency curves. *Messenger Math* 41:155–156
- Greene WH (2012) *Econometric Analysis*, 7th edn. Prentice Hall, Pearson
- Hald A (1999) On the history of maximum likelihood in relation to inverse probability and least squares. *Stat Sci* 14(2):214–222
- Heckert NA, Filliben JJ (2003) NIST/SEMATECH e-handbook of statistical methods; chapter 1: exploratory data analysis. <https://www.itl.nist.gov/div898/handbook/eda/section3/eda3652.htm>
- Panchenko D (2006) 18.650 Statistics for Applications. Massachusetts Institute of Technology. MIT OpenCourseWare. <https://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006/lecture-notes/lecture3.pdf>
- Pratt JW (1976) FY Edgeworth and RA Fisher on the efficiency of maximum likelihood estimation. *Ann Stat* 4(3):501–514
- Sagar BSD, Cheng Q, Agterberg F (2018) *Handbook of mathematical geosciences: fifty years of IAMG*. Springer Nature, Cham, Switzerland

McCammon, Richard B.

Michael E. Hohn
Morgantown, WV, USA



Fig. 1 Richard B. McCammon, courtesy of Dr. Richard McCammon

Biography

Richard McCammon, founding member of the International Association for Mathematical Geology (IAMG), Krumbein Medalist, editor, longtime scientist with the US Geological Survey (USGS), and researcher in geologic resources, was born on December 3, 1932, in Indianapolis, Indiana (USA), son of Bert McCammon, a pharmacist in that city, and wife Mary. He graduated from Broad Ripple High School in 1951 and went on to college at the Massachusetts Institute of Technology, receiving a Bachelor's degree in 1955. His initial ambition was to receive a degree in chemical engineering, but in his junior year he changed his major to geology, realizing the appeal of working outdoors. He received a Master's degree in geology from the University of Michigan in 1956 and a Ph.D. from Indiana University in 1959. His dissertation led to an interest in probability and statistics and a postdoctoral fellowship in statistics at the University of Chicago.

Following several academic positions and 7 years at Gulf Research and Development Company, Richard took up a position with the USGS in the mid-1970s, where he spent the balance of his career, achieving the position of Chief of the Branch of Resource Analysis in 1992. He conducted research in assessment of mineral resources and application of numerical methods to mineral exploration.

Author or coauthor of more than 100 publications and technical reports, Richard was an early user of cluster analysis and other multivariate methods for purposes of classification in the 1960s, when these methods were first being applied to geological problems. His research reports and papers at the USGS addressed numerical methods for estimating number and size of undiscovered mineral resources in a region.

His activities in the IAMG include serving as Western Treasurer, 1980–1984; Secretary General, 1984–1989; President, 1989–1992; and Past President 1992–1994. He was the second Editor in Chief for *Mathematical Geology* (now *Mathematical Geosciences*), serving in that capacity from 1976–1980. He negotiated with a publisher for works too long for journals, resulting in the *Monograph Series in Mathematical Geology* (now *Studies in Mathematical Geosciences*), which he edited 1987–1998. His interest in founding a journal dedicated to the appraisal of natural resources led him to again work with a publisher to create *Nonrenewable Resources* (now *Natural Resources Research*) and serve as its first editor in chief, 1992–1998.

In recognition for his research contributions, service to the profession, and service to the IAMG, Richard B. McCammon received the prestigious Krumbein Medal in 1992.

Acknowledgments I obtained personal, academic, and career details from Ricardo Olea's biography of Richard McCammon (*Mathematical Geology*, Vol. 27, p. 463). The IAMG website (www.iamg.org) lists his terms as an IAMG officer and editor. I would like to thank Frits Agterberg for confirming that Richard was present at the inaugural meeting in Prague, Czech Republic, in 1968; and B. S. Daya Sagar for providing the photograph.

Membership Functions

Candan Gokceoglu

Department of Geological Engineering, Hacettepe University, Ankara, Turkey

Definition

Unlike the classical set theory, the transitions in a universe can be smooth, and the definitions of elements can be mixed and diversified with the help of membership definitions of a given set. An element defined with fuzzy sets can be members of different classes, which ensure the gradual transition between them. The fuzzy sets can unfold and model the vagueness and the uncertainty in the universe by designing various degrees of memberships at the boundaries. Hence, a function needs to be designed for defining the memberships of the elements, and

these functions are defined as membership functions. Figure 1 shows a general difference between characteristic function for crisp set A and membership function for fuzzy set A.

Introduction

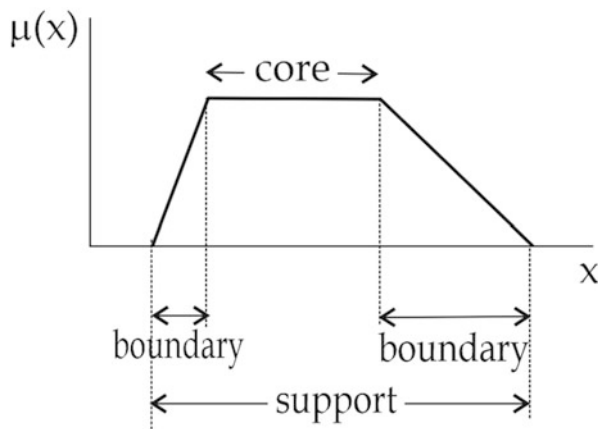
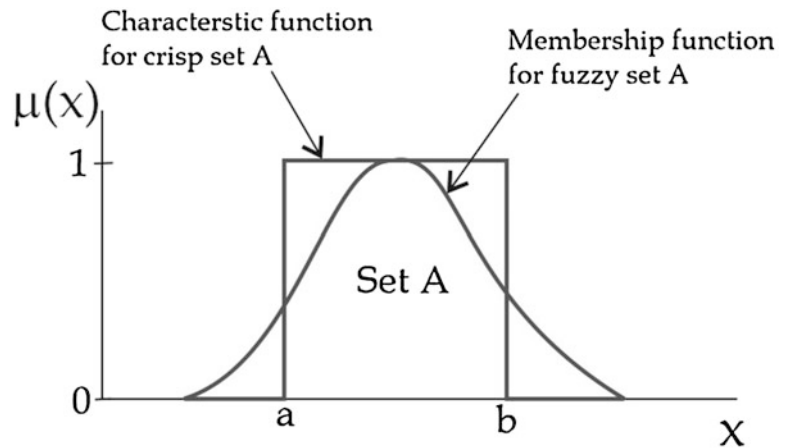
A fuzzy set is a class of object with a continuum of grades of membership, and such a set is characterized by a membership function which assigns to each object a grade of membership ranging between 0 and 1 (Zadeh 1965). Modeling has become an important research problem in Earth Sciences due to the development of data collection and processing methods in the last few decades. However, the models utilized in Earth Sciences are extremely complex, since they deal with natural events and processes. In addition, it is clear that expressing the elements employed in a model as crisp sets would not be adequate to eliminate the uncertainty. Instead, expressing each element with a degree of membership will decrease the uncertainty and yield to more representative models. Therefore, the use of membership functions in modeling problems in Earth Sciences has a significant contribution for addressing the natural phenomena or processes properly.

Mapping is an important concept in relating set-theoretic forms to function-theoretic representations of information (Ross 2004). The main characteristics of membership functions were described by Dombi (1990). According to the membership function properties described by Dombi (1990), all membership functions are continuous, and all membership functions map an interval $[a, b]$ to $[0, 1]$, $\mu[a, b] \rightarrow [0, 1]$. In addition, the membership functions are either increasing or decreasing monotonically or can be divided into a monotonically increasing or decreasing part (Dombi 1990). The monotonous membership functions on the whole interval are either convex functions or concave functions, or there exists a point c in the interval $[a, b]$ such that $[a, c]$ is convex and $[c, b]$ is concave (called S-shaped functions) (Dombi 1990). Another property of membership functions is that monotonically increasing functions have the property $u(a) = 0$, $u(b) = 1$, while monotonically decreasing functions have the property $u(a) = 1$, $u(b) = 0$ (Dombi 1990). Finally, the linear form or linearization of the membership function is important (Dombi 1990). Figure 2 shows the general properties of a trapezoidal membership function.

As can be seen from Fig. 2, the “core” of a membership function for a fuzzy set A is defined as the region of a universe that is characterized by a complete and full membership in the set A, and hence, the core comprises those elements x of the universe such that $\mu_A(x) = 1$. The “support” of a membership function for some fuzzy set A is defined as that region of the universe that is characterized by nonzero membership

Membership Functions,

Fig. 1 General view of a characteristic function for a crisp set and membership function for a fuzzy set. (Rearranged after Berkan and Trubatch (1997))



Membership Functions, Fig. 2 General properties of a membership function. (Rearranged after Ross (2004))

in the set A, and it means that the support comprises those elements x of the universe such that $\mu_A(x) > 0$ (Ross 2004).

Types of Membership Functions

In this entry, the most common forms of membership functions are presented. The most commonly used membership functions are singleton, Gaussian, generalized bell, sigmoidal, and triangular and trapezoidal membership functions (Rutkowski 2004). Their mathematical descriptions are provided below.

- (i) “Singleton membership function” is defined as unity at a particular point and zero everywhere else.

$$\mu(x) = \begin{cases} 1 & \text{for } x = \bar{x} \\ 0 & \text{for } x \neq \bar{x} \end{cases}$$

- (ii) “Gaussian membership function” is controlled by two parameters such as $\bar{x}$ (controls the center) and σ (controls the width).

$$\mu(x) = \exp\left(-\left(\frac{x - \bar{x}}{\sigma}\right)^2\right)$$

- (iii) “Generalized bell membership function” is governed by three parameters. These are a (responsible for its width), c (responsible for its center), and b (responsible for its slopes).

$$\mu(x) = \frac{1}{1 + \left|\frac{x-c}{a}\right|^{2b}}$$

- (iv) “Sigmoidal membership function” is controlled by two parameters such as a (controls for its slope) at the cross-over point $x = c$.

$$\mu(x) = \frac{1}{1 + \exp[-a(x - c)]}$$

- (v) “Triangular membership function” is completely controlled by three parameters such as a , b , and c ; and $a < b < c$.

$$\mu(x) = \max\left(\min\left(\frac{x-a}{b-a}, \frac{c-a}{c-b}\right), 0\right)$$

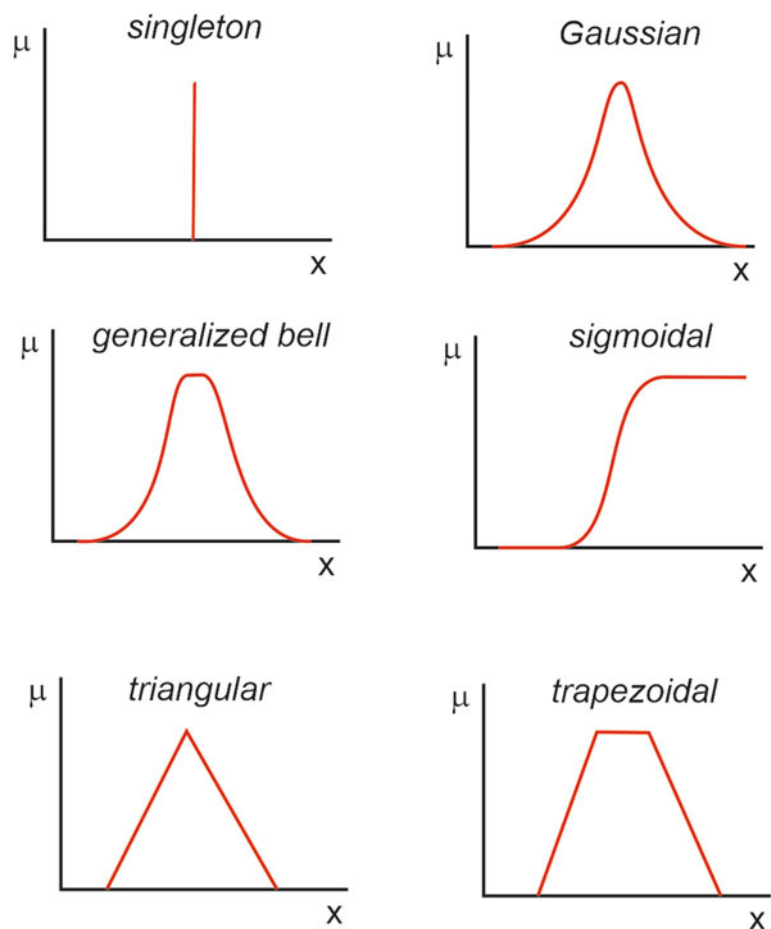
- (vi) “Trapezoidal membership function” is controlled by four parameters such as a , b , c , and d ; and $a < b < c < d$.

$$\mu(x) = \max\left(\min\left(\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}\right), 0\right)$$

The representative views of the membership functions mentioned above are given in Fig. 3.

Use of Membership Functions in Earth Sciences

When the materials are natural rock, the only known fact about the certainty is that this material can never be known

Membership Functions,**Fig. 3** Graphical form of the most commonly used membership functions

with certainty (Goodman 1995). This idea is valid for all natural materials, events, and processes. For this reason, classifying natural materials or explaining the natural processes involves uncertainty at different levels. Although the issue of uncertainty (historic, present, or future) is often mentioned, it is mostly as a side note, and it is still rarely used for quantitative and predictive purposes (Caers 2011). In the recent two decades, various fuzzy and neuro-fuzzy modeling studies for classification and prediction problems in Earth Sciences have been demonstrated, and all the models have membership functions. In fact, the features of membership functions fit the classification and prediction of events and processes in Earth Sciences. One of the pioneering fuzzy studies was performed by den Hartog et al. (1997) for the performance prediction of a rock-cutting trencher, in which trapezoidal membership functions were used. Subsequently, several studies in the international literature of the Earth Sciences have been published. Prediction of susceptible zones to landsliding (Akgun et al. 2012) and the prediction of rockburst (Adoko et al. 2013) can be listed as examples for such studies. Akgun et al. (2012) used trapezoidal and triangular membership functions, while Adoko et al. (2013) employed Gaussian

membership functions. The type of membership functions depends on the nature of problem, and the membership functions can be created based on expert opinion or by using data-driven approaches.

Summary or Conclusions

This entry presents the definition of membership functions with examples to their types commonly used in Earth Sciences. Due to the complexity of natural events or processes, it is not an easy task to classify natural materials and define their properties, and to predict natural events by employing crisp classes and sets. Due to this reason, the solutions proposed for such problems have uncertainty. To minimize the uncertainty, fuzzy sets including membership functions have been used in real-world problems. Use of expert opinion is one of the major advantages of membership functions because representative and reliable data cannot be obtained each time to solve the problems in Earth Sciences. In addition, the membership functions provide a flexible and transparent tool for many inference systems.

Cross-References

- [Fuzzy C-Means Clustering](#)
- [Fuzzy Set Theory in Geosciences](#)
- [Uncertainty Quantification](#)

Bibliography

- Adoko AC, Gokceoglu C, Wu L, Zuo QJ (2013) Knowledge-based and data-driven fuzzy modeling for rockburst prediction. *Int J Rock Mech Min Sci* 61:86–95
- Akgun A, Sezer EA, Nefeslioglu HA, Gokceoglu C, Pradhan B (2012) An easy-to-use MATLAB program (MamLand) for the assessment of landslide susceptibility using a Mamdani fuzzy algorithm. *Comput Geosci* 38:23–34
- Berkan RC, Trubatch SL (1997) *Fuzzy system design principles*. IEEE Press, New York, p 495
- Caers J (2011) *Modeling uncertainty in the earth sciences*. Wiley, Hoboken, p 229
- den Hartog MH, Babuska R, Deketh HJR, Alvarez Grima M, Verhoef PNW (1997) Knowledge-based fuzzy model for performance prediction of a rock-cutting trencher. *Int J Approx Reason* 16:43–66
- Dombi J (1990) Membership function as an evaluation. *Fuzzy Sets Syst* 35(1):1–21
- Goodman RE (1995) Block theory and its application. *Geotechnique* 45(3):383–423
- Ross TJ (2004) *Fuzzy logic with engineering applications*, 2nd edn. Wiley, Chichester, p 652
- Rutkowski L (2004) *Flexible neuro-fuzzy systems: structures, learning and performance evaluation*. Kluwer Academic, London, p 279
- Zadeh LA (1965) Fuzzy sets. *Inf Control* 8:338–353

Merriam, Daniel F.

D. Collins

Emeritus Associate Scientist, The University of Kansas,
Lawrence, KS, USA

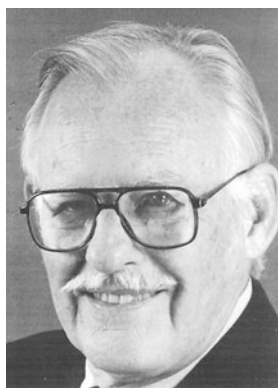


Fig. 1 Dan Merriam (© public domain)

Biography

Daniel (Dan) Francis Merriam (1927–2017) is known as one of the leading pioneers of mathematical geology and one of the foremost exponents of computer applications in the geosciences.

Following distinguished service with the US Navy in the Pacific in World War II, Dan's professional development began at the University of Kansas where he received an undergraduate degree in geology in 1949. Following 2 years of field work with Union Oil, Dan took a job with the Kansas Geological Survey (KGS) at the University of Kansas and enrolled in graduate school, receiving his MS degree in geology in 1953 and his PhD in 1961. He subsequently received the M.Sc. degree (1969) and D.Sc. degree (1975) from Leicester University, England.

Dan became the Survey's Chief of Geologic Research in 1963, where his promotion of quantitative analysis of geologic data began in earnest. With Dan as editor of the Survey's *Special Distribution Series* (1964–1966), numerous papers on quantitative analysis were published with computer programs in ALGOL, BALGOL, and FORTRAN II. Dan then served as editor of the KGS series *Computer Contributions*, an irregular publication of 50 issues (1966–1970) with computer programs for analysis of geologic data (available online at www.kgs.ku.edu). In 1969, while at the KGS, Dan also became the editor for the new Plenum Press series, *Computer Applications in the Earth Sciences*.

During the dramatic events of the Prague Spring in 1968, Dan participated as a founding member of the International Association for Mathematical Geology (IAMG - now International Association for Mathematical Geosciences). This led to Dan's creation of IAMG's international journals, *Mathematical Geology* (now *Mathematical Geosciences*) in 1969 and, with Elsevier, *Computers & Geosciences* in 1975, both edited by Dan for many years. He also edited another IAMG journal, *Natural Resources Research*, starting in 1999. In addition, Dan edited 11 books and published more than 100 papers during his career.

Dan was elected Secretary General of IAMG in 1972 and President in 1978. Among his many awards and distinctions, IAMG presented Dan the sixth William Christian Krumbein Medal (1981) for his scientific contributions and support of the geologic profession in general and the IAMG in particular (Davis 1982). The Geologic Society (London) presented him with the William Smith Medal in 1992.

In 1971, Dan moved to Syracuse University as Chairman of the Department of Geology. With Dan's leadership, Syracuse University became known as a leading center for quantitative studies in the geosciences. In 1981, Dan accepted a position as Chairman of the Department of Geology at Wichita State University in Kansas, where he stayed until 1991 when he rejoined the KGS in Lawrence, Kansas. Retiring in

1997, Dan remained active as an Emeritus Scientist for most of the last two decades of his life.

Bibliography

- Davis JC (1982) Association announcement. *Math Geol* 14(6):679–681
- Harbaugh JW, Merriam DF (1968) *Computer applications in stratigraphic analysis*. Wiley, New York
- Merriam DF (ed) (1966) *Computer applications in the earth sciences: Colloquium on classificational procedures*. Kansas Geological Survey, computer contribution series 7. Available online. <http://www.kgs.ku.edu/Publications/Bulletins/CC/7/CompContr7.pdf>
- Merriam DF (ed) (1967) *Computer applications in the earth sciences: Colloquium on time-series analysis*. Kansas Geological Survey, computer contribution series 18. Available online. <http://www.kgs.ku.edu/Publications/Bulletins/CC/18/CompContr18.pdf>
- Merriam DF (1999) Reminiscences of the editor of the Kansas geological survey computer contributions, 1966–1970 and a byte. *Comput Geosci* 25:321–334

Metadata

Simon J. D. Cox
CSIRO, Melbourne, VIC, Australia

Definition

Metadata is information about a data resource or transaction. Metadata provides contextual and summary information about a dataset or data item. Metadata may be useful in discovery, selection, management, access, and use of the described resources.

Metadata Standards

Metadata refers to information *about* other data, as distinct from information embodied *within* it. While descriptive records may be constructed for a wide variety of resources, including services, processes, and physical objects, those relating to **data and datasets** are often denoted **metadata**. Private metadata is created and stored as a routine process in data management systems. However, for external or public use, standard metadata structures and formats should be used.

Libraries

Libraries and archives had a long-standing practice of summary descriptions of the “information resources” which they manage, for example, in the form of index cards in a

catalogue. Metadata for books was standardized through Cataloguing-in-Publication records to be prepared by publishers, the MARC standards from Library of Congress, and in the context of library management software developed by organizations such as OCLC, alongside local and disciplinary standards. These practices had a strong influence on the development of “generic” metadata standards through the **Dublin Core** Metadata Initiative (Kunze and Baker 2007; DCMI Usage Board 2020), which emerged in the web era, but have been influential quite broadly and are used directly in many metadata records and systems. Dublin Core now includes nearly 80 terms, but the classic 15 terms, which are most widely used, indicate the overall scope.

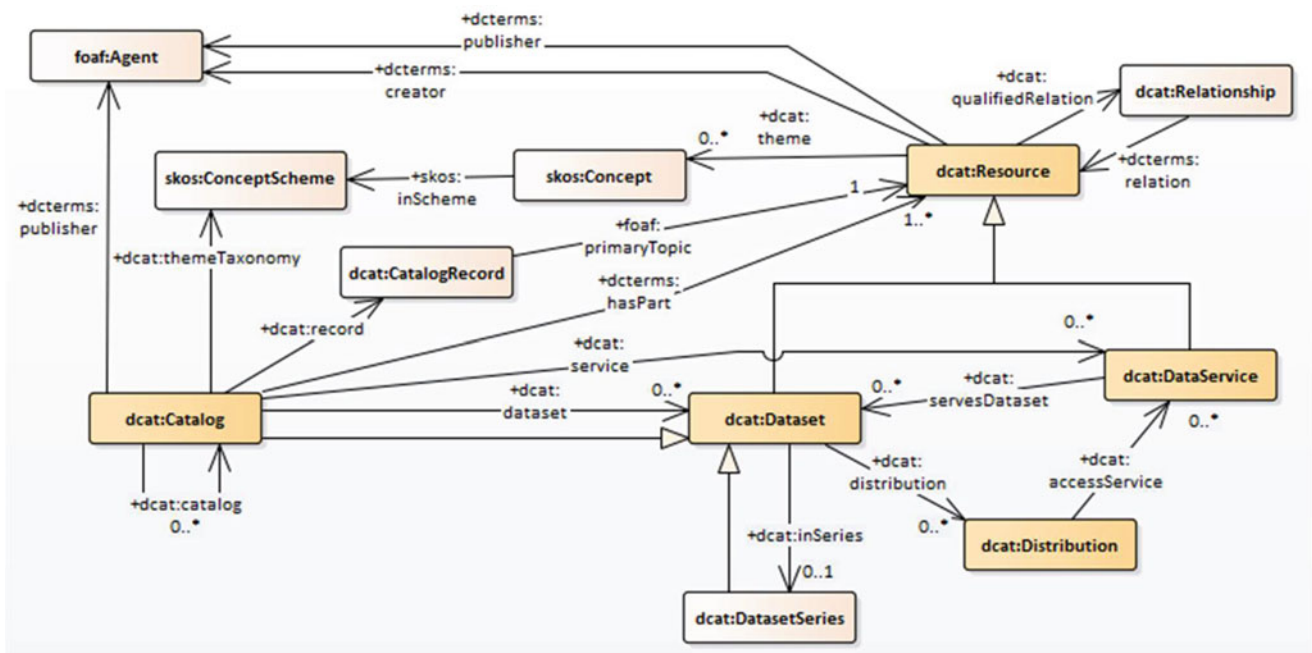
Dublin core element	Key application(s)
Title	Discovery and selection
Description	Discovery and selection
Subject	Discovery and selection
Coverage (<i>spatial and temporal scope</i>)	Discovery and selection
Language	Discovery and selection, use
Type	Discovery and selection, management
Identifier	Management, access
Creator, contributor, and publisher	Management, provenance
Date	Provenance, discovery, and selection
Format	Access, use
License, rights	Access, use
Relation	Discovery and selection, provenance

Web Pages

Metadata also appears in web pages, where it may be used for web search indexing. Several of the most prominent search systems sponsored the development of a more comprehensive set of metadata tags in **Schema.org** (Guha et al. 2015). This is maintained through an open community process. Schema.org now includes hundreds of elements, though only a subset of these are used for the general-purpose indexes. However, a profile known as **Science-on-Schema.org** has been developed by the earth and environmental science community and is published and harvested in some specialized indexes (Jones et al. 2021).

Geospatial

Much geoscience data is georeferenced. Geospatial metadata standards were originally developed somewhat



Metadata, Fig. 1 The DCAT metadata model, showing relationships between datasets, dataset-series, data-services, and catalog entries

independently, driven by requirements of map and image libraries, which were influenced by rapid growth of non-proprietary remote sensing data. Important general geospatial metadata standards were designed by the US Federal Geographic Data Committee (FGDC), and the Australia and New Zealand Land Information Council (ANZLIC). These were consolidated and standardized internationally as ISO 19115 (ISO/TC 211 2014, 2016, 2019). Data discovery systems based on these ISO standards are available from many statutory data providers in the Geosciences (e.g., geological surveys).

Semantic Web

The W3C's Resource Description Framework (RDF) provides a generic platform for metadata, which enables component vocabularies that have been defined for various concerns to be mixed and matched as required. For example, there are W3C-conformant vocabularies covering a number of specific concerns as follows:

Vocabulary	Concern	Reference
SKOS	Organizing keywords and classifications	Isaac and Summers (2009)
DQV	Data quality	Albertoni and Isaac (2016)
ODRL	Rights and licensing	Iannella et al. (2018)
PROV-O	Provenance	Lebo et al. (2013)

(continued)

Vocabulary	Concern	Reference
GeoSPARQL	Geospatial descriptions	Perry and Herring (2012)
OWL-Time	Temporal descriptions	Cox and Little (2017)

The Data Catalog Vocabulary (Albertoni et al. 2020) combines these with Dublin Core to cover most of the concerns of a comprehensive metadata record for datasets and services, including geospatial (see Fig. 1 for a summary). The threads are brought together in GeoDCAT (Perego and van Nuffelen 2020), which is effectively an RDF implementation of the ISO 19115 Geospatial metadata standard, accomplished by using elements from many of these vocabularies.

Keywords and Controlled Vocabularies

Metadata for geoscience data usually includes keywords and classifiers from domain specific vocabularies, such as minerals, lithology, the geological timescale, and taxonomic references, as well as lists of procedures, sensors, software, and algorithms that may have been used in the production of a dataset (i.e., data provenance), and therefore be important in discovery and selection of datasets by potential users. Geoscience vocabularies are provided by agencies such as geological surveys. Maximum interoperability and transparency are achieved by the use of FAIR vocabularies, where each term or vocabulary item is denoted by a unique, persistent web identifier or IRI (Wilkinson et al. 2016).

Summary

Metadata describes datasets in order to support a number of functions, including finding, selecting, accessing, using, and managing data. Metadata intended for public use should follow standards. Specialist geospatial metadata standards are used by many agencies and data repositories, while general metadata standards have been developed by the web community.

Cross-References

- [FAIR Data Principles](#)
- [Ontology](#)
- [Spatial Data Infrastructure and Generalization](#)

Bibliography

- Albertoni R, Isaac A (2016) Data on the web best practices: data quality vocabulary. W3C Working Group Note. World Wide Web Consortium. <https://www.w3.org/TR/vocab-dqv/>
- Albertoni R, Browning D, Cox SJD, Gonzalez-Beltran A, Perego A, Winstanley P (2020) Data Catalog Vocabulary (DCAT) – Version 2. W3C Recommendation. World Wide Web Consortium. <https://www.w3.org/TR/vocab-dcat/>
- Cox SJD, Little C (2017) Time ontology in OWL. W3C Recommendation. World Wide Web Consortium. <https://www.w3.org/TR/owl-time/>
- DCMI Usage Board (2020) DCMI metadata terms. DCMI Recommendation. 20 January 2020. <https://dublincore.org/specifications/dublin-core/dcmi-terms/>
- Guha RV, Brickley D, Macbeth S (2015) Schema.Org: evolution of structured data on the web. ACM Queue 13(9) 28 pp
- Iannella R, Steidl M, Myles S, Rodriguez-Doncel V (2018) ODRL vocabulary & expression 2.2. W3C Recommendation. World Wide Web Consortium. <https://www.w3.org/TR/odrl-vocab/>
- Isaac, Antoine, and Ed Summers. 2009. SKOS simple knowledge organization system primer. W3C Working Group Note. World Wide Web Consortium. <https://www.w3.org/TR/skos-primer/>
- ISO/TC 211 (2014) ISO 19115-1:2014 geographic information – metadata – Part 1: Fundamentals. International Standard ISO 19115-1. Geographic Information. International Organization for Standardization, Geneva. <https://www.iso.org/standard/53798.html>
- ISO/TC 211 (2016) ISO/TS 19115-3:2016 geographic information – metadata – Part 3: XML schema implementation for fundamental concepts. International Standard, Geneva. <https://www.iso.org/standard/32579.html>
- ISO/TC 211 (2019) ISO 19115-2:2019 geographic information – metadata – Part 2: Extensions for acquisition and processing. International Standard, Geneva. <https://www.iso.org/standard/67039.html>
- Jones M, Richard SM, Vieglais D, Shepherd A, Duerr R, Fils D, McGibbney L (2021) Science-on-Schema.Org v1.2.0. Zenodo. <https://doi.org/10.5281/zenodo.4477164>
- Kunze J, Baker T (2007) The Dublin core metadata element set. IETF RFC 5013. Internet Engineering Task Force (IETF). <https://datatracker.ietf.org/doc/html/rfc5013>

- Lebo T, Sahoo S, McGuinness DL (2013) PROV-O: the PROVontology. W3C Recommendation. World Wide Web Consortium, Cambridge, MA. <http://www.w3.org/TR/prov-o/>
- Perego A, van Nuffelen B (2020) GeoDCAT-AP – Version 2.0.0. Recommendation. European Commission. <https://semiceu.github.io/GeoDCAT-AP/releases/2.0.0/>
- Perry M, Herring J (2012) OGC GeoSPARQL – a geographic query language for RDF data. OGC Standard 11-052r4. OGC 11-052r4. Open Geospatial Consortium, Wayland. <https://portal.opengeospatial.org/files/47664>
- Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N et al (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3(1):160018. <https://doi.org/10.1038/sdata.2016.18>

Mine Planning

N. Caciagli
Metals Exploration, BHP Ltd, Toronto, ON, Canada
Barrick Gold Corp, Toronto, ON, USA

Synonyms

Mine design

Definition

Mine planning encompasses the design and implementation of safe, sustainable, and profitable plans for mining operations, with the aim of maximizing value.

Introduction

The extraction of natural resources is fundamental to modern society. Almost everything we touch and use has some component that was mined. However, the ore bodies that are being discovered are increasingly lower grade and more difficult to mine. Additionally, the social and environmental impacts to mining need to be addressed and in many jurisdictions are becoming more regulated. As a result, the extraction and processing methods must become more intricate and the planning process and models more detailed and robust.

Inputs into Mine Planning

Inputs into the mine planning process include:

- Orebody characteristics contained within geological, geo-technical, and geometallurgical models
- Mining method and layouts, taking into account the local mining and environmental regulations regarding safety, noise, dust, and other impacts to workers and the environment
- Technical inputs such as cut-off grades and requirements for blending or stockpiling
- Operational inputs such as ore loss and dilution factors, mining, and processing costs
- Processing plant capacity and feed requirements
- Final product requirements and specifications (e.g., penalty element content)
- Commodity value based on both short- and long-range forecasting

Much of the mine planning process aims to generate to a set of high-level options around these inputs, building several hypothetical models or plans to be tested to determine which plan yields the highest value (Whittle 2011). Any one of the inputs or decisions can affect all the others; for example, constraints around the processing method will affect the optimum cut-off grades, changes in cut-off grade will affect the mining schedules, and changes in schedules affect which pit shells or mine levels should be selected as mine phases. A change in the commodity price will affect everything, right back to the mine design (Whittle 2010). Additionally, mine planning has the added dimension of time and needs to consider when and how each phase or structure is implemented.

Mine planning examines specific problems at different stages of the mining value chain and can be broken down into three views to better address these:

- Long-term mine planning
- Medium-term mine planning
- Short-term mine planning

Long-term mine planning aims to define the “best” mine plan over the full life of mine, subject to the constraints imposed by physical and geological conditions, as well as corporate and regulatory policies. The term “best” is typically defined as maximizing the monetary value and guarantying a safe operation but also includes quality characteristics of the product extracted and delivered (Goodfellow and Dimitrakopoulos 2017). Long-term mine planning also considers capital investment decisions, sovereign risk issues as well as market demands for the material to be mined.

Medium-term mine planning focuses on issues such as mobile equipment capital investment, strategic positioning for optimization of product quality, and processing plant

capacity. This also considers stope access or drift development for underground mines or pit phases and pushbacks for open pit mines.

Short-term mine planning focuses on production scheduling. Where to mine next week, is it accessible? Is there enough blasted inventory to feed the crusher and the plant?

Methodology and Tools

Most commercial open pit mine planning tools use a Nested-Pit methodology based on the Lerchs and Grossmann algorithm (Lerchs and Grossmann 1965). The Lerchs-Grossmann algorithm determines the economic three-dimensional shape of the pit considering block grades, pit slopes, costs, recoveries, and metal prices and is based on the valuation of each block in the model. Also known as the Nested Pit method this procedure breaks down the problem into two parts: it determines a set of push-packs (or phases) to be used as guides to construct a production schedule and then schedules the blocks in each push-back, bench by bench.

Underground mine planning has no equivalent methodology for determining an “ultimate pit limit,” rather the models or plans to be tested and assessed focus on the on the choice of “best” mining method (e.g., stoping or caving methods), development and layout of infrastructure, design of stope boundaries or envelopes, and production scheduling. For underground mining there are a few optimization algorithms developed in the 1990s and early 2000s. Alford (1995) developed the heuristic floating stope algorithm to define the optimum stope envelopes, analogous to the floating cone algorithm used in open pit optimization. Others have presented variants on the heuristic method first proposed by Alford utilizing a maximum value neighborhood (MVN) heuristic algorithm, integer programming techniques, or 3D modeling (Musingwini 2016 and references therein). Other work has focused on selecting optimal mining methods or planning mine development and layout using neural networks or genetic algorithms (Newman 2010 and references therein). However, a method to integrate the layout and development planning, sizing of stope envelopes, production scheduling, equipment selection, and utilization with the aim to enable just-in-time development analogous to open-pit planning remains elusive.

It is understood that the uncertainties associated with the input parameters, especially metal grades or commodity prices, can lead to large deviations from expected budgets. However, accounting for these in mine planning remains an active area of research and development (Musingwini 2016 and references therein, Newman et al. 2010 and references therein, Davis and Newman 2008 CRC mining conference).

Sources of Uncertainty

Orebody characterization attempts to quantify the sources of uncertainty that contribute most to the differences between planned and operational outcomes. Geological uncertainty represents the level of confidence in the mineralogical characterization, grade, continuity, as well as of the extent and position of geological units. Resource model estimates are continuous interpolations of discretely obtained data, but the models do not capture the real variability of the deposit and tend to reduce the variability of the deposit attributes (Jélvez et al. 2020; Goodfellow and Dimitrakopoulos 2017) which in turn leads to cascading impacts within the mine planning outputs.

Accounting for these uncertainties allows for the development of models that are more realistic and creates the opportunity to develop options within the mine plan (Davis and Newman 2008 CRC mining conference).

Jélvez et al. (2020) looked at incorporating uncertainty into the mine planning process by representing geological uncertainty as a series of stochastic simulations of the block model. This methodology developed strategies for the definition of the ultimate pit limit, pushback selection, and production scheduling.

Goodfellow and Dimitrakopoulos (2017) implemented a stochastic approach to optimizing the entire mining complex incorporating the variability of materials and metal content into the simultaneous production scheduling, ore dispatching, and processing (stockpiling or blending) decision variables.

In both case studies a stochastic design is better able to meet the production targets and simultaneously reduce the risk in production profiles and increase the expected NPV, highlighting the importance of characterizing uncertainty and incorporating it into mine planning work.

Conclusions

Mining has become more challenging given the nature of the orebodies being discovered, health, safety, and environmental requirements and the sophisticated extraction and processing methods available. This has demanded the mine planning process solve increasingly more complex problems and necessitating more intricate and integrated solutions.

Cross-References

- [Geotechnics](#)
- [Spatial Statistics](#)
- [Three-Dimensional Geologic Modeling](#)

Bibliography

- Goodfellow R, Dimitrakopoulos R (2017) Simultaneous stochastic optimization of mining complexes and mineral value chains. *Math Geosci* 49:341–360
- Goycoolea M, Espinoza D, Moreno E, Rivera O (2015) Comparing new and traditional methodologies for production scheduling in open pit mining. Application of computers and operations research in the mineral industry. In: *Proceedings of the 37th international symposium: APCOM 2015*
- Jélvez E, Morales N, Ortiz JM (2020) Impact of geological uncertainty at different stages of the open-pit mine production planning process. In: *Proceedings of the 28th international symposium on mine planning and equipment selection 2019*, pp 83–91
- Lerchs H, Grossmann I (1965) Optimal design of open-pit mines. *Trans CIM* 68:17–24
- Musingwini C (2016) Presidential address: optimization in underground mine planning- developments and opportunities. *J S Afr Inst Min Metall* 116(9):1–12
- Newman AM, Rubio E, Caro R, Weintraub A, Eurek K (2010) A review of operations research in mine planning. *Interfaces* 40(3):222–245
- Whittle J (1989) *The facts and fallacies of open pit optimization*. Whittle Programming, North Balwyn
- Whittle G (2010) *Enterprise optimisation. Mine planning and equipment selection, 2010*. The Australasian Institute of Mining and Metallurgy publication series no 11/2010. The Australasian Institute of Mining and Metallurgy, Melbourne
- Whittle D (2011) *Open-pit planning and design*. In: Darling P (ed) *Mining engineering handbook*, 3rd edn. Society for Mining, Metallurgy, and Exploration, Englewood

Mineral Prospectivity Analysis

Emmanuel John M. Carranza

Department of Geology, Faculty of Natural and Agricultural Sciences, University of the Free State, Bloemfontein, Republic of South Africa

Definition

The phrase “mineral prospectivity” denotes the probability or chance that mineral deposits of a certain class can be discovered in an area because the geology of this area is permissive or favorable to the formation of such class of mineral deposits. The phrase is synonymous to the phrase “mineral favorability” or “mineral potential,” either of which denotes the possibility or likelihood that a certain class of mineral deposits exist in an area because the geology of this area is permissive or favorable to the formation of such class of mineral deposits. However, the phrase “mineral prospectivity” is used in this encyclopedia because the term “prospectivity” relates to the search, prospecting, or exploration of mineral deposits of economic importance. Mineral prospectivity analysis, therefore, aims to predict where undiscovered mineral deposits of a certain class exist in order guide mineral exploration.

Introduction

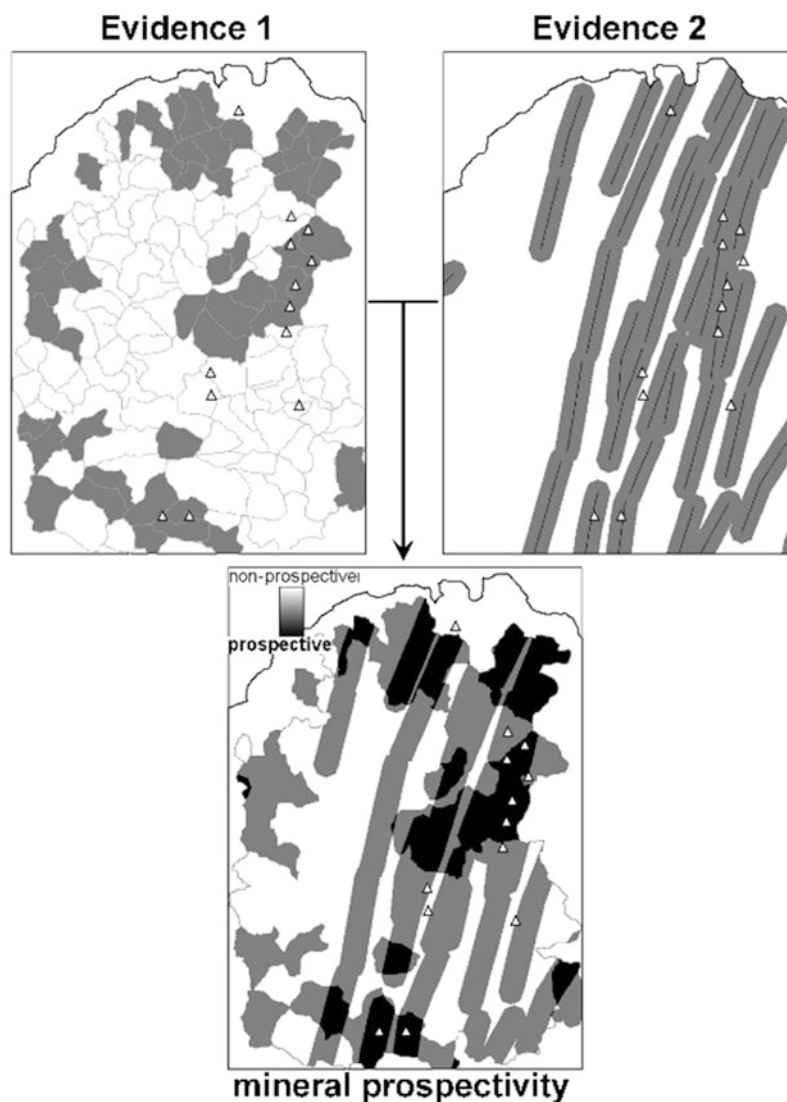
Because the occurrence of any mineral deposit is revealed by the existence of certain pieces of geological evidence (for example, abnormal concentrations of certain metals in certain earth materials), mineral prospectivity is associated to the amount of geological evidence. Therefore, as shown in Fig. 1, mineral prospectivity analysis dwells on two assumptions. First, an area is prospective if it is typified by pieces or layers of geological evidence that are the same or similar to those in areas where mineral deposits of a certain class exist. Second, if in one area there exist more pieces or layers of geological evidence than in another area in a geologically permissive or favorable terrane, then the degree of mineral prospectivity in the first area is higher than in the second area.

Mineral prospectivity analysis can be linked to every or any stage of mineral exploration, from broad regional-scale to

restricted local-scale, whereby it involves analysis and synthesis of pieces or layers of evidence, in map form, obtained from spatial data of different geoscientific types (e.g., geophysical, geological, and geochemical). The purpose of that is to outline exploration targets (Fig. 1) and to prioritize such targets for exploration of yet to be discovered mineral deposits of the class of interest. On a regional-scale, mineral prospectivity analysis endeavors to outline exploration targets within continental-scale geologically permissive regions; on district- to local-scales, mineral prospectivity analysis seeks to delineate exploration targets in more detail. These imply that the spatial resolution, detail, accuracy, and information content of individual sets of geoscientific data, as well as understanding of mineral deposit occurrence, employed in mineral prospectivity analysis for exploration targeting must increase progressively from regional- to local-scale.

Mineral Prospectivity Analysis,

Fig. 1 Level of mineral prospectivity is correlated with level of spatial coincidence of relevant layers of evidence. Areas with the same or similar level of mineral prospectivity as existing mineral deposits (triangles) are regarded as exploration targets



Conceptual Modeling of Mineral Prospectivity

The analysis of mineral prospectivity should relate to mineral deposits of a single class or related mineral deposit belonging to a single mineral system. Therefore, a map of mineral prospectivity for porphyry Cu deposits is not appropriate for guiding exploration for diamond deposits, or the other way around. At every scale of exploration targeting, however, mineral prospectivity analysis adheres to definite steps commencing with the conceptualization of prospectivity for mineral deposit class or mineral system of interest (Fig. 2). Theoretical and spatial relationships among various features, which are factors or controls on how and where mineral deposit of a certain class forms, comprise the mineral prospectivity conceptual model. This model prescribes in expressions and/or illustrations of the theoretical or hypothetical interplays among different geological processes in the measure of how and particularly where mineral deposits of the class of interest possibly form.

Conceptualization of prospectivity for mineral deposits in an area needs a thorough review of literature on characteristics and processes of formation of mineral deposits of the class of interest, such as described in mineral deposit models (e.g., Cox and Singer 1986). A robust mineral prospectivity conceptual model is one that also adheres particularly to the mineral systems approach to exploration targeting. This approach focuses on three main themes of geological elements (or processes) that are critical to the formation of mineral deposits (Walshe et al. 2005; McCuaig et al. 2010): (a) source of metals; (b) pathways for metal-bearing fluids; and (c) depositional traps. Analysis of the spatial distribution of mineral deposits of the class of interest in an area and analysis of the spatial relationships of such mineral deposits with certain geological features in the same area can provide supplementary knowledge to the conceptualization of prospectivity for mineral deposits (Carranza 2009).

The spatial (geophysical, geological, and geochemical) features of known mineral deposits of the class of interest, whether they exist in the area currently being explored or in

other areas with similar geological setting, comprise the criteria for recognition or modeling of prospectivity. These criteria and the mineral prospectivity conceptual model instruct the structure of mineral prospectivity analysis with regard to the choice of appropriate of the following:

- Sets of geoscientific spatial data to be employed.
- Evidential elements to be mapped from each set of geoscientific spatial data.
- Techniques for converting each map of evidential elements into a map of prospectivity recognition criterion.
- Techniques for weighting of each class or range of values in a map of prospectivity recognition criterion to generate a predictor map.
- Techniques of combining predictor maps to generate a mineral prospectivity map.

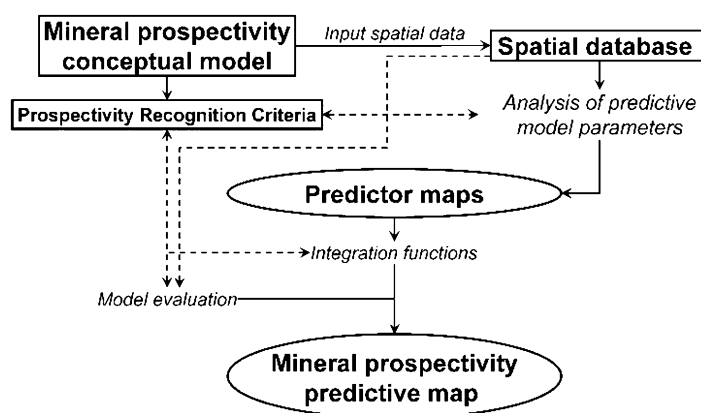
The items (b), (c), and (d) pertain to the analysis of mineral prospectivity parameters.

Preparation of Evidence Maps

A map that portrays a prospectivity recognition criterion is called an evidential map, or in Geographic Information System parlance, an evidential layer. Techniques for mapping evidential elements expressive of a criterion of mineral prospectivity are explicit to types of geoscientific spatial data (i.e., geophysical, geological, and geochemical). Concepts of mapping of anomalies from geochemical data for mineral exploration are discussed in the chapter for Exploration Geochemistry. Geological elements of specific classes of mineral deposits, such as hydrothermal alterations, are detectable and mappable by remote sensing (see chapters on Hyperspectral Remote Sensing, Mathematical Methods in Remote Sensing Data and Image Analysis, Remote Sensing). Geological elements such as structures or rock units considered as proxies of controls of mineral deposit formation can be extracted from

Mineral Prospectivity Analysis,

Fig. 2 Components of mineral prospectivity analysis. Solid arrows portray input-output workflow. Dashed arrows denote support among or feedback between components



existing geological maps or mapped from images of suitable geophysical data (e.g., Kearey et al. 2002).

Some geoscientific spatial data or some types of mapped evidential elements must be manipulated or transformed in order to portray them as evidential maps. The choice of technique to manipulate or transform specific relevant data, to generate an evidential map, depends on which criterion for recognition of prospectivity is to be represented. For example, representation of fluid pathways as a control on mineral deposit formation first needs extraction of certain structures from a spatial database, then generation of a map of distances to such structures, and finally classification of distances into ranges of proximity. This manipulation or transformation is generally known as buffering (see chapter on Geographic Information System). Similarly, representation of chemical traps as a control of mineral deposit formation may first need appropriate interpolation of metal concentration data at sample locations and then classification of the interpolated metal data into different levels of presence of chemical traps. For concepts of methods of interpolation, see chapters on Interpolation, Inverse Weighted Distance, or Kriging. Classification is also a typical operation in a Geographic Information System. The purpose of manipulation or transformation of a map of evidential elements or image of relevant spatial data to generate an evidential map is to represent the degree of presence of evidential elements per location.

Generation and Integration of Predictor Maps

Techniques for the assignment of weights to each class or range of values in an evidential map, in order to generate a predictor map, involve either data- or knowledge-driven analysis of their spatial relationships with discovered mineral deposits of the class of interest (Carranza 2008). On the one hand, data-driven techniques quantify spatial relationships between a mineral deposit location map and each class or range of values in an evidential map; quantified degrees of spatial relationship translate to weights of each class or range of values in an evidential map. On the other hand, knowledge-driven techniques make use of expert opinions to assign weights to each class or range of values in an evidential map. Not every technique for data-driven quantification of spatial relationships between a mineral deposit location map and an evidential map steers can be used for the generation and then combination of predictor maps to obtain a predictive mineral prospectivity map. Data-driven techniques that only quantify spatial relationships between classes or range of values in an evidential map are those that are useful in providing supplementary knowledge to the conceptualization of prospectivity for mineral deposits as mentioned above. In contrast, every technique for knowledge-driven predictor map generation can be used straight away for predictor map

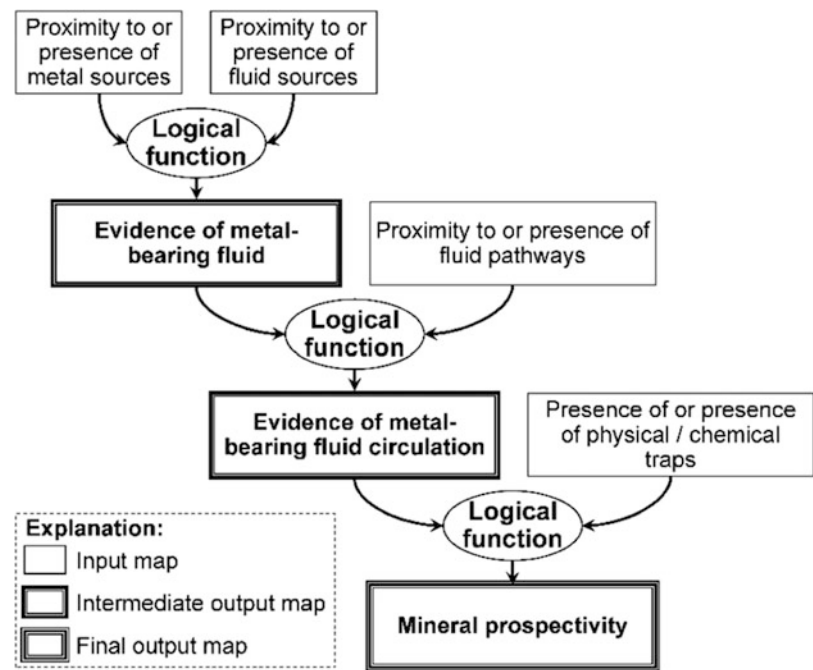
combination to generate a predictive mineral prospectivity map.

Therefore, mineral prospectivity analysis is either data- or knowledge-driven. The main distinction between data- and knowledge-driven mineral prospectivity analyses lies in the generation of predictor maps. Data-driven mineral prospectivity analysis uses known mineral deposits to assign (i.e., quantify) weights to each class or range of values in an evidential map. Knowledge-driven mineral prospectivity analysis uses expert opinion to assign (i.e., qualify) weights to each class or range of values in an evidential map. Data-driven mineral prospectivity analysis is suitable in moderately to well-explored geologically permissive areas (also known as “brownfields”) where several mineral deposits of the class of interest exist. Knowledge-driven mineral prospectivity analysis is suitable in less-explored geologically permissive areas (also known as “greenfields”) where no or very few mineral deposits of the class of interest are known to exist. In either case, a predictive mineral prospectivity map is generated by integrating at least two predictor maps (Fig. 1) using a mathematical (e.g., probability, likelihood, favorability, or belief) function that appropriately represents or models the interplays among geological processes, represented by each predictor map that leads to mineral deposit occurrence or formation, thus: *predictive mineral prospectivity map* = $f(\text{predictor maps})$. The mathematical function f used typically in mineral prospectivity analysis conveys empirical interactions among predictor maps (as proxies of factors or controls) of mineral deposit occurrence. On the one hand, knowledge-driven techniques of mineral prospectivity analysis commonly make use of logical functions (e.g., AND and/or OR operators; see Fuzzy Set Theory) for successive combination of predictor maps guided by an inference system (Fig. 3). On the other hand, data-driven techniques of mineral prospectivity analysis commonly make use of mathematical functions for instantaneous combination of predictor maps irrespective of knowledge on the interplays of geological factors or controls of mineral deposit occurrence. A few data-driven techniques use functions signifying logical operations (e.g., Dempster’s rule of integrating belief functions; see Carranza et al. 2008) successive combination of predictor maps guided by an inference system. Similarly, a few knowledge-driven techniques use mathematical (mainly arithmetic or logical) functions for instantaneous combination of predictor maps.

For mineral prospectivity analysis, the most commonly used data-driven technique is weight-of-evidence analysis, and the most commonly used knowledge-driven technique involves the application of the fuzzy set theory (for details, see Bonham-Carter 1994). Most published researches on mineral prospectivity analysis can be found in the following journal: Natural Resources Research, Ore Geology Reviews, Mathematical Geosciences and Computers & Geosciences.

Mineral Prospectivity Analysis,

Fig. 3 Example of inference system using suitable logical functions (e.g., AND and/or OR operators, etc.) for successive combination of predictor maps through knowledge-driven mineral prospectivity analysis. An inference system portrays an analyst's understanding or an expert's opinion of the interplays of geological factors or controls of mineral deposit occurrence



Evaluation of Results of Mineral Prospectivity Analysis

Every technique of mineral prospectivity analysis produces *systemic* (or procedure-related) errors vis-à-vis the interplays of geological factors or controls of mineral deposit occurrence. Moreover, each set of geoscientific spatial data or each evidential map employed in mineral prospectivity analysis usually carries *parametric* (or data-related) errors vis-à-vis the spatial distribution of mineral deposits. Because these types of errors are unavoidable in mineral prospectivity analysis and because they spread from input data to evidential map to predictor map to final output mineral prospectivity map, it is necessary to employ strategies for predictive map evaluation and, if needed, to reduce error in each step of mineral prospectivity analysis. The ideal strategy to evaluate the predictive capacity of a mineral prospectivity map is, of course, to use it as basis for guiding mineral exploration and hope that it leads to mineral deposit discovery. However, this is not a practical strategy because it usually takes several years before a mineral deposit is discovered through grassroots exploration.

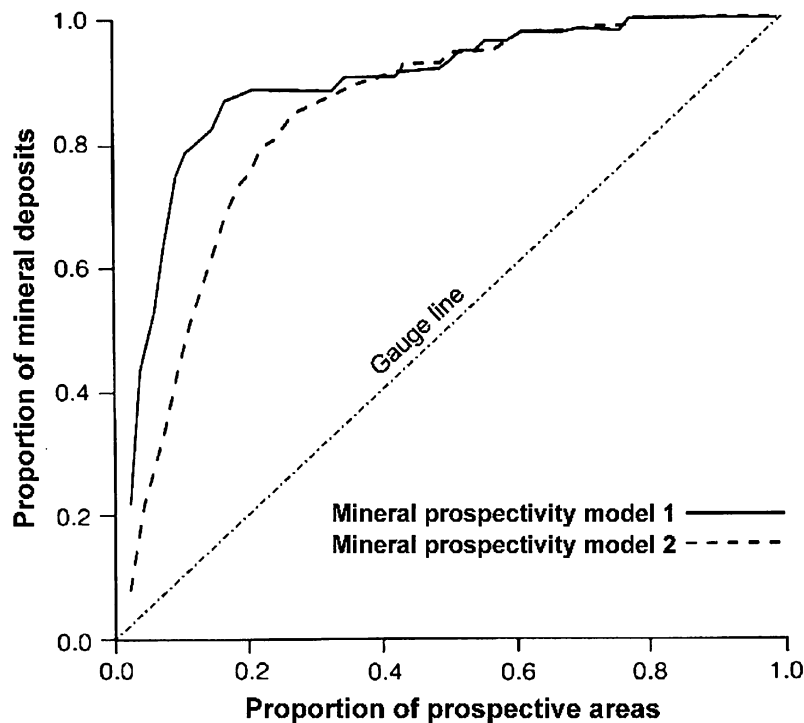
A practical way to evaluate results of mineral prospectivity analysis is to answer the following question. For at least two predictive mineral prospectivity maps (say, generated by different techniques or generated by the same techniques but using different parameters), which map (a) delineates the smallest proportion of prospective areas but (b) contains the largest proportion of existing mineral deposits of the class of interest? To answer and visualize the answer to this question, one must create a so-called *area-occurrence plot* per mineral prospectivity map (Fig. 4) (Agterberg and Bonham-Carter

2005). This question is relevant to results of either data- or knowledge-driven mineral prospectivity analysis.

For knowledge-driven mineral prospectivity analysis, the area-occurrence plot is appropriately called a *prediction-rate* plot because mineral deposits are not used to create predictor maps (i.e., assign weights to each class or range of values in an evidential map). Therefore, according to Fig. 4, the map based on mineral prospectivity model 1 has greater predictive capacity than the map based on mineral prospectivity model 2. For example, considering predicted prospective areas occupying 10% of a region, the predictive capacity of mineral prospectivity model 1 is about 75% whereas that of mineral prospectivity model 2 is just about 45%.

For data-driven mineral prospectivity analysis in which all existing mineral deposits are used to create predictor maps, the area-occurrence plot is appropriately a *success-rate* plot. That is, the success-rate plot portrays only the goodness-of-fit between evidential maps and the map of mineral deposits locations used for calculating the weights for each class or range of values in each evidential map. Therefore, according to Fig. 4, the map based on mineral prospectivity model 1 has greater goodness-of-fit with the mineral deposits based on which it was created compared to the map based on mineral prospectivity model 2.

To determine the predictive capacity of a data-driven mineral prospectivity model, one must answer the following question: Assume that, for a study area, we split the existing mineral deposits of the class of interest into two sets. The first set, called the *training set*, is used in mineral prospectivity analysis. What proportion of the second set, called the *validation set* (i.e., mineral deposits presumed to be undiscovered), coincides spatially with high prospectivity



Mineral Prospectivity Analysis, Fig. 4 Area-occurrence plots for evaluation of mineral prospectivity models. The procedure for creating such plots involves a series of delineation of prospective areas in a mineral prospectivity map using decreasing threshold values at narrow, say 5-percentile, intervals from the highest to the lowest prospectivity values, which yield, respectively, the lowest and the highest proportions of prospective areas. Then, the next step is to determine the cumulative increasing proportions of existing mineral deposits contained in

respective cumulative increasing prospective areas defined by decreasing prospectivity values. Finally, proportions of contained mineral deposits are plotted on the y-axis, and proportions of prospective areas are plotted on the x-axis. The gauge line represents a random spatial relationship between mineral deposits and prospectivity values in a mineral prospectivity map. For a mineral prospectivity map, the higher above its occurrence-area plot is relative to the gauge line, the stronger is its predictive capacity

values derived from the first set? This question is the principle of blind testing of data-driven mineral prospectivity analysis (cf. Fabbri and Chung 2008). It applies to evaluation of mineral prospectivity analyses using one technique or different techniques. Similarly, to answer and visualize the answer to this question, one must create a so-called *area-occurrence plot* per mineral prospectivity map. Moreover, the role of the first and second sets of mineral deposits can be switched to compare the predictive capacities of two mineral prospectivity models using the same technique of mineral prospectivity analysis. In this case, the area-occurrence plot is appropriately called a *prediction-rate* plot; therefore, according to Fig. 4, the map based on mineral prospectivity model 1 has greater predictive capacity than the map based on mineral prospectivity model 2.

Conclusions

A predictive mineral prospectivity map is generally an empirical model, portraying the chance of finding undiscovered mineral deposits of the class of interest. This is a critical

spatial information for management of mineral exploration. The routines in mineral prospectivity analysis can be tedious and complex, but nowadays these are facilitated with use of a Geographic Information System. However, mineral prospectivity analysis is a typical “garbage in, garbage out” activity because, as shown in Fig. 1, if the mineral prospectivity conceptual model is wrong, then the output mineral prospectivity map is wrong. Therefore, knowledge of how and where mineral deposits form is the most critical element of mineral prospectivity analysis. It is judicious to evaluate performance of mineral prospectivity maps before they are used to guide mineral exploration.

Cross-References

- [Exploration Geochemistry](#)
- [Favorability Analysis](#)
- [Fuzzy Set Theory in Geosciences](#)
- [Hyperspectral Remote Sensing](#)
- [Interpolation](#)
- [Inverse Distance Weight](#)

- [Kriging](#)
- [Object-Based Image Analysis](#)
- [Remote Sensing](#)
- [Spatial Analysis](#)

Bibliography

- Agterberg FP, Bonham-Carter GF (2005) Measuring the performance of mineral-potential maps. *Nat Resour Res* 14(1):1–17
- Bonham-Carter GF (1994) *Geographic Information Systems for Geoscientists: Modelling with GIS*. Pergamon, Oxford
- Carranza EJM (2008) Geochemical anomaly and mineral prospectivity mapping in GIS. *Handbook of exploration and environmental geochemistry*, vol 11. Elsevier, Amsterdam
- Carranza EJM (2009) Controls on mineral deposit occurrence inferred from analysis of their spatial pattern and spatial association with geological features. *Ore Geol Rev* 35:383–400
- Carranza EJM, van Ruitenbeek FJA, Hecker C, van der Meijde M, van der Meer FD (2008) Knowledge-guided data-driven evidential belief modeling of mineral prospectivity in Cabo de Gata, SE Spain. *Int J Appl Earth Obs Geoinf* 10(3):374–387
- Cox DP, Singer DA (eds) (1986) *Mineral deposit models*. U.S. geological survey bulletin 1693. United States Government Printing Office, Washington, DC
- Fabbri AG, Chung CJ (2008) On blind tests and spatial prediction models. *Nat Resour Res* 17(2):107–118
- Kearey P, Brooks M, Hill I (2002) *An introduction to geophysical exploration*, 3rd edn. Blackwell Scientific Publications, Oxford
- McCuaig TC, Beresford S, Hronsky J (2010) Translating the mineral systems approach into an effective exploration targeting system. *Ore Geol Rev* 38(3):128–138
- Walshe JL, Cooke DR, Neumayr P (2005) Five questions for fun and profit: a mineral systems perspective on metallogenic epochs, provinces and magmatic hydrothermal Cu and Au deposits. In: Mao J, Bierlein FP (eds) *Mineral deposit research: Meeting the global challenge*, vol 1. Springer, Heidelberg, pp 477–480

Minimum Entropy Deconvolution

Jaya Sreevalsan-Nair
Graphics-Visualization-Computing Lab, International
Institute of Information Technology Bangalore, Electronic
City, Bangalore, Karnataka, India

Definition

In reflection seismology, the reflected seismic waves are used to study the geophysical characteristics of the earth subsurface. Thus, in a seismogram model, deconvolution is routinely used to extract the earth reflectivity signal from a noisy seismic trace. The conventional method is used to whiten the spectra of the earth reflectivity signal, i.e., to assume that the signal has a constant value as the Fourier

transform. This, however, reconstructs only a part of the signal, as the assumption also implies that the signal is a Dirac delta function δ , i.e., a spiking function.

To adapt this method to include the practical characteristics of the signal of being noisy, band-limited, etc., minimum entropy deconvolution (MED) is used, where it finds the smallest number of the large spikes that are consistent with the data using minimum Wiggins' entropy measure (Wiggins 1978). MED was proposed by Wiggins in 1978, and MED has been inspired by the factor analysis techniques.

Overview

The seismic trace, x_t , that is measured has been observed to be a convolution between the earth reflectivity, y_t and a wavelet function, w_t . To obtain y_t , a straightforward signal processing process is to perform a deconvolution process, say using a linear scheme. However, in real life, the signal x_t tends to be contaminated with noisy signal, n_t . Thus, the normal incidence seismogram model is given by: $x_t = w_t * y_t + n_t$.

A filter f_t , which upon convolution with the wavelet function gives the Dirac delta function δ_t , is desired. Thus, $f_t * w_t = \delta_t$. However, the estimated filter $\hat{f}_t$ gives an averaging function, a_t , i.e., $\hat{f}_t * w_t = a_t$. Thus, to get the estimated reflectivity signal

$$\hat{y}_t = a_t * y_t + \hat{f}_t * n_t = y_t + (a_t - \delta_t) * y_t + \hat{f}_t * n_t.$$

The desirable solution is that the a_t is close to δ_t and n_t is relatively low. Given the band-pass nature of the seismic signals, only a part of the reflectivity signal is retrieved, i.e., the band-limited signal after removing the wavelet, $\hat{y}_t = y_t * a_t$. Thus, $\hat{y}_t$ and y_t are referred to as the band-limited and the full-band reflectivity signals. Estimating y_t from $\hat{y}_t$ is known to be a non-unique linear inverse problem. The Fourier transform gives $\hat{Y}_\omega = A_\omega \cdot Y_\omega$, for frequency ω . There are an infinite number of solutions for y_t that satisfies the Fourier transform equation, especially when $A_\omega \neq 0$.

MED has been proposed as a solution to reduce the non-uniqueness of the solution (Wiggins 1978) by finding those solutions which provide “minimum structure” to the reflectivity signal. The “minimum structure” is represented by “minimum entropy,” where entropy is proportional to the energy dispersal of the seismic signal. Entropy is also inversely proportional to normalized variance V . The difference between the traditional spiking deconvolution, i.e., predictive deconvolution, and MED is that the former whitens the signal, thus giving minimum normalized variance, whereas MED identifies isolated spikes, thus giving maximum normalized variance (Wiggins 1978).

Inspired by factor analysis techniques, the entropy term is given in terms of V of a vector $\mathbf{y}$ of reflection coefficients of length N , with amplitude-normalized reflectivity measure q'_i , and a monotonically increasing function $F(q'_i)$:

$$V(\mathbf{y}) = \frac{1}{N \cdot F(N)} \sum_{i=1}^N q'_i \cdot F(q'_i), \text{ where } q'_i = y_i^4 \cdot \left(\frac{1}{N} \cdot \sum_k y_k^2 \right)^{-1}.$$

Given this background (Sacchi et al. 1994), the following inequality holds: $\frac{F(1)}{F(N)} \leq V \leq 1$. Thus, the lower bound of V is when all the samples have equal value. The MED is obtained when $F(q'_i) = q'_i$, which is referred to as MED norm or Wiggins's entropy function. The original algorithm (Wiggins 1978) had used the varimax norm, given by $q'_i = \hat{y}_i^4 \cdot \left(\frac{1}{N} \cdot \sum_k y_k^2 \right)^{-2}$. To resolve the sensitivity of MED norm to strong reflections, other norms have been used, of which the logarithmic norm or the logarithm entropy function, i.e., $F(q'_i) = \ln(q'_i)$, has been found to be effective.

A trivial solution for the reflectivity signal y_t is $\hat{f}_t = x_t^{-1}$, $\hat{y}_t = \delta_t$, which requires inverting the seismic trace, which may not exist. Hence, a filter of fixed length (L) is applied to avoid the requirement of inverse computation. Thus, $y_n = \sum_{l=1}^L f_l \cdot x_{n-l}$.

Maximizing V implies $\frac{\partial V(\mathbf{y})}{\partial f_k} = 0$.

This optimization outputs a vector $\mathbf{b}$, with $b_i = G(q_i) y_i \cdot \left(\frac{1}{N} \sum_j G(q_j) q_j \right)^{-1}$. The final solution is solved using matrix system equation of $\mathbf{f} \cdot \mathbf{R} = \mathbf{g}(f)$, where $\mathbf{R}$ is the Toeplitz matrix of the input data, and the vector $\mathbf{g}(\mathbf{f})$ is the cross-correlation between $\mathbf{b}$ and x . Solving for the fixed length filter is achieved by using an iterative method, given by $\mathbf{f}^{(n)} = \mathbf{R}^{-1} \cdot \mathbf{g}(\mathbf{f}^{(n-1)})$ (Sacchi et al. 1994). In the original algorithm (Wiggins 1978), the reduction is $\mathbf{R} \cdot \mathbf{f} = \mathbf{g}$, where the matrix $\mathbf{R}$ is referred to as an autocorrelation matrix, that is a weighted sum of the autocorrelations of the input signals, and $\mathbf{g}$ is the weighted sum of the cross-correlations of the filter outputs cubed with the input signals. Each iteration involves solving the system using Levinson's algorithm and reproducing the features of the reflectivity signal. If the filter length L is chosen appropriately, the system maximizes V (Wiggins 1978).

A few significant improvements to the original MED method (Wiggins 1978) include the use of D norm instead of the varimax norm (Cabrelli 1984) and implementing the iterative solution in the frequency domain instead of the signal domain (Sacchi et al. 1994). The D norm is defined as $D(y) = \max_{1 \leq k \leq N} \frac{|y_k|}{\|\mathbf{y}\|}$, based on its equivalence to the geometrical interpretation of the varimax norm (Cabrelli 1984). Using MED with D norm, referred to as MEDD, has improved the signal-to-noise ratio of the outputs, and stability

in the presence of additive noise, when compared to MED. In a different kind of enhancement of MED using the varimax norm, the iterative solution is implemented in the frequency domain by minimizing the objective function Φ , which is the sum of the amplitudes of the signal values using a linear programming (LP) formulation (Sacchi et al. 1994). Using MED in the frequency domain, referred to as FMED, is an effective method for processing band-limited data. FMED additionally improves the estimation of higher frequencies compared to the original MED.

Applications

MED has been popularly used for seismic signal analysis in the 1980s and 1990s (Nose-Filho et al. 2018), after which the use of MED has increasingly shifted to applications outside of geophysics, e.g., speech signals (González et al. 1995) and early fault detection (EFD) in rotating machinery (Wei et al. 2019). In applications for seismic deconvolution, the predictive convolution and MED filters work with the single-input, single-output (SISO) systems, whereas the requirements have shifted to single-input, multiple-output (SIMO) systems (Nose-Filho et al. 2018). This shift has been required owing to the use of the common shot gather (CSG), which involves multiple receivers recording different shots from a single shot of the seismic wavelet. Thus, the analysis of SIMO seismic systems has paved the way to multichannel blind deconvolution (MBD) methods. Inspired by MED, sparse MBD (SMBD) accounts for sparsity in reflectivity signals.

As an example of extending MED to other forms of signals, it has been effectively used for period estimation of seismic signals, including natural, artificial, periodic, and quasiperiodic ECG and speech signals (González et al. 1995). MED has been effective and robust for individual period estimation.

Gear failure diagnosis involves analysis of vibration signals, where the quality of MED to identify signals with maximum kurtosis can be effectively used to deconvolve the impulsive sources in gear faults (Endo and Randall 2007). The existing autoregressive (AR) prediction method is used for detecting localized faults in gears, but owing to the use of autocorrelations, AR method is not sensitive to phase relationships. These relationships are necessary for differentiating noise from gear impulses. Since MED uses phase information in the form of kurtosis, MED in combination with AR filtering helps in detecting spalls and tooth fillet cracks in gears. This is an example of how MED is effectively introduced in a workflow of an application, entirely different from a seismic one.

MED has now been used in several such applications of combining with similar methods for EFD in rotating machinery (Wei et al. 2019). Analogous to the shift from SISO to SIMO systems in seismology, gear failure diagnosis uses

MED for solving the fault diagnosis for a single impulse as well as for an infinite impulse train (McDonald and Zhao 2017). Multi D norm, originally used in MEDD, has been used with an adjustment for EFD to generate non-iterative filters to be used on an infinite impulse train. This method is referred to as multipoint optimal MED adjusted (MOMEDA) method. MOMEDA has been further improved for complex working conditions by including product functions (Cai and Wang 2018). Thus, these examples demonstrate the potential of MED in providing stepping stones for improving analysis with increasing complexity in the problem. Another example of a combined method is where variational mode decomposition, L-Kurtosis, and MED are used in sequence for detecting mechanical faults, which have periodic impulses in the vibration signal (Liu and Xiang 2019).

Future Scope

In the recent past, MED has been effectively used in early fault diagnosis in mechanical systems. MED thus has the potential to work on applications involving impulse signals. Starting with seismological applications, MED has proved its mettle in identifying signals with minimum entropy or minimum structure since its inception (Wiggins 1978).

Bibliography

- Cabrelli CA (1984) Minimum entropy deconvolution and simplicity: a noniterative algorithm. *Geophysics* 50(3):394–413
- Cai W, Wang Z (2018) Application of an improved multipoint optimal minimum entropy deconvolution adjusted for gearbox composite fault diagnosis. *Sensors* 18(9):2861
- Endo H, Randall R (2007) Enhancement of autoregressive model based gear tooth fault detection technique by the use of minimum entropy deconvolution filter. *Mech Syst Signal Process* 21(2):906–919
- González G, Badra R, Medina R, Regidor J (1995) Period estimation using minimum entropy deconvolution (MED). *Signal Process* 41(1):91–100
- Liu H, Xiang J (2019) A strategy using variational mode decomposition, L-kurtosis and minimum entropy deconvolution to detect mechanical faults. *IEEE Access* 7:70564–70573
- McDonald GL, Zhao Q (2017) Multipoint optimal minimum entropy deconvolution and convolution fix: application to vibration fault detection. *Mech Syst Signal Process* 82:461–477
- Nose-Filho K, Takahata AK, Lopes R, Romano JMT (2018) Improving sparse multichannel blind deconvolution with correlated seismic data: foundations and further results. *IEEE Signal Process Mag* 35(2):41–50
- Sacchi MD, Velis DR, Cominquez AH (1994) Minimum entropy deconvolution with frequency-domain constraints. *Geophysics* 59(6):938–945
- Wei Y, Li Y, Xu M, Huang W (2019) A review of early fault diagnosis approaches and their applications in rotating machinery. *Entropy* 21(4):409
- Wiggins RA (1978) Minimum entropy deconvolution. *Geoexploration* 16(1–2):21–35

Minimum Maximum Autocorrelation Factors

U. Mueller

School of Science, Edith Cowan University, Joondalup, WA, Australia

Definition

Let $\mathbf{Z}(\mathbf{u}) = [Z_1(\mathbf{u}), Z_2(\mathbf{u}), \dots, Z_K(\mathbf{u})]$ be a K-dimensional random field on some region R in two- or three-dimensional space with covariance matrix Σ_0 and set of experimental semivariogram matrices $\{\Gamma(\mathbf{h}_i) : \mathbf{h}_i = i\mathbf{h}_0, \mathbf{h}_0 \in \mathbb{R}^2 (\mathbb{R}^3), i = 1, 2, \dots, \ell\}$ where $\mathbf{h}_0 \in \mathbb{R}^2 (\mathbb{R}^3)$ is fixed. The minimum/maximum autocorrelation factors $\mathbf{Y}$ of $\mathbf{X}$ are given by

$$\mathbf{Y}(\mathbf{u}) = \mathbf{Z}(\mathbf{u}) \Sigma_0^{-\frac{1}{2}} \mathbf{Q}$$

where $\mathbf{Q}$ is the matrix of eigenvectors of the covariance matrix $\mathbf{I} - \Sigma_0^{-\frac{1}{2}} \Gamma(\mathbf{h}_0) \Sigma_0^{-\frac{1}{2}}$ corresponding to $\mathbf{h}_0$ arranged so that the corresponding eigenvalues are decreasing in magnitude.

Introduction

Principal component analysis is a commonly used method of extracting the most relevant combinations from multivariate data with a view to maximize the explained variance (Davis 2002). However, spatial autocorrelation and spatial cross-correlation present in the data are ignored, and noise in the data stemming from spatial autocorrelation is also not captured. The method of minimum/maximum autocorrelation factors (MAFs) was first introduced by Switzer and Green (1984) as an approach to separate noise and signal in remote sensing data. MAF makes use of global spatial statistics, most notably the experimental covariance matrix and an experimental semivariogram matrix to characterize the short-scale variability of the data. Just as principal components, the MAF factors are linear combinations of the original variables. The first MAF contains maximum autocorrelation between neighboring observations. A higher-order MAF contains maximum autocorrelation subject to the constraint that it is orthogonal to lower-order MAFs and so the factors form a family of factors of decreasing autocorrelation. MAF analysis thus provides a more satisfactory method of orthogonalizing spatial data than PCA. The method has application whenever dimension reduction becomes an issue in geospatial data, and retaining the spatial autocorrelation is desired. MAF has been seen to be more effective than PCA when combined with

classification techniques such as K-NN (Drummond et al. 2010) or linear discriminant analysis (Mueller and Grunsky 2016; Grunsky et al. 2017). For example, Grunsky et al. (2017) showed that it was possible to use the first six MAF factors of the geochemical composition of the Australian surface regolith, followed by LDA to map the underlying crustal architecture, despite secondary modifications due to transport and weathering effects. It should be noted that this represents reduction in the number of variables to 6 from a total of 50 geochemical variables compared to 8 in the case of PCA and 10 if raw variables were used.

Apart from dimension reduction, the most common use in the geosciences lies in the geostatistical estimation and simulation of multivariate data. The MAF method was first applied by Desbarats and Dimitrakopoulos (2000) to simulate pore size distributions. Since this application, MAF has become one of the standard tools for spatial decorrelation as a preprocessing step for geostatistical simulation. The incorporation of MAF was extended to block simulation by Boucher and Dimitrakopoulos (2012), further highlighting the usefulness of the approach for large-scale mining operations. MAF was also used for factorial kriging of geochemical data from Greenland (Nielsen et al. 2000). The authors noted that a combination of maps of MAFs 1, 2, and 3 in a color ternary image resulted in good discrimination of the major provinces of South Greenland highlighting the potential of MAF analysis to identify geochemical boundaries in a region.

There are two approaches which can be used to derive MAFs, data driven or model driven (Rondon 2012). The data-driven approach is based on the sample covariance matrix and a semivariogram matrix at a suitably chosen lag spacing; in contrast, the model-driven approach is based on a linear model of coregionalization (LMC) with two model structures, and the coefficient matrices of the model are used to derive the factors. The model-based approach has been investigated extensively to determine whether extensions to more than two model structures could be used to derive factors; however, an extension to more complex LMCs is not possible (Vargas-Guzman and Dimitrakopoulos 2003; Bandarian et al. 2008), which is a consequence of the non-commutativity of square matrices. One exception is the case when the LMC consists of multiple structures whose coefficient matrices are such that all but one form a commuting family in which case a model-based approach will lead to an LMC that is diagonal (Tran et al. 2006).

Calculation of MAFs

A typical workflow for the application of data-driven MAF is as follows:

1. Compute the variance-covariance matrix Σ_0 of the centered data and determine its principal component decomposition

$$\Sigma_0 = P\Lambda P^T$$

with the entries in Λ arranged in decreasing order, and the matrix P is the corresponding orthogonal matrix of eigenvectors.

2. Compute the PCA scores $\hat{\mathbf{z}}(\mathbf{u}_i) = \mathbf{z}(\mathbf{u}_i)P\Lambda^{-1/2} = \mathbf{z}(\mathbf{u}_i)\Sigma_0^{-1/2}$. At this stage, the new variables have variance 1.
3. Compute the semivariogram matrix $\hat{\Gamma}(\mathbf{h}_0) = \Lambda^{-1/2}P^T\Gamma(\mathbf{h}_0)P\Lambda^{-1/2}$ for $\hat{\mathbf{z}}$ at the chosen lag $\mathbf{h}_0$, and determine its spectral decomposition again ordering the eigenvalues in descending order: $\hat{\Gamma}(\mathbf{h}_0) = Q\Lambda_1Q^T$
4. Put $\mathbf{y}(\mathbf{u}_i) = \hat{\mathbf{z}}(\mathbf{u}_i)Q$. Thus the overall transformation is $\mathbf{y}(\mathbf{u}_i) = \mathbf{z}(\mathbf{u}_i)A$ with

$$A = P\Lambda^{-1/2}Q.$$

The matrix A diagonalizes Σ_0 and $\Gamma(\mathbf{h}_0)$ simultaneously by congruence:

$$A^T\Sigma_0A = Q^T\Lambda^{-1/2}P^T\Sigma_0P\Lambda^{-1/2}Q = I$$

and

$$A^T\Gamma(\mathbf{h}_0)A = Q^T\Lambda^{-1/2}P^T\Gamma(\mathbf{h}_0)P\Lambda^{-1/2}Q = Q^T\hat{\Gamma}(\mathbf{h}_0)Q = \Lambda_1.$$

In the model-driven approach, the variables are first standardized or converted to normal scores. Experimental direct and cross semivariograms are computed based on a suitably chosen lag spacing as represented by the vector $\mathbf{h}_0$. This step is followed by fitting an LMC composed of two structures to the experimental semivariograms:

$$\Gamma(h) = B_1g_1(h) + B_2g_2(h)$$

Here g_1 and g_2 are allowed covariance model functions, and B_1 and B_2 are positive semi-definite coefficient matrices. Their sum is assumed to represent the covariance matrix of the data. It is further assumed that the range of the first covariance model function is less than that of the second; in many cases, the first model is the nugget covariance. The sum of the matrices $B = B_1 + B_2$ is treated as the variance-covariance matrix of the data, and the MAF transformation is determined from B and B_1 using steps 1 and 3 from the procedure above. The construction of the MAF transformation can also be seen as the solution of the generalized eigenvalue problem (Bandarian and Mueller 2008).

Irrespective as to whether the data-driven or model-driven approach is adopted, the factors are ordered in decreasing continuity in contrast to PCA, where spatial behavior is ignored.

The matrix $\mathbf{Q}$ derived in step 3 can be interpreted as containing the correlations between the PCA factors derived in step 1 and the MAF factors (Rondon 2012). The matrix $\mathbf{Q}$ is an orthogonal matrix and thus a composition of rotations and reflections; as a consequence, PCA and MAF factors are rarely equivalent, and the proportion of variance of the original data explained via MAF differs from that explained via the factors or the PCA.

It can also be shown that the MAF factors and the raw data are related by

$$\begin{aligned} E\left(Y(u)^T Z(u)\right) &= E\left(A^T Z(u)^T Z(u)\right) = A^T \Sigma_0 \\ &= \mathbf{Q}^T \Lambda^{-1/2} \mathbf{P}^T \mathbf{P} \Lambda \mathbf{P}^T = \mathbf{Q}^T \Lambda^{1/2} \mathbf{P}^T = A^{-1}. \end{aligned}$$

Thus the inverse of A which is required for back-transformation in the case of simulation or estimation does not need to be computed through inversion, but simply through matrix multiplication (Rondon 2012). In the case when the data are normalized, the expectation $E(Y(u)^T Z(u))$ is nothing other but the correlation matrix between the MAF factors and the raw data, which is analogous to the interpretation in the case of PCA.

A guided application is provided in Rondon (2012) where the normal scores of cadmium (Cd), cobalt (Co), chromium (Cr), and nickel (Ni) metal concentrations (ppm) from the Jura data set (Atteia et al. 1994) are spatially decorrelated with the data-driven and model-driven approaches. The author provides a detailed comparison of the outputs from a simulation study which shows that overall there is little difference in the performance of the two approaches.

Since the data-driven approach does not make any assumptions about the model of continuity, and so is much less restrictive than the model-driven approach, it is the preferred method in geostatistical studies as well as exploratory work.

MAF Biplots

As is the case for PCA, we may use the MAF scores and loadings to construct form biplots based on selected MAF factors. The scores are given by $Y(u) = Z(u)A$, and the matrix A is the corresponding loading matrix. The columns of Y are the principal coordinates, represented as dots, and the rows of the inverse loadings matrix are the analogues of the standard coordinates (represented as arrows).

MAF Analysis for Geochemical Surveys

The analysis of geochemical survey data is one specific area of application for MAF. The motivation is that typically many variables need to be analyzed jointly and for process discovery or hypothesis building, exploratory multivariate statistical methods are applied including PCA. However, the application of a PCA ignores any spatial dependence among variables which might make MAF more appropriate. When analyzing geochemical survey data, attention needs to be paid to the compositional nature of these, and this in turn requires pre-processing of the data through an appropriate log-ratio transformation, such as an additive (alr), a centered (clr), or an isometric (ilr) log-ratio transformation.

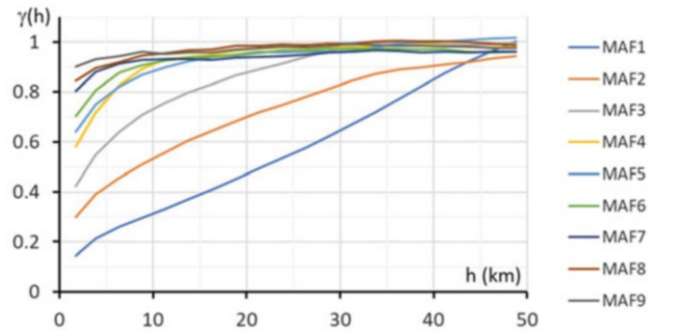
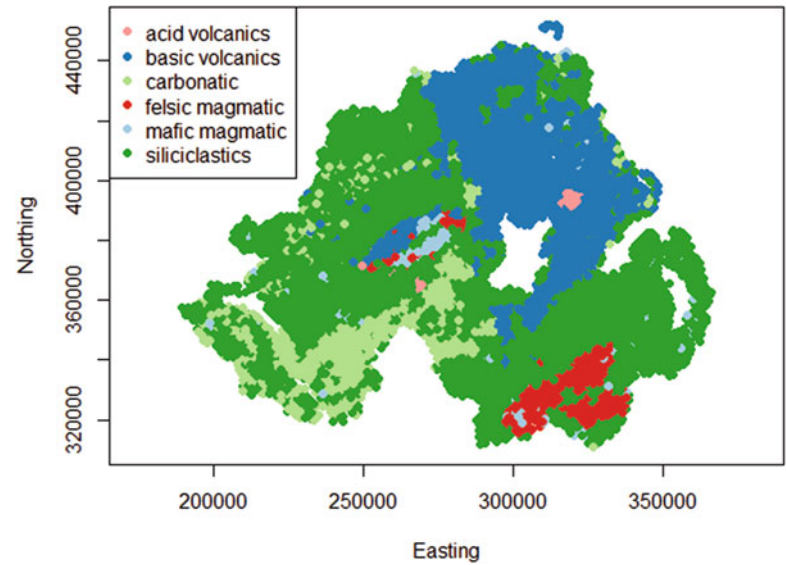
The most common choice is to use the clr-transformation. In this case, the covariance matrix Σ_{clr} of the transformed data is singular, and the PCA of Σ_{clr} can be written as $\Sigma_{clr} = [U_H, D^{-1/2} \mathbf{1}_D] \Lambda [U_H, D^{-1/2} \mathbf{1}_D]^T$, where D denotes the number of clr-variables. The matrix $U = [U_H, D^{-1/2} \mathbf{1}_D]$ is orthogonal and contains the eigenvectors of Σ_{clr} . The matrix $\Lambda = \text{diag}(-\lambda_1, \dots, \lambda_{D-1}, 0)$ is the diagonal matrix of eigenvalues arranged in descending order, with the last eigenvalue equal to 0 corresponding to the eigenvector $\mathbf{u}_D = D^{-1/2} \mathbf{1}_D$ and H is subspace orthogonal to $\mathbf{u}_D$. To complete the MAF transformation, the second decomposition is based on a suitably chosen semivariogram matrix chosen from the set $\{\Gamma_H(h_\ell) = U_H^T \Gamma_{clr}(h_\ell) U_H, \ell = 1, \dots, L\}$ of positive definite $(D-1) \times (D-1)$ matrices (Mueller et al. 2020).

Application

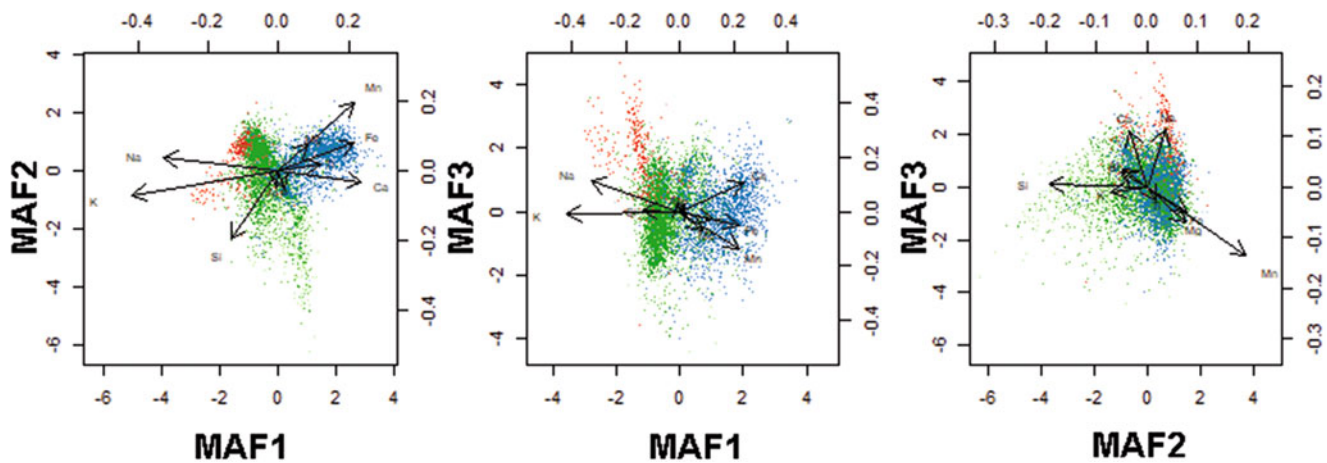
The Tellus geochemical survey was conducted between 2004 and 2007 and covers the region of Northern Ireland, UK (GSNI 2007; Young and Donald 2013). It consists of 6862 rural soil samples (X-ray fluorescence (XRF) analyses), collected at 20-cm depth, with average spatial coverage of 1 sample site every 2 km². Each soil sample site was assigned one of six broadly defined lithological classes (acid volcanics, felsic magmatics, basic volcanics, mafic magmatics, carbonatic, silicic clastics) as described in Tolosana-Delgado and McKinley (2016) and Mueller et al. (2020). A map of the sample locations colored by lithology is shown in Fig. 1.

Nine of the 50 geochemical variables available were chosen for analysis. They represent the bulk of the variability and make up the subcomposition Al, Fe, K, Mg, Mn, Na, P, Si, and Ti. The subcomposition was closed through the addition of a rest variable, and experimental direct and cross variograms of the clr data were computed for 30 lags at a nominal spacing of 1 km. The MAF transform was based on an estimate of the

Minimum Maximum Autocorrelation Factors,
Fig. 1 Map of Northern Ireland covered by broad lithological classes



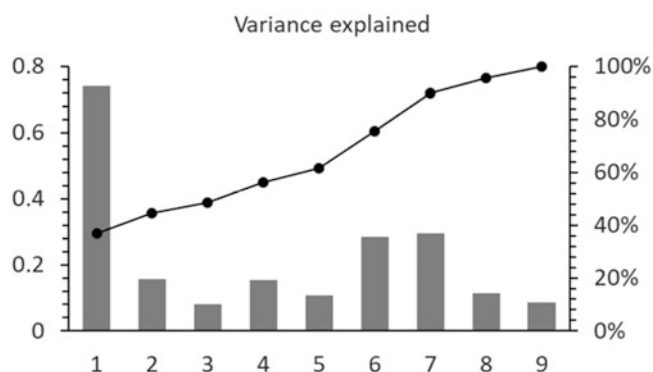
Minimum Maximum Autocorrelation Factors, Fig. 2 Experimental semivariograms of the direct MAF semivariograms



Minimum Maximum Autocorrelation Factors, Fig. 3 Biplots of the first three MAFs; scores are color coded using the classification in Fig. 1

covariance matrix Σ_{clr} and the semivariogram matrix for the first lag $\Gamma_H(h_1)$. The direct variograms (Fig. 2) of the resulting nine factors show clear destructure with increasing factor index, with an increase in the nugget-to-sill ratio from about 0.1 for the first to 0.95 for the ninth factor and a decrease in range from 90 km to 15 km. Already for the fifth factor, the contribution of the nugget is at about 60% for the overall variability.

Biplot for MAF show a ternary system of mafics, felsics, and siliciclastic material (Fig. 3). The biplots of the first and second and first and third MAFs show a separation of

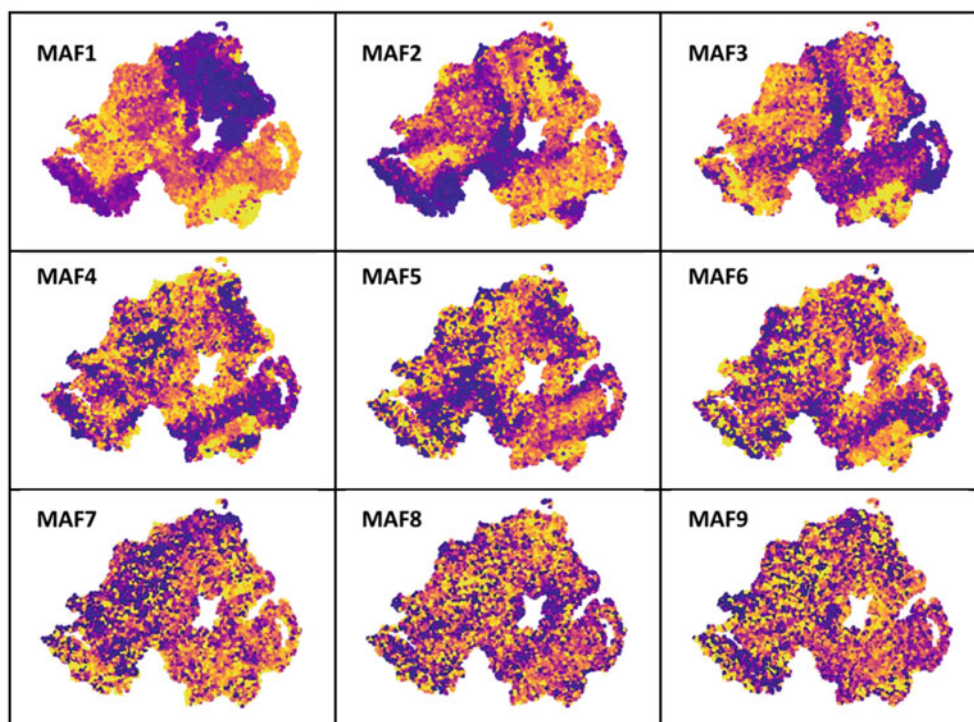


Minimum Maximum Autocorrelation Factors, Fig. 4 Scree plot for factorization based on MAF and cumulative variance explained

lithologies, which is no longer the case for the biplot of second and third MAF. The MAF biplot 1–2 indicates collinearity of Mn, Fe, and Ca, which is also still apparent in the MAF1–3 biplot; in contrast, there is no evidence of separation in the biplot contrasting MAF2 and MAF3.

The scree plot in Fig. 4 shows that the first factor explains approximately 36% of total variability, but in contrast to a PCA scree plot, the function showing the eigenvalues as a function of the MAF-index is not a decreasing function of the index. When compared with the standard PCA, the MAF clearly prioritizes spatial continuity (see Mueller et al. 2020).

The spatial maps of the MAF factors (Fig. 5) identify the lithologies reasonably well. The scores of the first two factors (MAF1 and MAF2) can be interpreted as presenting a broad balance between mafic elements and felsic elements. This feature is related to the contrast of the Paleocene Antrim basalts in the north west with the older rocks across the remainder of Northern Ireland. The intrusive igneous granite and granodiorite are also highlighted (MAF1, MAF2, and MAF3). Carbonates and clastics of different ages are differentiated by the fourth and fifth scores (MAF4 and MAF5). The scores of the sixth, seventh, and eight factors are less clear. The maps also highlight the decrease in spatial continuity of the factors within particular MAFs 1 and 2 having broader spatial features than the remaining MAFs.



Minimum Maximum Autocorrelation Factors, Fig. 5 Maps of MAF factors; top, 1 to 3; center, 4 to 6; bottom, 7–9 showing the decrease in spatial continuity; blue and yellow colors indicate low and high values, respectively

Summary

Decomposition of multivariate spatial data via MAF provides a powerful means to analyze spatial features of data and to characterize spatial scale. The resulting factors allow for univariate instead of multivariate geostatistical simulation and so provide more freedom in the choice of spatial covariance models than what is readily available in the multivariate case.

Cross-References

- Compositional Data
- Geostatistics
- Multivariate Analysis
- Principal Component Analysis
- Spatial Autocorrelation

Bibliography

- Atteia O, Dubois JP, Webster R (1994) Geostatistical analysis of soil contamination in the Swiss Jura. *Environ Pollut* 86:315–327
- Bandarian E, Mueller U (2008) Reformulation of MAF as a generalised eigenvalue problem. *Geostats* 2008:1173–1178
- Bandarian EM, Bloom LM, Mueller UA (2008) Direct minimum/maximum autocorrelation factors within the framework of a two structure linear model of coregionalisation. *Comput Geosci* 34:190–200
- Boucher A, Dimitrakopoulos R (2012) Multivariate block-support simulation of the Yandi iron ore deposit, Western Australia. *Math Geosci* 44:449–468. <https://doi.org/10.1007/s11004-012-9402-9>
- Davis JC (2002) *Statistics and data analysis in geology*, 3rd edn. Wiley, Hoboken
- Desbarats AJ, Dimitrakopoulos R (2000) Geostatistical simulation of regionalized pore-size distributions using min/max autocorrelation factors. *Math Geol* 23(8):919–941
- Drummond, RD, Vidal AC, Bueno JF, Leite EP (2010) Maximum autocorrelation factors applied to electrofacies classification. In: Society of exploration geophysicists international exposition and 80th annual meeting 2010, SEG 2010 Denver, 17 October 2010–22 October 2010, 139015, pp 1560–1565
- Grunsky EC, de Caritat P, Mueller UA (2017) Using Regolith geochemistry to map the major crustal blocks of the Australian continent. *Gondwana Res* 46:227–239
- GSNI (2007) Geological survey Northern Ireland Tellus project overview. <https://www.bgs.ac.uk/gsni/Tellus/index.html>. Accessed 7 Mar 2017
- Mueller UA, Grunsky EC (2016) Multivariate spatial analysis of Lake sediment geochemical data. Melville Peninsula, Nunavut, Canada. *Appl Geochem*. <https://doi.org/10.1016/j.apgeochem.2016.02.007>
- Mueller U, Tolosana-Delgado R, Grunsky EC, McKinley JM (2020) Biplots for compositional data derived from generalised joint diagonalization methods. *Appl Comput Geosci* 8:100044. <https://doi.org/10.1016/j.acags.2020.100044>
- Nielsen A, Conradsen K, Pedersen J, Steenfelt A (2000) Maximum autocorrelation factorial kriging. In: Kleingeld W, Krige D (eds) *Geostats 2000 Cape Town*, Vol 2: 548–558, Geostatistical Association of Southern Africa
- Rondon O (2012) Teaching aid: minimum/maximum autocorrelation factors for joint simulation of attributes. *Math Geosci* 44:469–504
- Switzer P, Green AA (1984) Min/max autocorrelation factors for multivariate spatial imagery. Technical report SWI NSF 06. Department of Statistics, Stanford University
- Tolosana-Delgado R, McKinley J (2016) Exploring the joint compositional variability of major components and trace elements in the Tellus soil geochemistry survey (Northern Ireland). *Appl Geochem* 75:263–276
- Tran TT, Murphy M, Glacken I (2006) Semivariogram structures used in multivariate conditional simulation via minimum/maximum autocorrelation factors. In: *Proceedings XI international congress. IAMG, Liège*
- Vargas-Guzman JA, Dimitrakopoulos R (2003) Computational properties of min/max autocorrelation factors. *Comput Geosci* 29(6): 715–723. [https://doi.org/10.1016/S0098-3004\(03\)00036-0](https://doi.org/10.1016/S0098-3004(03)00036-0)
- Young M, Donald A (2013) A guide to the Tellus data. Geological Survey of Northern Ireland, Belfast. 233pp

Mining Modeling

Youhei Kawamura

Division of Sustainable Resources Engineering, Hokkaido University, Sapporo, Hokkaido, Japan

Definition

A “digital twin” is defined as a virtual representation that serves as the real-time digital counterpart of a physical object or process. Digital twin technology is the result of continuous improvements in product design and engineering. Product drawings and engineering specifications have progressed from handmade drafting to computer-aided design and then to model-based systems engineering. On the other hand, the mining modeling is a relatively universal method of mathematical model construction and application intended to aid managerial personnel at various management levels in decision-making situations, which are frequently characterized by complicated relations of a quantitative as well as logical character.

The digital twin of a physical object is dependent on the digital thread – the lowest-level design and specification for a digital twin – for accuracy to be maintained. Changes to product design are implemented using engineering change orders (ECO). An ECO issued for a component item results in a new version of the item’s digital thread, and correspondingly, of the digital twin.

The concept of digital twins has started attracting attention due to the expansion of the Internet of Things (IoT) and the development of augmented reality (AR) and virtual reality (VR). However, a digital twin is different from general models and simulations. The concept of reproducing and

simulating an entity in a digital space using real-world data is not novel. However, the difference between a digital twin and general simulations is that changes in the real world can be reproduced in the digital space and are linked to each other in real time. With the expansion of IoT, real-world data is being automatically collected in real time and immediately reflected in digital spaces through networks. Thus, the similarity between real-world objects and digital-space models can be maintained. From this perspective, a digital twin is considered to be a “dynamic” virtual model in the real world. Digital twin technology for mining utilises can enable “mining modeling” for various purposes. Moreover, a digital twin for mining is a virtual representation of the physical world and is stored in a representative structure on a cloud data platform.

Introduction

The idea of “digital twin” is attracting considerable attention in the present-day mining industry and is expected to be used in various applications. Unlike older programmable logic controllers (PLCs), distributed control systems (DCSs), and mining execution systems (MESs), a digital twin leverages the latest updates in user interface (UI) and advanced visualization to allow operators recognitions in mine. The digital twin will be at the core of the interface in “smart mining,” which is rapidly gaining acceptance in the global mining industry. This industry is becoming increasingly serious regarding the environment, the economy, and safety.

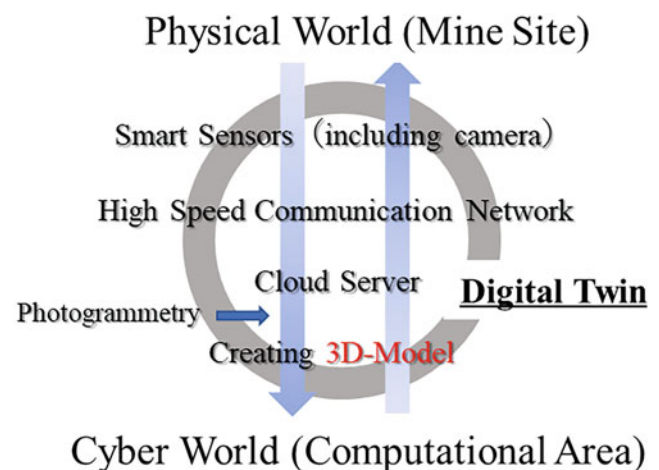
Visualizing the development of the mine, environmental information, the position of workers, and the operational status of trucks at the central control room on the ground surface using communication systems is advantageous for mine operations. Rio Tinto has developed RTVis (Rio Tinto Visualization), a unique 3D visualization tool that can acquire geological information of open-pit mines, the progress of excavation and blasting, and the position information of trucks in real time (Rio Tinto 2020). RTVis superimposes information on a 3D model, determines the development status in real time in the central control room, and supports decision-making. This technology is an example of the use of digital twins for open-pit mining. These are the areas that “mining modeling” has been in charge of so far. The digital twin makes mining modeling easier to apply and is expected to evolve as a field. It is important to create a 3D model that serves as the “reflection destination,” providing an interface for sensing data of a mine whose shape keeps changing over time. CAD-based BIM/CIM in civil engineering and architecture cannot capture shape changes. Currently, two main methods are used for constructing a 3D model. One involves using a laser scanner, and the other is based on photogrammetry. With the rapid progress in drone technology in recent years, it has become possible to capture images and record

videos of large-scale objects, such as open-pit mines. Presently, photogrammetry-built 3D models are being used in open-pit mining projects (Obara et al. 2018).

Structure of Digital Twin

The structure of a digital twin is shown in Fig. 1. The “physical world” (mine site) is measured using a set of smart sensors, including a camera (image sensor), and the obtained data is transmitted through a wireless communication system, such as Wi-Fi or 5G, to a cloud server functioning as a data storage and sharing hub. The on-site analogue data is converted to digital data at the time of measurement by the sensors. A digital twin is created from the digital data on the cloud server. Any individual with access rights and sufficient network speed can utilize the created digital twin since it already exists in the cyber world (computational domain).

Comprehensive cloud data structures require a representative visualization layer for exploring the entire mining site. Thus, digital twins can be used to build mining sites virtually and obtain detailed information regarding mining processes, assets, and key settings (both current and recommended). In summary, by leveraging digital twins to collect and store data, spatial intelligence graphs can be created to model relationships and interactions between individuals, spaces, and devices at mining sites. In addition, a digital twin with advanced simulation capabilities allows for leveraging artificial intelligence and machine learning models from historical data, thereby simplifying future predictive testing. This methodology, unlike general simulations, allows for simulating extraction of information and optimisation of processing equipment in a virtual environment Fig. 1.



Mining Modeling, Fig. 1 Structure of digital twin

Modeling Technologies for Creating Digital Twin

Photogrammetry is the most effective technology for inexpensive and convenient 3D modeling of large-scale spaces. Photogrammetry uses structure from motion (SfM) and patch-based multi-view stereo (PMVS) as constituent technologies. In computer vision, methods for reconstructing an object in three dimensions on a computer based on images captured from various angles (multi-viewpoint images) have been studied extensively. SfM is the most versatile and accurate method, although several other reconstruction methods, such as using shadows in the image (shape from shading) and using focus (shape from defocus), are available.

Currently, a new and effective method is being developed to reconstruct 3D shapes using SfM to obtain models closer to the real object. This method is called multi-view stereo (MVS). When combined with patch-based technology, it is called “patch-based multi-view stereo.” As mentioned earlier, it is necessary to roughly divide modelling process into two steps to create a 3D model. Three-dimensional reconstruction using a multi-view image can yield information on the depth lost by projecting from a 3D image to a 2D image (Furukawa and Ponce 2010). A 3D model of the entire mine, similar to RTVis, is needed to visualize various types of information.

3D Model Reconstruction from Multi-view Images (Structure from Motion)

SfM is a method for simultaneously estimating the position and orientation of a camera and a sparse 3D shape (a collection of points, a sparse point cloud) from multiple images based on multi-view geometry. To estimate the position and orientation of the camera and the sparse point cloud from the multi-viewpoint image, mainly four steps are performed. In sequence, the feature points of the image are extracted and matched, the position and orientation of the camera are estimated, the 3D points are reconstructed via triangulation, and the camera orientation and the 3D point positions are optimized. These four steps are initially performed on two-viewpoint images as a pair. When the camera orientation and 3D points are reconstructed from the two-view images, one new image is added, and processing is performed again between the three images. When the posture estimation and restoration are completed, the images are added, and the process is repeated for all the multi-viewpoint images. Each process is explained below.

- Step 1: Extraction and Matching of Feature Points in Images.

To estimate the relative positions of the two cameras that have captured the two different viewpoint images, it is necessary to be able to first confirm multiple corresponding points between the images. Image feature

extraction technology (also called local image features) is used to search for corresponding feature points between two images. This technique considers a part of an image (a small set of pixels) composed of innumerable pixel values and determines characteristic points (key points) that indicate the overall tendency of that area. Algorithms such as scaled invariance feature transform (SIFT) can automatically calculate the feature points of images and associate them with each other (Lowe 2004). The image-to-image mappings that these algorithms automatically establish often contain errors. In other words, the feature points calculated between the two images are different, but they are related to each other. Random sample consensus (RANSAC) and epipolar constraints are used to eliminate such false correspondences (Isack and Boykov 2012). After performing RANSAC, only the feature points that satisfy the epipolar constraint (the corresponding points of the two images are located along a straight line) are recorded as the official corresponding points.

- Step 2: Estimation of Camera Position and Orientation.

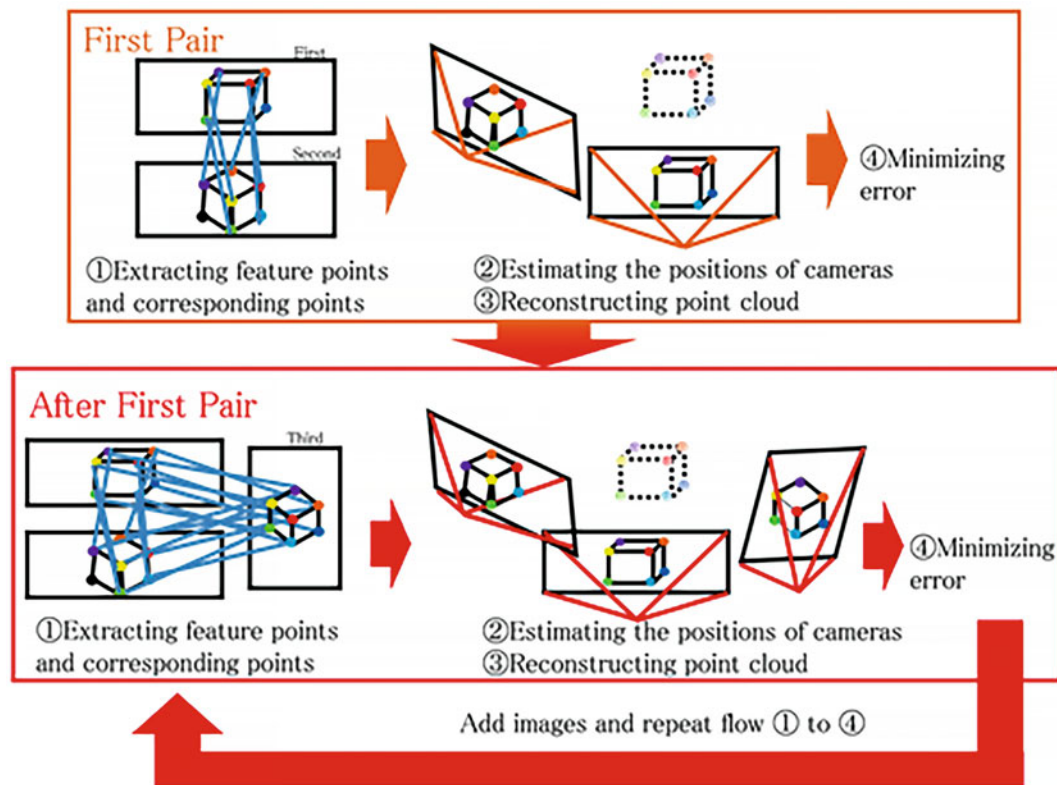
When the search for the corresponding points of the two different viewpoint images is completed, the positions and orientations of the two cameras are estimated. At this stage, some correspondence is found between the two images. Therefore, the least-squares method is applied on these corresponding points to estimate the elementary matrix. Since the basic matrix contains the position and orientation of the cameras, these parameters can be estimated conveniently from more than two images.

- Step 3: Reconstruction of 3D Points via Triangulation.

Based on Steps 1 and 2, the corresponding points of the two images are searched, and the positions and orientations of the two cameras are estimated. The corresponding points satisfy an epipolar constraint as they lie on a straight line between the two images. Therefore, since one side and two angles are known, triangulation is possible, and the three-dimensional point position can be calculated. This calculation is repeated for each corresponding point to ascertain the position of the 3D point.

- Step 4: Optimization of Camera Orientation and 3D Point Position.

After Steps 1–3, the reconstruction of the 3D point is completed. The estimated camera position and orientation and the coordinates of the three-dimensional point overlap with each other as each process is performed. The deviation between the restored 3D point is projected onto the image plane, and the observation point on the corresponding image is called the reprojection error. This reprojection error is optimized, and the camera positions and orientation and the 3D points are estimated again Fig. 2.



Mining Modeling, Fig. 2 Flow of structure from motion (SfM)

Patch-Based Multi-View Stereo (PMVS)

MVS is a method of reconstructing the dense 3D shape of a subject using multi-view geometry under the condition that the position and orientation of the camera have already been clarified by SfM. MVS generates a dense point cloud from a sparse point cloud by estimating the pixel depth of a multi-viewpoint image. Specifically, the depth of each pixel is estimated by exploring the corresponding points in the paired images on the epipolar line using the camera orientation estimated by SfM. Thus, MVS searches for the corresponding points again separately from SfM. PMVS is an advanced method developed by Furukawa et al. to improve the accuracy of conventional MVS (Furukawa and Ponce 2010). PMVS estimates the normal vector and the depth of each pixel simultaneously. First, the feature points are detected, and the initial patch is calculated from the correspondence of the characteristic image area. Thereafter, the patch already obtained is extended to nearby pixels, and the process of removing erroneous corresponding points is repeated to generate a highly accurate and dense patch.

Surface Reconstruction from Point Cloud

The dense point cloud obtained after the processing using SfM and PMVS is automatically inferred from the color information, and the multi-viewpoint image of the object is obtained. After surface reconstruction, the 3D model is finally

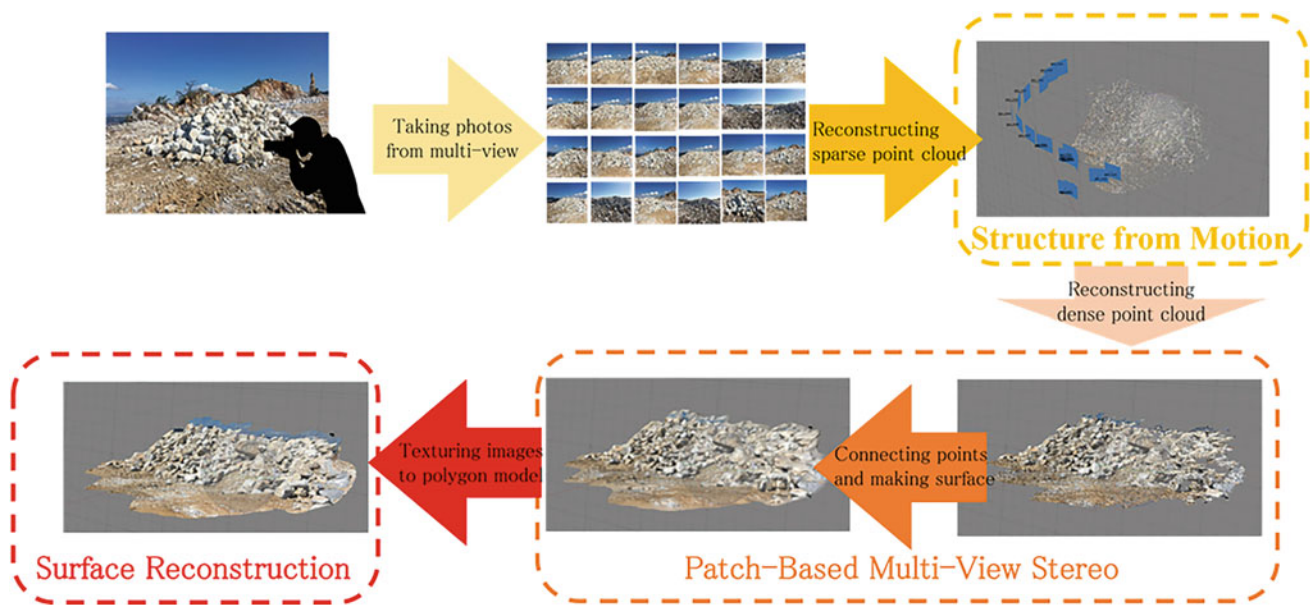
completed. Figure 2 shows a series of steps from a multi-viewpoint image, summarizing the information in Sects. 4.1, 4.2, and 4.3. The lower-left model is the final model with an image pasted on the surface via surface reconstruction Fig. 3.

Examples of Cyber-Physical Implementation (Applications of Digital Twin for Mining)

This section presents specific examples of cyber-physical implementation in mining, including the multiple elemental technologies shown in Fig. 2. Cyber-physical implementation connects the physical world with the cyber world and improves work efficiency using the digital twin. The development of the digital twins for mining will enable rapid implementation of new technology at the mine site through actualization of the cyber-physical platform, which can appropriately allocate the work costs among the physical and cyber worlds. Specific usage examples are introduced separately for open-pit mines and underground mines.

Particle Size Distribution (PSD) Analysis and Blasting Optimization in Open-Pit Mine

The particle size distribution (PSD) of blasted rocks in hard rock mining has significant effects on the subsequent mine-to-mill process. For example, regions of fines and oversized



Mining Modeling, Fig. 3 Constructing 3D model from multi-view Images of mock piles

rocks substantially reduce the loading and hauling productivity. Given that the material transportation cost may reach 60% of the total operating costs, maintaining an appropriate rock fragment PSD is the ultimate objective of mine productivity optimization. Furthermore, the total milling process energy can be reduced by feeding rock fragments within the optimum particle size ranges. For nearly three decades, the mining industry has been reliant on conventional 2D photo-based rock fragmentation measurement methods, which analyze the PSD of rock fragments from 2D surface images of rock piles. Since the 1980s, various 2D photo-based rock fragmentation measurement systems have been introduced, such as IPACS, WipFrag, SPLIT, and PowerSieve. The drawbacks of conventional 2D photo-based rock fragmentation measurement (Con2D) methods have consistently been highlighted by researchers and practitioners. Therefore, fragmentation management is regarded as an essential task for mining engineers.

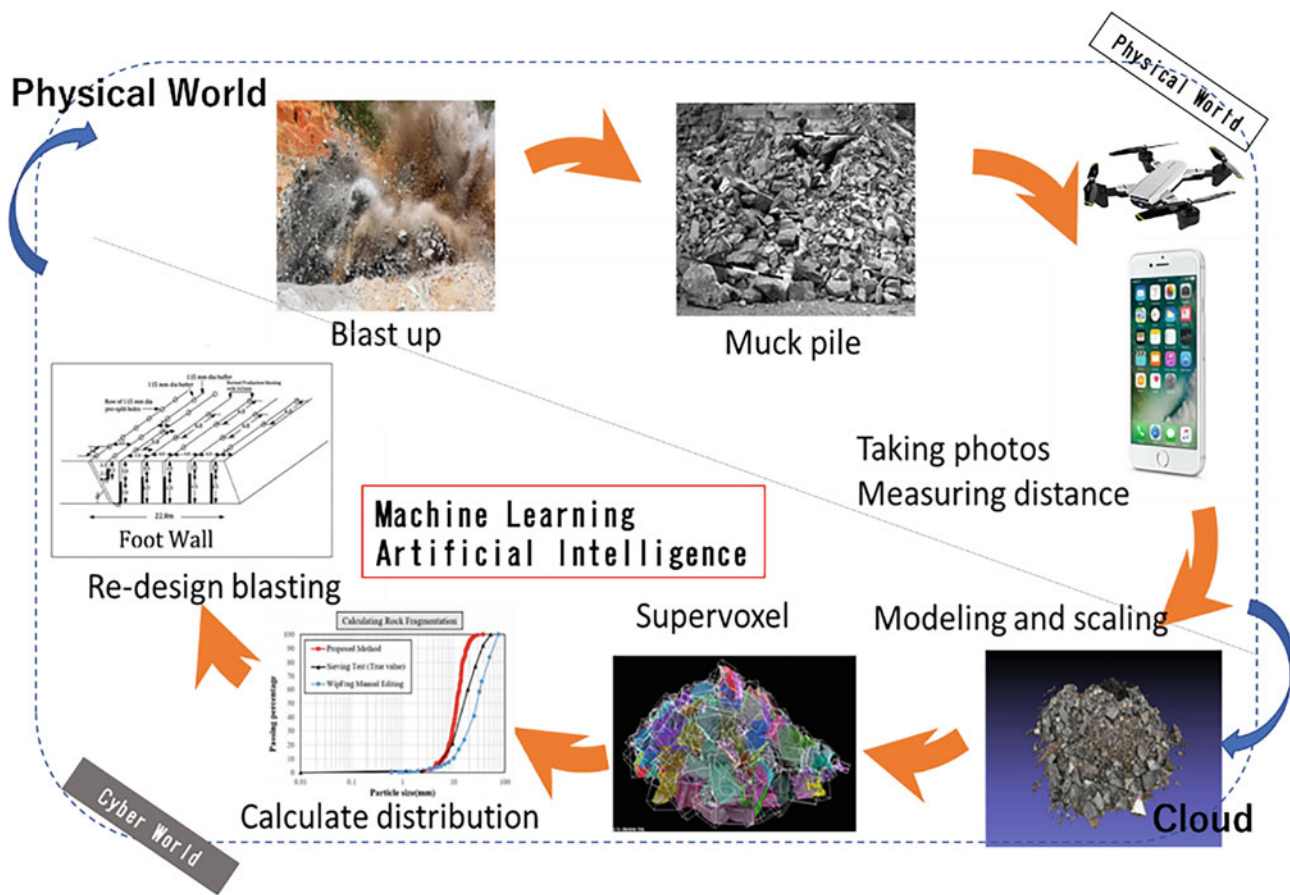
To overcome the limitations of Con2D, a 3D rock fragmentation measurement system (3DFM) based on photogrammetry technology has been proposed. 3DFM extracts the particle sizes from a 3D rock pile model, which facilitates accurate PSD measurement without scale objects and eliminates the need for excessive manual editing. Furthermore, a representative fragmentation of an entire blasting shot can easily be analyzed by generating a 3D model of an entire blasted rock pile (Jang et al. 2019).

In the field of computer vision, which involves equipping a computer with human-like visual function, multi-viewpoint images of an object captured from various angles are integrated to reconstruct the 3D shape of the object. Research on image-based modeling and rendering (IBMR) is being

actively conducted. The objective of our specific study is to develop a particle size distribution (PSD) estimation method that is both accurate and convenient, by applying IBMR to multi-view images of mock piles. In this system, a video or multiple photographs of the mock pile generated by blasting are captured by devices such as a drone or a smartphone. These multi-view images are uploaded to a cloud server via a network such as Wi-Fi or 5G. Then, information is passed from the physical world to the cyber world. A 3D model of the mock pile is generated from the multi-viewpoint images using the aforementioned photogrammetry technique. Thus, digital twins comprising polygons having the same scale as that of the real field are built in the cyber world. Subsequently, it becomes possible to automatically calculate the particle size of each fragmented rock using machine learning (e.g., through supervoxel processing) and to obtain an accurate particle size distribution curve. Since it is possible to create artificial intelligence (AI) with blasting design as the input and particle size distribution as the output, the subsequent blasting design can be optimized, and the new design can be adopted. Thereafter, the information is transferred to the physical world for use. The digital twin enables such blasting optimization Fig. 4.

Visualization of Ground Stress Concentration in Underground Mine

In future, underground mines are expected to become deeper. Therefore, the rock stress at the face will increase. To handle this situation, it is necessary to develop a solution that fully utilizes the latest technology. The main factor affecting underground work safety is rock stress. The purpose of this study was to improve the



Mining Modeling, Fig. 4 PSD analysis and blasting optimization

efficiency and safety of underground work by constructing a rock stress monitoring/visualization system based on a 3D model of an underground mine and AR technology.

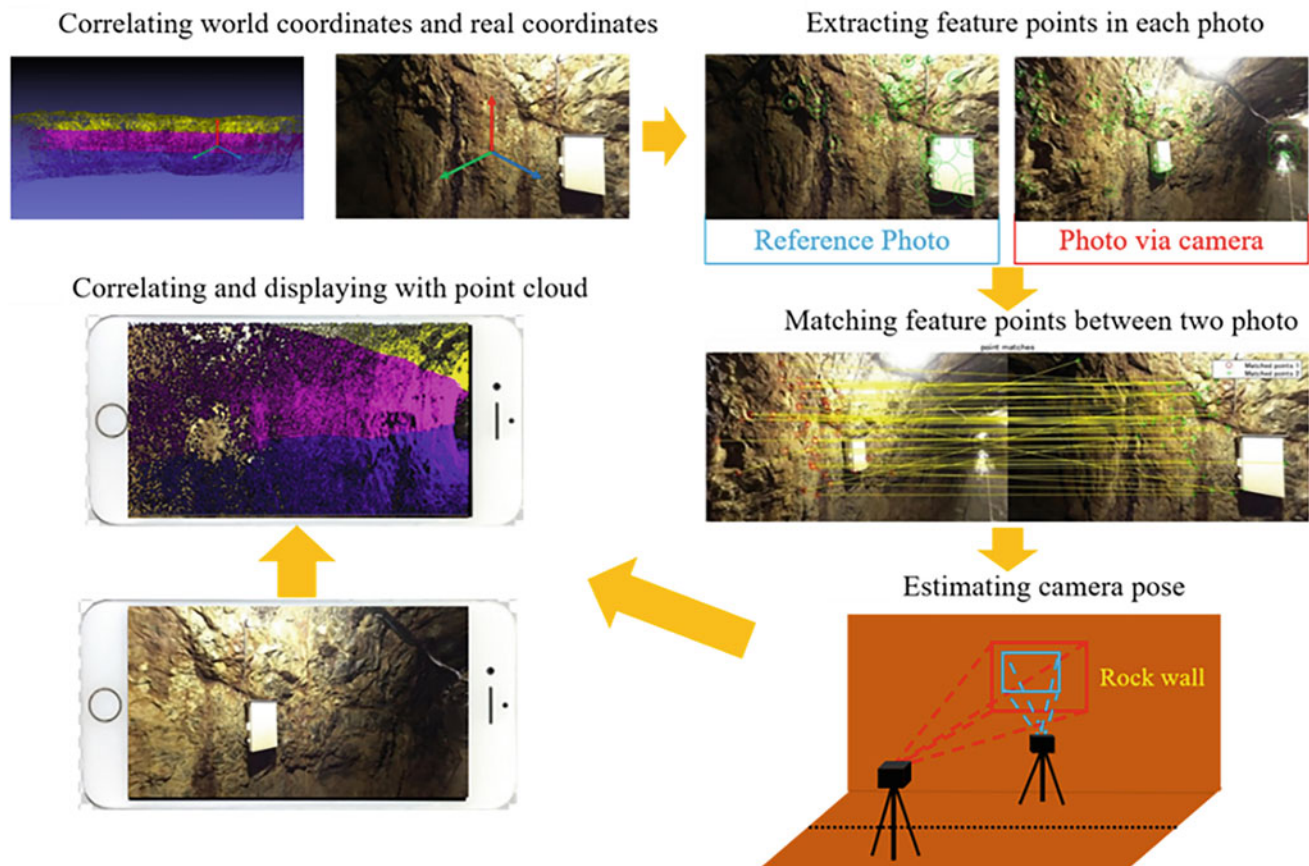
We investigated the technology for visualizing the stress information of the rock mass in a tunnel using the natural feature-tracking method, which is a marker-less tracking method, and a point cloud 3D model of the underground tunnel. To construct the 3D model of an underground mine using photogrammetry, numerous images are required, and the shooting time becomes excessively long. To solve these problems, the methods of constructing a three-dimensional model of the mine using a 360° camera were compared with the method of using a conventional camera. From the experimental results, it was confirmed that reconstruction using a 360° camera is superior in terms of shadow time and effective restoration rate. An optimal method for building a 3D model in an underground mine has thus been developed.

With the point cloud model of the tunnel constructed using the proposed method, the visualization of rock stress by AR was examined. Initially, assuming that the rock stress distribution in the underground tunnel is divided into three regions, the point cloud model in the tunnel was edited in three colors. GPS connectivity is unavailable in underground tunnels, but

the amount of light is constant. Therefore, the natural feature-tracking method was adopted to match the real world with the point cloud model as a virtual object to confirm the extraction and collation of feature points between the images of the rock wall and to estimate the camera attitude and coordinates. From this information, the absolute coordinates of the point cloud model of the mine were determined. By associating these coordinates with the relative coordinates of the camera, the point cloud models of the corresponding colors were superimposed according to the wall surface projected on the camera. An AR system based on this point cloud model can be the basis for visualizing rock stress in real time after the underground monitoring system is completed. This can enable underground workers to identify changes in stress in real time. The system is ultimately expected to provide a safer and more productive work environment for underground workers Fig. 5.

Conclusion

The development of new technologies to improve efficiency and safety in mine development is accelerating. This is further



Mining Modeling, Fig. 5 AR visualisation system for rock stress in an underground mine

boosted by the inevitable fusion of mining with other disciplines. Among them, smart mining (Mining 4.0) represents a fusion of mining with information engineering, offering new possibilities in the mining industry. Furthermore, digital twins are becoming more important as the core technology for smart mining. A digital twin connects the physical and cyber worlds as an effective interface and contributes to the improvement of efficiency and safety by enabling visualization of mining operations.

References

- Furukawa Y, Ponce J (2010) Accurate, dense, and robust multiview stereopsis. *IEEE Trans Pattern Anal Mach Intell* 32(8):1362–1376. <https://doi.org/10.1109/TPAMI.2009.161>
- Isack H, Boykov Y (2012) Energy-based geometric multi-model fitting. *Int J Comput Vis* 97:123–147. <https://doi.org/10.1007/s11263-011-0474-7>
- Jang H, Kitahara I, Kawamura Y, Endo T, Degawa R, Mazara S (2019) Development of 3D rock fragmentation measurement system using photogrammetry. *Int J Min Reclam Environ* 34(4):294–305. <https://doi.org/10.1080/17480930.2019.1585597>
- Lowe DG (2004) Distinctive image features from scale-invariant key points. *Int J Comput Vis* 60:91–110. <https://doi.org/10.1023/B:VISI.0000029664.99615.94>
- Obara Y, Yoshinaga T, Hamachi M (2018) Measurement accuracy and case studies of monitoring system for rock slope using drone. *J MMIJ* 134(12):222–231. <https://doi.org/10.2473/journalofmmij.134.222>
- Rio Tinto (2020) Smart mining. <https://www.riotinto.com/en/about/innovation/smart-mining>. Accessed 2020/12/24

Modal Analysis

Frits Agterberg
Central Canada Division, Geomathematics Section,
Geological Survey of Canada, Ottawa, ON, Canada

Definition

Modal analysis is the study of system components in the frequency domain. In the geosciences, it is especially important in igneous and sedimentary petrography. It assumes that minerals

are randomly distributed in the rocks being studied and that proportions of the rock-forming crystals (or other types of rock particles) can be determined by point counting using a special microscope or microprobe. For points forming a regular grid, the presence or absence of the rock-forming minerals is counted to measure the volume percentages of the minerals considered. If distances between points are sufficiently long, exceeding linear dimensions of the mineral grains considered, the binomial distribution and its normal approximation can be used to model these presence-absence data.

Introduction

In the geosciences, the technique of modal analysis was introduced by Chayes (1956) for widespread use in petrography; it had several precursors. Delesse (1848) introduced the principle of measuring volume percentage values of constituents of rocks by measuring the presence of every constituent at equally spaced points along lines. However, the initial method was exceedingly laborious as described by Johannsen (1917). Delesse's preferred method was to trace, on oiled paper, the outline of each constituent as shown on polished slabs. After coloring the resulting drawing, it was pasted on tinfoil and cut apart for each component, soaked off the paper, and weighed. Half a century later, Rosiwal (1898) introduced his improved method of linear measurements. Subsequently Shand (1916) described the method of employing a mechanical stage as part of the microscope that became generally used, because it requires very few measurements. Other early applications of modal analysis include Glagolev (1932) and Krumbein (1935).

Chayes (1956) brought mathematical statistics to bear on how to use the stage. His technique is based on the binomial distribution, which is a discrete probability distribution with parameters n (= number of points counted) and p (= number of points on mineral considered). It is assumed that n independent experiments are performed, each answering a yes-no question with probability p of obtaining a Boolean-value (0 or 1) outcome. The probability of no success satisfies $q = 1 - p$. A single success-failure experiment is also called a Bernoulli trial.

Binomial Distribution

The probability $P(k, n)$ of k successes in n experiments satisfies the binomial frequency distribution:

$$P(k, n) = \binom{n}{k} p^k q^{n-k} \text{ where } \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

The mean μ of this distribution satisfies $\mu = np$, and the standard deviation is $\sigma = \sqrt{npq}$.

The following example illustrates the application of the binomial distribution in petrographic modal analysis. Suppose that 12% by volume of a rock consists of a given mineral "A." The method of point counting is applied to a thin section of this rock. Suppose that 100 points are counted. From $p = 0.12$ and $n = 100$, it follows that $q = 0.88$ and

$$\mu = np = 12; \sigma = \sqrt{npq} = \sqrt{12 \times 0.88} = 3.25.$$

If the experiment of counting 100 points would be repeated many times, the number of times (K) that the mineral "A" is counted would describe a binomial distribution with mean 12 and standard deviation 3.25. According to the so-called central-limit theorem, a binomial distribution approaches a normal distribution if n increases. Hence, we can say for a single experiment that:

$$P(\mu - 1.96\sigma < K \leq \mu + 1.96\sigma) = 0.95$$

or

$$P(5.6 < K \leq 18.4) = 95\%$$

The resulting value of K is between 5.6 and 18.4 with a probability of 95%.

This precision can be increased by counting more points. For example, if $n = 1000$, then $\mu = 120$ and $\sigma = 10.3$ and

$$P(99.8 < K \leq 140.2) = 95\%$$

This symmetrical probability can also be written as $k = 120 \pm 20.2$. To compare the experiment of counting 1000 points to that for 100 points, the result must be divided by 10, giving $k' = 12 \pm 2.02$. This number represents the estimate of volume percentage for the mineral "A." It is $3.25/(10.3/10) = 3.2$ times as precise as the first estimate for 100 points only. It is noted that this increase in precision agrees with the formula $\sigma^2(\bar{x}) = \sigma^2(x)/n$ for the population variance of $\bar{x}$ representing the mean of n values. The method of modal analysis has been discussed in more detail by Chayes (1956).

Antarctic Meteorite Case Study

Petrographic modal analysis is usually performed in conjunction with other methods such as chemical analysis of rock considered, as seen in McKay et al. (1999), that is taken here

as an example. This particular study describes a preliminary examination of the Antarctic Meteorite LEW 87051 (total weight 0.6 g) which is porphyritic in texture, with equant, euhedral to subhedral olivine phenocrysts ~0.5 mm wide set in a line-grained groundmass of euhedral plagioclase laths and interstitial pyroxene, Fe-rich olivine, kirschsteinite, and some minor constituents. Microprobe modal analysis was performed on these minerals by counting 721 points ranging from a maximum of 34.1% plagioclase to a minimum of 4.7% kirschsteinite, plus traces of several other minerals.

The information given in the preceding paragraph can be used to determine the approximate precision of the estimated volume percentage values by using the normal approximation of the binomial model as discussed in the preceding section. For very small frequencies, it would be better to approximate the binomial model by an asymmetric Poisson distribution model instead of by the normal distribution model. This is because, for very small frequencies, the binomial model approaches its Poisson distribution limit. According to Hald (1952), this limit can be set at $\theta < 0.1$, where $\theta = \mu/n$. This lower limit criterion is not met for kirschsteinite in the current example with $\theta = 0.047$ suggesting that use of the Poisson model would be better. Consulting tables published by Mantel (1962), and later reproduced by Johnson and Kotz (1969), then give its 95% confidence interval as $P(23.6 < K \leq 48.3) = 95\%$ instead of $P(22.6 < K \leq 45.3) = 95\%$ as results from the normal approximation of the binomial distribution. This is a very small improvement only.

Summary and Conclusions

The purpose of petrographic modal analysis is to determine the volume percentages of rock constituents from thin sections under the microscope. The spacing between points at which the constituents are counted should be sufficiently wide to avoid counting the same grain more than once. Precision of the results can be determined by using the binomial distribution model. For rare constituents, the Poisson model can be used. The Antarctic Meteorite LEW 87051 (total weight 0.6 g) was taken for example.

Bibliography

- Chayes F (1956) Petrographic modal analysis. Wiley, New York, p 113
 Delesse M (1848) Procédé mécanique pour déterminer la composition des roches. Ann Min XIII(4):379–388
 Glagolev AA (1932) Quantitative mineralogical analysis of rocks with the microscope. Gosgeolizdat, Leningrad, pp 1–25
 Hald A (1952) Statistical theory with engineering applications. Wiley, New York, p 783

- Johannsen A (1917) A planimeter method for the determination of the percentage compositions of rocks. J Geol 25:276–283
 Johnson NI, Kotz S (1969) Distributions in statistics: discrete distributions. Houghton Mifflin, Boston, p 328
 Krumbein WC (1935) Thin-section mechanical analysis of indurated sediments. J Geol 43:482–496
 Mantel N (1962) Lung cancer mortality as related to residence and smoking histories. I. White males. J Natl Cancer Inst 28:947–997
 McKay G, Crozaz G, Wagstaff J, Yang SR, Lundberg L (1999) A petrographic, electron microprobe, and ion microprobe study of mini-angrite Lewis Cliff 87051. In: Abstracts of the LPSC XXI, pp 771–772
 Rosiwal A (1898) Über geometrische Gesteinsanalysen. Ein einfacher Weg zur ziffermässigen Feststellung des Quantitätsverhältnisses der Mineral bestandtheile gemengter Gesteine. Verh der k-k Geol Reichsanstalt, Wien, pp 143–175
 Shand J (1916) A recording micrometer for geometrical rock analysis. J Geol 24:394–401

Moments

Jessica Silva Lomba¹ and Maria Isabel Fraga Alves²

¹CEAUL, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal

²CEAUL & DEIO, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal

Definition

Statistical moments can be introduced as features of (the probability distribution of) a random variable (RV). These numerical characteristics describe the behavior of the RV's distribution, such as location and dispersion, playing an important role in its identification, fitting, and estimation of parameters. Theoretical (population) moments are generally defined as expectations of functions of the RV, while their empirical counterparts, the sample moments, are basically computed as averages of similar functions of observations. The most common statistical moments in Geosciences are the following: the conventional first moment – *mean*, the second central moment – *variance*, and the standardized third and fourth central moments – *skewness* and *kurtosis*. However, Probability Weighted Moments (PWM) and L-moments have also been gaining visibility in this field.

Introduction

The term *statistical moment* refers to application of the mathematical concept *moment* as a tool in the study of probability, or frequency, distributions. First introduced by Karl Pearson (1893) in a short study of asymmetrical frequency curves, moments became widely spread probabilistic features across

all data-based pursuits. The choice of name *moment* connects strongly to the physical interpretation of the moments of mass distributions in Mechanics.

Broadly speaking, for a continuous real function of real variable $f(x)$ and $c, k \in \mathbb{R}$, consider the moment integral

$$\int_{-\infty}^{+\infty} (x - c)^k f(x) dx. \quad (1)$$

This is specially interesting when $k \in \mathbb{N}_0$, as (1) then defines the k^{th} moment of f about c . We restrict our exposition to the continuous case, for practical relevance, but discrete analogues can be found in the literature (e.g., Rohatgi and Saleh 2015). This concept is also extendable to higher dimensions (e.g., see “Multivariate Analysis of Variance”), though our main focus will be the univariate setting.

Illustrating the physical interpretation, consider the geophysical study of the Earth’s gravity field by Jekeli (2011), exploring the relationship between low-degree spherical harmonics of the Earth’s gravitational potential and the moments of its mass density (Eq. 50 therein). It is seen that the 0th order density moment is proportional to the Earth’s *total mass*, the first moments are proportional to its *center of mass*, and the second moments define the *inertia tensor* – which can indicate asymmetry in mass distribution. We borrow these ideas for interpretation of statistical moments.

Throughout, we think of an RV X with continuous, univariate distribution and density functions F and $F' := f$ (see “Probability Density Function” – PDF). Moment systems allow identification and description of the distribution of X in terms of location (analogous to center of mass) and shape (including dispersion and symmetry about the center), as functions of its parameters. Historically, moments have been recommended in the context of Geosciences since the 1900s.

Since Pearson’s seminal work, generalizations of the conventional moments (or simply Moments) have been suggested, addressing shortcomings of the primary system. However, the conventional remains the most widely known. We present three systems’ definitions and usage, illustrated with applications in Geosciences.

Conventional Moments

Traditionally, the k^{th} moment of X refers to the expectation of X^k

$$\mu'_k := \mathbb{E}[X^k] = \int_S x^k f(x) dx, \quad (2)$$

also known as *raw moment*, stemming from (1) with f the PDF of X , $c = 0$, and integrating over the variable’s support S . Notice $\mathbb{E}[X^k]$ in (2) exists only if $\mathbb{E}|X|^k := \int |x|^k f(x) dx < \infty$.

It becomes apparent that $\mu'_0 = 1$ gives the total probability ($\sim$ total mass). The *first moment* of X corresponds to the *mean* or *expected value* of the distribution, $\mu := \mathbb{E}[X]$, if it exists, and provides a measure of location, a central value of the distribution ($\sim$ center of mass).

Some distributions have all finite moments (e.g., see “Normal Distribution”), but not all moments of a distribution necessarily exist (e.g., Cauchy distribution); however, existence of the k^{th} moment implies existence of all moments of lower order $0 < n < k$. The *Moment Generator Function* (MGF) of X is defined as

$$M(s) := \mathbb{E}[e^{sX}], \quad (3)$$

provided the expectation on the right exists around the origin, and so the moments of X appear as successive derivatives of $M(s)$ at that point: $M'(0) = \mathbb{E}[X]$; $M''(0) = \mathbb{E}[X^2]$; in general, $M^{(k)}(0) = \mathbb{E}[X^k]$, for $k \geq 1$. The MGF may not exist, but existence of $M(s)$ implies existence of all order μ'_k , uniquely determining F .

If $\mathbb{E}[X] < \infty$, useful features arise by considering the moments centered about the mean, the *central moments*, of X

$$\mu_k := \mathbb{E}[(X - \mu)^k] = \int_S (x - \mu)^k f(x) dx, \quad (4)$$

as they directly inform on the distribution’s shape. These can be computed from raw moments, and *vice versa*. The first central moment is trivially 0, but, when they exist, higher order ones are specially relevant:

– Provided $\mathbb{E}|X|^2 < \infty$, then

$$\sigma^2 := \mu_2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \quad (5)$$

is the *variance* of X , and measures dispersion around the mean ($\sim$ moment of inertia); the square root, σ , is the *standard deviation*, expressed in the units of X , and its ratio by the mean results in the *coefficient of variation*

$$c = \frac{\sigma}{\mu}; \quad (6)$$

(see “Variance,” “Standard Deviation”);

- For symmetric distributions, all odd central moments are 0.
- The third and fourth *standardized* central moments give dimensionless measures of asymmetry and peakedness of distributions, respectively, the *coefficient of skewness*

$$\gamma_1 := \frac{\mu_3}{\sigma^3} \text{ and kurtosis } \gamma_2 = \frac{\mu_4}{\sigma^4}.$$

This topic is comprehensively addressed in Chapter 3 of Rohatgi and Saleh (2015).

The primary practical interest of these features is to compare them, in a way, to their empirical counterparts. For RV collection $X_1, X_2, \dots, X_n$, the k^{th} sample moment and sample central moment (cf. Rohatgi and Saleh 2015, Chapter 6) are defined, respectively, by

$$m'_k := \frac{1}{n} \sum_{i=1}^n X_i^k \quad \text{and} \quad m_k := \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k \quad (7)$$

where the common average $\bar{X} := m'_1$ denotes the *sample mean*. When $\{X_i\}_{i=1}^n$ constitute a random sample (RS) – independent, identically distributed (i.i.d) RVs – if μ'_k exists, the statistic m'_k is a consistent and unbiased estimator for that parameter. In fact, $\bar{X}$ is μ 's linear unbiased estimator of minimum variance (see “Best Linear Unbiased Estimator”). The sample central moments are generally not unbiased for their population analogues. Hence, the *sample variance* is commonly defined as

$$S^2 := \frac{n}{n-1} m_2, \quad (8)$$

an unbiased estimator for σ^2 (not so for $S = \sqrt{S^2}$). Regarding higher orders, the *sample coefficient of skewness* and *sample kurtosis*, respectively,

$$g = \frac{n^2}{(n-1)(n-2)} \frac{m_3}{S^3} \quad \text{and} \quad (9)$$

$$k = \frac{n^2}{(n-2)(n-3)} \left[\left(\frac{n+1}{n-1} \right) \frac{m_4}{S^4} - 3 \frac{(n-1)^2}{n^2} \right] + 3,$$

are possibly markedly biased estimators of γ_1 and γ_2 , with undesirable algebraic bounds determined by the sample size, $|g| \leq \sqrt{n}$ and $k \leq n+3$ (Hosking and Wallis 1997), and are highly outlier-sensitive. Mind that software usually computes the sample *excess kurtosis*, $k-3$, by comparison with the Normal distribution ($\gamma_2 = 3$).

Sampling distributions and asymptotic properties of these statistics have been widely studied and yield extremely useful results, like the Central Limit Theorem and Monte Carlo methods.

Other functions of moments relevant in applications can be listed, especially in the context of multivariate and spatial problems. Chapter 4 of Rohatgi and Saleh (2015) includes a

detailed exposition of generalizations of (2)–(4) to multiple RVs. We briefly mention a few aspects:

- The *covariance* between two RVs X and Y

$$\text{cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \quad (10)$$

(if the expectations exist) provides a measure of joint variation, particularly, $\text{cov}(X, X) = \sigma_X^2$; this concept is generalized for n -dimensional vectors by the (symmetric) *variance-covariance* matrix, where the main diagonal shows individual variances and off-diagonal entries show pairwise covariances;

- While $\text{cov}(X, Y)$ points if Y tends to increase or decrease when X increases (respectively, $\text{cov}(X, Y) > 0$ or $\text{cov}(X, Y) < 0$), the *correlation coefficient*

$$\rho_{X,Y} = \text{corr}(X, Y) := \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \quad (11)$$

indicates the relationship's strength. Note $|\rho_{X,Y}| \leq 1$, the equality equivalent to linear association between X and Y (see “Correlation Coefficient,” “Multiple Correlation Coefficient”).

- If X and Y are independent, $\text{cov}(X, Y) = 0 = \rho_{X,Y}$, i.e., X and Y are uncorrelated; the converse is not necessarily true;
- For a bivariate sample $\{(X_i, Y_i)\}_{i=1}^n$, these characteristics are estimated by the *sample covariance* and *correlation coefficient*, respectively,

$$S_{X,Y} := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \quad \text{and} \quad (12)$$

$$R_{X,Y} := \frac{S_{X,Y}}{S_X S_Y};$$

- In Geostatistics, for a stochastic process $Z(s)$, a common generalization is the *variogram*

$$2\gamma(s_1, s_2) := \mathbb{E}[(Z(s_1) - \mathbb{E}[Z(s_1)]) - (Z(s_2) - \mathbb{E}[Z(s_2)])]^2, \quad (13)$$

representing the variance of the difference of the process at distinct points s_1, s_2 . If $Z(s)$ is stationary, the variogram is simply the function of the distance $h = s_2 - s_1$, since $\gamma(s_1, s_2) = \gamma(0, s_2 - s_1) := \gamma(h)$. Assuming intrinsic stationarity

of $Z(s)$, for points $\{s_i\}_{i=1}^n$, the prevailing *empirical/experimental (semi)variogram* is computed as

$$2\hat{\gamma}(h) := \frac{1}{n(h)} \sum_{i=1}^{n(h)} (Z(s_i) - Z(s_i + h))^2, \quad (14)$$

with $n(h)$ the number of pairs of points at distance h , and several theoretical models are available to fit this estimate (see “Variogram,” “Stochastic Process,” and “Stationarity”).

Method of Moments

“Method of Moments”(MM) is a broad category including various estimation methodologies, applied in several contexts, presented in different ways, but built on the same basis: essentially, equating sample moments to their population match – a substitution principle.

We focus on estimation of parameters of an assumed distribution F from a RS $X_1, \dots, X_n$, although the principle can be generally applied to statistical models. Let $\theta = (\theta_1, \dots, \theta_p) \in \Theta$ be the parameter vector of F . Knowing population μ'_k are functions of θ , the MM estimate $\hat{\theta}$ is the solution to an equation system of the type

$$\mu'_k(\theta) = m'_k \text{ for } k = 1, \dots, m \quad (15)$$

with $m \geq p$ the smallest integer guaranteeing a determined system. In general, the MM estimator of $\lambda = g(\mu'_1, \mu'_2, \dots, \mu'_k)$, a real-valued function of moments, is simply $\hat{\lambda} = g(m'_1, m'_2, \dots, m'_k)$; if g is continuous, $\hat{\lambda}$ is a consistent estimator for λ , asymptotically normal under mild conditions (Rohatgi and Saleh 2015, Chapter 8). Hence, some raw moment equations in (15) are frequently replaced by central moment relatives.

Compared to other methods (e.g., “Maximum Likelihood”), MM estimators are more straightforwardly computed, despite being usually less efficient and possibly resulting in inadmissible estimates $\hat{\theta} \notin \Theta$.

Applications in Geosciences

Moments and MM have been vastly applied in many fields of Geosciences. We mention a few examples:

- A frequent use of the variogram is *Kriging* – an optimized spatial interpolation method (see “Kriging”), prolific in Geostatistics’ literature since the 1960s.

- Building on the concept of *cumulant* – a combination of moments, generated by the logarithm of the MGF – Dimitrakopoulos et al. (2010) suggested a modeling (and, in subsequent work, simulation) approach for non-linear, non-Gaussian spatial data by introducing high-order *spatial cumulants*. These were found to efficiently capture complex spatial patterns and geological characteristics. The associated simulation methodology for spatially correlated attributes is shown to be advantageous compared to previous multiple-point (MP) approaches (see “Multiple Point Statistics”).
- Osterholt and Dimitrakopoulos (2018) addressed limitations of traditional variogram-based stochastic simulation techniques in capturing non-linear geological complexities of orebodies; the authors revisit MP simulation as an algorithm for the simulation of the geology for mineral deposits, an approach based on high-order spatial statistics and a type of extension of *kriging* systems. This method, which allows for resource uncertainty assessment, is applied to the Yandi channel iron ore deposit.
- Recently, Wu et al. (2020) studied how soil properties’ distribution characteristics influence probabilistic behavior of shallow foundations’ total and differential settlements. Focusing on the influence of the first four moment statistics, used to define the PDF of elastic modulus E , on the settlement’s distribution, the authors suggest a moment-based simulation method for modeling E as a 2-dimensional non-Gaussian homogeneous field. The study (via Monte Carlo simulations) shows that mainly skewness of E significantly influences total/differential settlements and suggests greater attention be given to identifying low-skewed situations, which may result in dangerous unexpectedly large settlements.

Probability Weighted Moments

The PWM comprise another valuable system to characterize probability distributions, whose theory parallels that of conventional moments. Introduced by Greenwood et al. (1979), the primary aim was deriving expressions for parameters of continuous distributions with explicit inverse function $\chi(\cdot)$ – the *quantile function*, satisfying $\chi(F(x)) = x$, $x \in S$. PWM have been extensively used in Geosciences, counteracting some conventional moments’ drawbacks leading to unsatisfactory inference. See Hosking and Wallis (1997) and references therein for further details.

The PWM of X are defined as

$$\begin{aligned} M_{p,r,s} &:= \mathbb{E}[X^p F(X)]^r (1 - F(X))^s \\ &= \int_S x^p (F(x))^r (1 - F(x))^s f(x) dx, \text{ with } p, r, s \in \mathbb{R}, \end{aligned} \quad (16)$$

where similarities with (2) are clear: $\mu'_k = M_{k,0,0}$, $k \in \mathbb{N}_0$. Their usefulness stems from being translated in terms of $\chi(u)$, $0 < u < 1$: contrast the notable PWM

$$\alpha_s := M_{1,0,s} = \int_0^1 \chi(u) (1 - u)^s du, s \in \mathbb{N}_0 \quad (17)$$

$$\beta_r := M_{1,r,0} = \int_0^1 \chi(u) u^r du, r \in \mathbb{N}_0 \quad (18)$$

against the redefined (2)

$$\mu'_k = \int_0^1 (\chi(u))^k du, k \in \mathbb{N}_0. \quad (19)$$

Higher complexity of μ'_k is evident, as successively higher powers of $\chi(\cdot)$ are employed, absent from α_s and β_r .

Appealing properties of PWM include the following:

- $M_{p,r,s}$ exists for all $r, s \geq 0$ if and only if $\mathbb{E}|X|^p < \infty$.
- Complete determination and characterization of F by the sets $\{\alpha_s\}_{s \in \mathbb{N}_0}$ or $\{\beta_r\}_{r \in \mathbb{N}_0}$ if $\mathbb{E}|X| < \infty$ (note the two moments are functions of each other); however, interpretation as distribution's features is not apparent.
- Connection to expectations of *order statistics* $X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$; particularly, $n\alpha_{n-1} = \mathbb{E}[X_{1:n}]$ and $n\beta_{n-1} = \mathbb{E}[X_{n:n}]$.
- Straightforward estimation – for i.i.d. $\{X_i\}_{i=1}^n$, the *sample PWM*

$$a_s := \frac{1}{n} \sum_{i=1}^n \binom{n-i}{s} X_{i:n} \binom{n-1}{s}^{-1}, s = 0, 1, \dots, n-1 \quad (20)$$

$$b_r := \frac{1}{n} \sum_{i=r+1}^n \binom{i-1}{r} X_{i:n} \binom{n-1}{r}^{-1}, r = 0, 1, \dots, n-1 \quad (21)$$

are unbiased estimators of α_s and β_r , also asymptotically normal if μ'_2 of F exists.

Applications: PWM Methods

The “PWM method” falls in the MM category, following the same principle as presented above: matching population $M_{p,r,s}$, expressed in terms of the parameters $\theta = (\theta_1, \dots, \theta_p)$ of F , with their empirical functionals, for example,

$$\alpha_s(\theta) = a_s, \text{ for } s = 0, 1, \dots, m \geq p. \quad (22)$$

These equations are, again, interchangeable with other convenient equalities of (functions of) PWM.

PWM-type estimators have been shown to be more robust to *outliers*, less subject to sampling variability and estimation bias, and frequently more efficient than conventional moment estimators, potentially even more accurate than Maximum Likelihood estimators for some smaller samples.

Generalized PWM (GPWM) methods have been suggested. We mention the GPWM estimators proposed by Diebolt et al. (2008) for the Generalized Extreme Value distribution (GEVd), employed in analysing annual daily precipitation maxima (see “Extrema”), as prescribed by *Extreme Value (EV) Statistics* framework.

PWM-GEVd estimates are valid under the shape parameter condition $\gamma < 1$ (equivalently, $\mathbb{E}|X| < \infty$); the goal was to extend the method's validity. The GPWM $\nu_\omega := \mathbb{E}[X\omega(F(X))]$, with convenient auxiliary functions $\omega(\cdot)$, written in terms of the GEVd's F three parameters $(\theta_1, \theta_2, \gamma)$. The GPWM estimators are $\hat{\nu}_{\omega_{ab}} := \int_0^1 \mathbb{F}_n^{-1}(u) \omega_{ab}(u) du$, with $\mathbb{F}_n^{-1}$, the empirical quantile function.

Solving the moment-equations system $\nu_\omega(\theta_1, \theta_2, \gamma) = \hat{\nu}_{\omega_{ab}}$ for three suggested $\omega_{ab}(\cdot)$ functions results in the desired GPWM parameter estimators. These are asymptotically normal, have broader validity domain and improved performance over PWM, especially for large values of γ (common in hydrology and climatology), while conserving conceptual simplicity and easy implementation.

L-moments

Since 1986, when J.R.M. Hosking first introduced the *L-moments*, a myriad of papers has been published detailing and applying this system and its properties. The fundamental information, for the purpose of this exposition, is compiled in Hosking and Wallis (1997), where L-moments are a stepping-stone for several Regional Frequency Analysis (RFA) techniques.

Derived as *linear combinations* of PWM, *L-moments* inherit the appealing features mentioned above. Particularly, estimation-wise, robustness to measurement errors and sample variability, while accounting for ordering in weighting observations, suggests appropriateness of L-moments for

analysis of EVs, paramount to various fields (e.g., seismology).

Define the k^{th} *L-moment* of X as

$$\lambda_k := \int_0^1 \chi(u) \cdot \sum_{i=0}^{k-1} (p_{k,i}^* u^i) du, \quad (23)$$

where

$$p_{k,i}^* := (-1)^{k-i} \binom{k}{i} \binom{k+i}{i} = \frac{(-1)^{k-i} (k+i)!}{(i!)^2 (k-i)!}. \quad (24)$$

With respect to the PWM (17) and (18), redefine (23) as

$$\lambda_{k+1} := (-1)^k \sum_{i=0}^k p_{k,i}^* \alpha_i = \sum_{i=0}^k p_{k,i}^* \beta_i. \quad (25)$$

The complete L-moment set exists and uniquely determines F if $\mathbb{E}|X| < \infty$, which is not true for conventional moments.

L-moments resolved the PWM interpretability issue, as they meaningfully describe a distribution: explicitly, the first four L-moments are

$$\begin{aligned} \lambda_1 &= \alpha_0 &= \beta_0 \\ \lambda_2 &= \alpha_0 - 2\alpha_1 &= 2\beta_1 - \beta_0 \\ \lambda_3 &= \alpha_0 - 6\alpha_1 + 6\alpha_2 &= 6\beta_2 - 6\beta_1 + \beta_0 \\ \lambda_4 &= \alpha_0 - 12\alpha_1 + 30\alpha_2 - 20\alpha_3 &= 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0 \end{aligned} \quad (26)$$

where $\lambda_1 \in \mathbb{R}$ and $\lambda_2 \geq 0$ are the *L-location* and *L-scale*; for dimensionless, scale-independent measures of shape, consider the *L-moment ratios*, $\tau_r = \lambda_r / \lambda_2$, $r = 3, 4, \dots$, with highlight to the *L-skewness* and *L-kurtosis*

$$\tau_3 = \frac{\lambda_3}{\lambda_2} \quad \text{and} \quad \tau_4 = \frac{\lambda_4}{\lambda_2}. \quad (27)$$

Moreover, the *L-CV* $\tau = \lambda_2 / \lambda_1$ is akin to c in (6).

L-moment ratios are easier to interpret, as they satisfy $|\tau_r| < 1$ for finite-mean distributions, and, again, all odd-order τ_r of symmetric distributions are zero; also τ_4 is bounded relatively to τ_3 by

$$\frac{1}{4} (5\tau_3^2 - 1) \leq \tau_4 < 1. \quad (28)$$

The *L-statistics* are simply computable from linear combinations of the order statistics. *Sample L-moment* counterparts of (26) result from imputing the estimators (20) or (21) into the linear combinations in (25):

$$\begin{aligned} \ell_1 &= a_0 &= b_0 \\ \ell_2 &= a_0 - 2a_1 &= 2b_1 - b_0 \\ \ell_3 &= a_0 - 6a_1 + 6a_2 &= 6b_2 - 6b_1 + b_0 \\ \ell_4 &= a_0 - 12a_1 + 30a_2 - 20a_3 &= 20b_3 - 30b_2 + 12b_1 - b_0 \end{aligned} \quad (29)$$

with corresponding *sample L-skewness* and *sample L-kurtosis*

$$t_3 = \frac{\ell_3}{\ell_2} \quad \text{and} \quad t_4 = \frac{\ell_4}{\ell_2}. \quad (30)$$

These natural estimators enjoy attractive properties:

- ℓ_k are unbiased and t_r are asymptotically unbiased estimators, both being asymptotically normal if μ'_2 of X exists.
- ℓ_k behave nicely under linear transformations of data.
- Contrary to g and k in (9), t_3 and t_4 are not algebraically bounded w.r.t. sample size.
- t_3 and t_4 are generally less biased than g and k .
- The joint distribution of (t_3, t_4) is near-normal, a useful feature for some RFA procedures.

Alternative, less frequent, *plotting-position estimators* of L-moments can be found in Hosking and Wallis (1997).

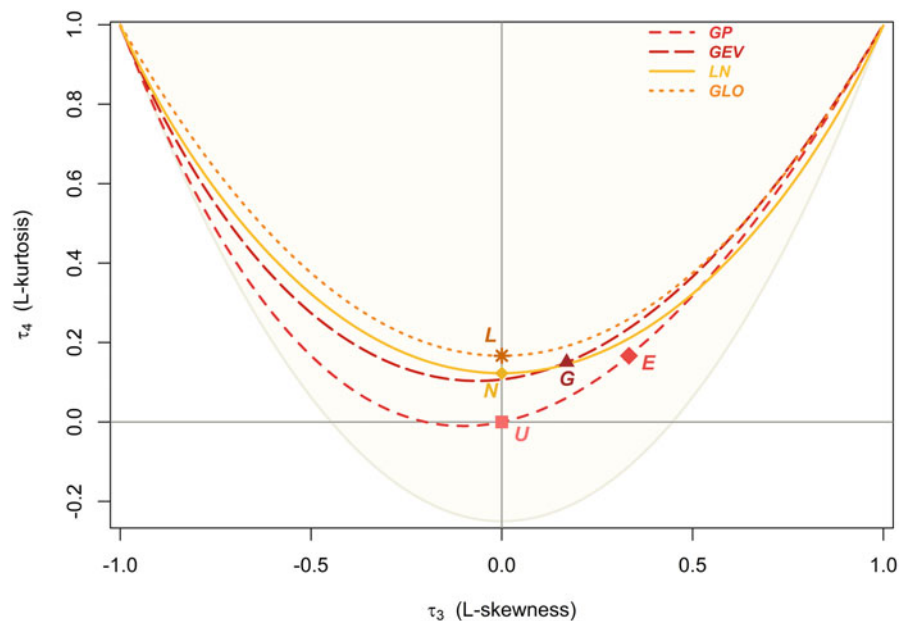
Again, it is possible to establish the “L-moment method” as a MM fitting approach. It follows the substitution principle introduced above, equating the first $m \geq p$ L-statistics to their population parallels. For brevity, we refer to the Appendices of Hosking and Wallis (1997), showing L-moment-derived expressions for common use distributions’ parameters, including the Lognormal (see “Lognormal Distribution”) and GEVd, relevant in several Geosciences areas.

Among the most significant contributions of L-moments is the identification of distributions given a reduced set of summary statistics. For this purpose, heavily featured in RFA, the *L-moment Ratio Diagram* (LMRD) was introduced. This visual tool presents the relationship between two L-moment ratios for specific distributions. Usually, L-skewness *versus* L-kurtosis is plotted, as shown in Fig. 1 for some families of interest; this is specially useful for skewed distributions, but higher order ratios could be chosen. A similar plot of conventional skewness *versus* kurtosis would be uninformative.

Notice that 2-parameter distribution families correspond to a point in the LMRD (e.g., Normal), 3-parameter ones draw a curve (e.g., GEVd), different points on the curve imply different shape parameter values, and families with more parameters appear as two-dimensional regions. Marking the estimates (t_3, t_4) from a given RS on the LMRD, and evaluating its distance to the several distribution-specific curves, may point to appropriateness of fitting one distribution over another; this has multiple practical applications, as we

Moments, Fig. 1 LMRD:

Shaded region – general bounds (28); key to distributions: E – Exponential; G – Gumbel; GEV – GEVd; GLO – Generalized Logistic; GP – Generalized Pareto; L – Logistic; LN – Lognormal; N – Normal; U – Uniform (cf. Hosking and Wallis 1997, Appendix A.12)



mention below. In practice, $0 \leq \tau_3 \leq 0.5$ and $0 \leq \tau_4 \leq 0.4$ define the commonly appropriate region of the LMRD.

Regional Frequency Analysis and Other Applications

The aim of the RFA approach suggested by Hosking and Wallis (1997), using an index-flow procedure, is to estimate quantiles of a frequency distribution (FD) at a given site by pooling data, or summary statistics, from sites judged to have similar FD. This is potentially problematic since FDs may not be exactly identical and magnitudes may be dependent across sites. Thus, steps are taken to guarantee accuracy and robustness of the analysis:

1. Screening the data – sites where data behave differently, needing closer inspection, are detected using a discordance measure built on L-moments.
2. Identifying and testing homogeneous regions – after grouping sites where the FDs are judged as approximately the same, L-moments are used in a heterogeneity measure to test if the region is acceptably close to homogeneous.
3. Choosing the FD – a goodness-of-fit L-moment-based measure, relating to the LMRD, is used to evaluate the fitting of candidate distributions to the data.
4. Estimating the FD – a regional L-moment algorithm is suggested for estimating regional and at-site quantiles.

L-moments clearly play a pivotal role at all stages of this process.

Additionally, we briefly mention recent papers with applications of L-moment methodologies to data of interest from Geosciences:

- Lee et al. (2021) have used the L-moment method to estimate the Gumbel distribution's parameters and related probabilities, while developing a new, EV theory-based approach of temporal probability assessment of future landslide occurrence from limited rainfall and landslide records.
- Silva Lomba and Fraga Alves (2020) developed an automatic procedure, based on the LMRD's Generalized Pareto-specific curve, for estimation of a crucial entity in EV Statistics, the threshold, showcased by accurate and efficient inference regarding significant wave height data.
- Papalexiou et al. (2020) evaluated the reliability of models' simulations for reproduction of historical global temperatures by comparing, based on L-moments and the LMRD (perceived as robust criteria), the distributional shape of historical/simulated temperatures, aggregated at different time scales.

Summary

As we have seen, common usefulness of moment-based methodologies includes data description, distribution identification, parameter estimation, dependence analysis, and further benefits could be mentioned (e.g., "Hypothesis Testing," "Autocorrelation"). We have explored advantages and drawbacks of three moment systems. The presented recent

applications to Geosciences show their appropriateness for diverse endeavors. Moreover, there is leeway for generalizations of the mentioned moments to be suggested (e.g., *Trimmed* L-moments). Moments are, in all, essential devices in Mathematical Geosciences.

Cross-References

- [Dimensionless Measures](#)
- [Expectation-Maximization Algorithm](#)
- [Frequency Distribution](#)
- [Geostatistics](#)
- [Monte Carlo Method](#)
- [Multivariate Analysis](#)
- [Random Variable](#)
- [Robust Statistics](#)
- [Shape](#)
- [Simulation](#)
- [Spatial Statistics](#)
- [Statistical Bias](#)
- [Statistical Computing](#)
- [Statistical Outliers](#)
- [Univariate](#)

Acknowledgments JSL and MIFA gratefully acknowledge support from Fundação para a Ciência e a Tecnologia, I.P. through project UIDB/00006/2020 and PhD grant SFRH/BD/130764/2017 (JSL).

Bibliography

- Diebolt J, Guillou A, Naveau P, Ribereau P (2008) Improving probability-weighted moment methods for the generalized extreme value distribution. *REVSTAT-Stat J* 6(1):33–50
- Dimitrakopoulos R, Mustapha H, Gloaguen E (2010) High-order statistics of spatial random fields: exploring spatial cumulants for modeling complex non-Gaussian and non-linear phenomena. *Math Geosci* 42(1):65–99
- Greenwood JA, Landwehr JM, Matalas NC, Wallis JR (1979) Probability weighted moments: definition and relation to parameters of several distributions expressible in inverse form. *Water Resour Res* 15(5):1049–1054
- Hosking JRM, Wallis JR (1997) *Regional frequency analysis: an approach based on L-moments*. Cambridge University Press, Cambridge
- Jekeli C (2011) Gravity field of the Earth. In: Gupta HK (ed) *Encyclopedia of solid earth geophysics*. Springer Netherlands, Dordrecht, pp 471–484
- Lee JH, Kim H, Park HJ, Heo JH (2021) Temporal prediction modeling for rainfall-induced shallow landslide hazards using extreme value distribution. *Landslides* 18:321–338
- Osterholt V, Dimitrakopoulos R (2018) Simulation of orebody geology with multiple-point geostatistics – application at Yandi channel iron ore deposit, WA, and implications for resource uncertainty. In: Dimitrakopoulos R (ed) *Advances in applied strategic mine planning*. Springer International Publishing, Cham, pp 335–352
- Papalexiou SM, Rajulapati CR, Clark MP, Lehner F (2020) Robustness of CMIP6 historical global mean temperature simulations: trends, long-term persistence, autocorrelation, and distributional shape. *Earth's Future* 8(10):e2020EF001667
- Pearson K (1893) Asymmetrical frequency curves. *Nature* 48(1252):615–616
- Rohatgi VK, Saleh AME (2015) *An introduction to probability and statistics*. Wiley, New Jersey
- Silva Lomba J, Fraga Alves MI (2020) L-moments for automatic threshold selection in extreme value analysis. *Stoch Env Res Risk A* 34(3):465–491
- Wu Y, Gao Y, Zhang L, Yang J (2020) How distribution characteristics of a soil property affect probabilistic foundation settlement – from the aspect of the first four statistical moments. *Can Geotech J* 57(4):595–607

Monte Carlo Method

Klaus Mosegaard

Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark

Definition

Any random process can, in principle, be represented by a computer program using pseudorandom numbers. If near-perfectly independent random numbers between 0 and 1 can be produced, appropriate functions of these numbers can simulate realizations of practically any probability distribution needed in technical and scientific applications. The development of Monte Carlo methods grew out of our ability to find those functions and algorithms.

The literature on Monte Carlo methods is vast, and the methods are many. We cannot cover everything here, but the intention is to provide a brief description of some of the most important techniques and their principles. After having reviewed basic techniques for random simulation, we shall move on to three high-level categories of Monte Carlo methods: testing, inference, and optimization. We shall also present some hybrid methods, as there is often no sharp borderline between the categories.

In the following, unless otherwise stated, we will consider continuous random variables over bounded (measurable) subsets of $\mathcal{R}^N$ and their probability densities, although many of the methods described are applicable more generally.

Methods

Basic Simulation Methods

Exact sampling of probability distributions is an important component of all Monte Carlo algorithms. Most important are the following methods (formulated for a 1 dimensional space):

Algorithm 1 The “Inverse Method” for a Continuous Random Variable with a Given, Everywhere Nonzero Probability Density

Let p be a probability density, and P its distribution function:

$$P(s) = \int_{-\infty}^s p(s)ds, \quad (1)$$

and let r be a random number chosen uniformly at random between 0 and 1. Then the random number x generated through the formula

$$x = P^{-1}(r) \quad (2)$$

has probability density p .

Algorithm 2 The Box-Muller Method for Gaussian Distributions

Let r_1 and r_2 be random numbers chosen uniformly at random between 0 and 1. Then the random numbers

$$x_1 = \sqrt{-2 \ln r_2} \cos(2\pi r_1) \quad (3)$$

$$x_2 = \sqrt{-2 \ln r_2} \sin(2\pi r_1) \quad (4)$$

are independent and Gaussian distributed with zero mean and unit variance.

As building blocks for some of the more sophisticated algorithms we shall discuss in the coming sections, we need to mention the following two algorithms for sampling in high-dimensional spaces:

Algorithm 3 Sequential Simulation

Consider a probability density p over an N -dimensional space decomposed into a product of N univariate conditional probability densities:

$$p(x_1, x_2, \dots, x_N) = p(x_1)p(x_2|x_1)p(x_3|x_2, x_1) \dots p(x_N|x_{N-1}, \dots, x_1). \quad (5)$$

Exact simulation of realizations from $p(x_1, x_2, \dots, x_N)$ can now be performed by first generating a realization $\hat{x}_1$ from $p(x_1)$, then a realization $\hat{x}_2$ from $p(x_2|\hat{x}_1)$, then a realization $\hat{x}_3$ from $p(x_3|\hat{x}_2, \hat{x}_1)$, and so on until we have a realization $\hat{x}_N$ from $p(x_N|\hat{x}_N, \dots, \hat{x}_1)$. The point $(\hat{x}_1, \dots, \hat{x}_N)$ is now a realization from $p(x_1, \dots, x_N)$.

The above type of algorithm is widely used for simulation of image realizations where the conditional probabilities are obtained from training images (see Strebel 2002).

The simplest way to sample a constant probability density is through a simple application of Algorithm (3) where all 1D conditional probability densities are constant functions. However, in the Markov Chain Monte Carlo algorithms we shall describe below, we need to perform uniform sampling via random walks whose design can be adapted to our sampling problem. This can be done in the following way:

Algorithm 4 Sampling a Constant Probability Density by a Random Walk

Given a uniform (constant) probability density, a sequence of random numbers $U_n (n = 1, 2, \dots)$ in $[0, 1]$, and a PROPOSAL DISTRIBUTION $q(\mathbf{x}^{(n+1)}|\mathbf{x}^{(n)})$ defining the probability that the random walk goes to $\mathbf{x}^{(n+1)}$, given that it starts in $\mathbf{x}^{(n)}$ (possibly with $q(\mathbf{x}^{(n+1)}|\mathbf{x}^{(n)}) \neq q(\mathbf{x}^{(n)}|\mathbf{x}^{(n+1)})$), then the points visited by the iterative random function V :

$$V(\mathbf{x}^{(n)}) = \begin{cases} \mathbf{x}^{(n+1)} & \text{if } U \leq \min \left[1, \frac{q(\mathbf{x}^{(n)}|\mathbf{x}^{(n+1)})}{q(\mathbf{x}^{(n+1)}|\mathbf{x}^{(n)})} \right] \\ \mathbf{x}^{(n)} & \text{otherwise} \end{cases} \quad (6)$$

asymptotically converge to a sample of the uniform distribution as n goes to infinity.

The above algorithms (1)–(4) are easily demonstrated and straightforward to use in practice. Algorithms (1) and (2) can without difficulty be generalized to higher dimensions.

Monte Carlo Testing

If we are interested in testing the validity of a parametric statistical model of, say, an inhomogeneous earth structure, we can select an appropriate function (a “statistic”) summarizing important properties of the structure, for instance, a variogram. We can now generate Monte Carlo realizations of the structure and compute values of the statistic for each realization. By comparing the distribution of the statistics with the observed (or theoretical) value of the statistics, we can decide if our choice of model is acceptable (Mrkvicka et al. 2016). The decision relies on a statistical test measuring to what extent the observed statistics coincides with high probabilities in the histogram of simulated values.

The above testing approach can be seen as a tool to determine acceptable models satisfying given observations. This is related, but in principle more general, than the Monte Carlo inference methods described in the following. In these methods, the statistical model is fixed beforehand.

Monte Carlo Inference

Inference problems are problems where we seek parameters of a given, underlying stochastic model, designed to explain observable data $\mathbf{d}$. In physical sciences, technology, and economics, inference problems are often *inverse problems* which, in a probabilistic formulation, result in the definition of a *posterior* probability distribution $p(\mathbf{x}|\mathbf{d})$ describing our state of information about model parameters $\mathbf{x}$ after inference. Integrals of p over the parameter space can then provide estimates of event probabilities, expectations, covariances, etc. Such integrals are of the form

$$I = \int_{\chi} h(\mathbf{x})p(\mathbf{x})d\mathbf{x}, \quad (7)$$

where $p(\mathbf{x})$ is the posterior probability density. They can be evaluated numerically by generating a large number of random, independent realizations $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}$ from $p(\mathbf{x})$. The sum

$$I \approx \frac{1}{N} \sum_{n=1}^N h(\mathbf{x}^{(n)}). \quad (8)$$

is then an estimate of the integral. The fact that the integral (7) in high-dimensional spaces is more efficiently calculated by Monte Carlo algorithms than by any other method means that Monte Carlo integration has become important in numerical analysis.

For many practical inference problems, the model space is so vast, and evaluation of the probability density $p(\mathbf{x})$ is so computer intensive, that full information about $p(\mathbf{x})$ remains unavailable. In this case, we need algorithms that can do with values of $p(\mathbf{x})$ from one point at a time. The simplest of these algorithms is the rejection algorithm (von Neumann 1951):

The Rejection Algorithm

Algorithm 5 The Rejection Algorithm

Assume that $p(\mathbf{x})$ is a probability density with an upper bound $M \geq \max(p(\mathbf{x}))$. In the n th step of the algorithm, choose with uniform probability a random candidate point $\mathbf{x}_c$. Accept $\mathbf{x}_c$ only with probability

$$p_{\text{accept}} = \frac{p(\mathbf{x}_c)}{M}. \quad (9)$$

The set of thus accepted candidate points is a sample from the probability distribution $p(\mathbf{x})$.

The rejection algorithm is ideal in the sense that it generates independent realizations from bounded probability densities, but for distributions in high-dimensional spaces with vast areas of negligible probability, the number of accepted points is small. This makes the algorithm very inefficient, and for this reason random-walk-based algorithms were developed. The original version of this was the Metropolis Algorithm (Metropolis et al. 1953), but it was later improved by Hastings (1970):

Markov-Chain Monte Carlo (MCMC)

Algorithm 6 (Metropolis-Hastings).

Given a (possibly un-normalized) probability density $p(\mathbf{x}) > 0$, a sequence of random numbers $U_n (n = 1, 2, \dots)$ in $[0,1]$, and an iterative random function $V(\mathbf{x})$ sampling a constant probability density using Algorithm (4):

$$\mathbf{x}^{(n+1)} = V(\mathbf{x}^{(n)}). \quad (10)$$

Then the distribution of samples produced by the iterative random function W :

$$\begin{aligned} \mathbf{x}^{(n+1)} &= W(\mathbf{x}^{(n)}) \\ &= \begin{cases} V(\mathbf{x}^{(n)}) & \text{if } U_n \leq \min \left[1, \frac{p(V(\mathbf{x}^{(n)}))}{p(\mathbf{x}^{(n)})} \right], \\ \mathbf{x}^{(n)} & \text{otherwise} \end{cases} \end{aligned} \quad (11)$$

will asymptotically converge to $p(\mathbf{x})$ as n goes to infinity.

An extended form of this algorithm is obtained if $V(\mathbf{x})$ is sampling an arbitrary probability density $r(\mathbf{x})$, in which case the algorithm will sample the product distribution $p(\mathbf{x})r(\mathbf{x})$. This was suggested by Mosegaard and Tarantola (1995) as a way of sampling the posterior distribution in Bayesian/probabilistic inverse problems, where $r(\mathbf{x})$ is the prior probability density and $p(\mathbf{x})$ is the likelihood. In this way, a closed-form expression for the prior was unnecessary, allowing V to be available only as a computer algorithm.

It is important to realize that the design of the proposal algorithm (4) is critical for the performance of the Metropolis-Hastings algorithm. The proposal distribution $q(\mathbf{x}^{(n+1)}|\mathbf{x}^{(n)})$ will work efficiently only if it locally resembles the sampling distribution $p(\mathbf{x})$ (except for a constant). In this way, the number of rejected moves will be minimized.

In practice, any sampling with the Metropolis-Hastings algorithm needs to run for some time to reach equilibrium

sampling with statistically stationary output. The time (number of iterations) needed to reach this equilibrium is called the *burn-in time*, and it is usually found by experimentation. Another problem is that the algorithm is based on a random walk, and hence, sample points will not be independent when sampled closely in time. To determine the time spacing between approximately independent samples, a correlation analysis may be needed. Ways of avoiding poor sampling results were discussed by Hastings (1970).

Geman and Geman (1984) introduced the *Gibbs Sampler*, another variant of the Metropolis algorithm where sample points were picked along coordinate axes in the parameter space according to the conditional distribution $p(x_j|x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_N)$, x_j being the parameter to be perturbed. Although this strategy gives acceptance probabilities equal to 1, it may not be practical when calculation of the conditional probability is computationally expensive.

Monte Carlo Inference for Sparse Models

The Metropolis-Hastings algorithm (6) can be formulated to allow variable dimension of the parameter vector $\mathbf{x}$ (Green 1995). This version of the algorithm is used in Bayesian inference when the number of parameters N is treated as an unknown (Sambridge et al. 2006), making it possible to work with sparse models, and thereby reducing the computational burden:

Algorithm 7 Reversible-Jump Monte Carlo

Let $p(\mathbf{x}, N)$ be a probability density over a subset of $\mathcal{R}^N$, augmented with its dimension parameter N . The Metropolis-Hastings algorithm (6) operating in the extended space with

$$V(\mathbf{x}^{(n)}, N^{(n)}) = \begin{cases} (\mathbf{x}^{(n+1)}, N^{(n+1)}) & \text{if } U_n \leq \min \left[1, \frac{q(\mathbf{x}^{(n)}, N^{(n)} | \mathbf{x}^{(n+1)}, N^{(n+1)})}{q(\mathbf{x}^{(n+1)}, N^{(n+1)} | \mathbf{x}^{(n)}, N^{(n)})} |J| \right], \\ (\mathbf{x}^{(n)}, N^{(n)}) & \text{otherwise} \end{cases} \quad (12)$$

where the proposal distributions $q(\mathbf{x}^{(q)} | \mathbf{x}^{(p)})$ and $q(\mathbf{x}^{(p)} | \mathbf{x}^{(q)})$ are expressed through the invertible, differential mappings:

$$\begin{aligned} \mathbf{x}^{(q)} &= h(\mathbf{x}^{(p)}, \mathbf{u}^{(p)}) \\ \mathbf{x}^{(p)} &= h'(\mathbf{x}^{(q)}, \mathbf{u}^{(q)}), \end{aligned} \quad (13)$$

$\mathbf{u}^{(p)}$ and $\mathbf{u}^{(q)}$ being random vectors with generally nonuniform distributions, and where $|J|$ is the Jacobian

$$|J| = \left| \frac{\partial(\mathbf{x}^{(p)}, \mathbf{u}^{(p)})}{\partial(\mathbf{x}^{(q)}, \mathbf{u}^{(q)})} \right|, \quad (14)$$

will sample the extended space according to $p(\mathbf{x}, N)$.

This algorithm will not only sample the space of models $\mathbf{x}$ (with varying dimension), but also the distribution N of the dimension itself.

Sequential Monte Carlo Methods

Estimating the evolution of a dynamic system is important in science and technology, for instance, in predicting the movement of an airplane in motion, or in numerical weather prediction. An important analytical tool in this field is the Kalman filter (Kalman 1960) which iteratively provides

deterministic, Bayesian, and linear updates to past predictions of system parameters. For large systems, however, the many parameters make the analytical handling of large covariance matrices computationally intractable. This led to the development of the ensemble Kalman filter (EnKF) (Evensen 1994, 2003) a Monte Carlo algorithm that avoids the covariance matrices and instead represents the Gaussian distributions of noise, prior and posterior by ensembles of realizations. A basic formulation is the following:

Algorithm 8 Ensemble Kalman Filter (EnKF)

Assume that data $\mathbf{d}$ at any time are related to the system parameters $\mathbf{x}$ through the linear equation $\mathbf{d} = \mathbf{H}\mathbf{x}$. If $\mathbf{X}$ is an $n \times N$ matrix of N realizations from the initial (prior) Gaussian distribution of system parameters, and if $\mathbf{D}$ is an $m \times N$ matrix with N copies of data $\mathbf{d} + \mathbf{e}_m$ where $\mathbf{d}$ is the noise-free data and each $\mathbf{e}_m$ is a realization of the Gaussian noise, it can be shown that if $\mathbf{K} = \mathbf{CH}^T(\mathbf{HCH}^T + \mathbf{R})^{-1}$, the columns of the matrix

$$\hat{\mathbf{X}} = \mathbf{X} + \mathbf{K}(\mathbf{D} - \mathbf{H}\mathbf{X}) \quad (15)$$

are realizations from the posterior probability density. Iterative application of these rules, where the posterior realizations in one time step are used as prior realizations in the next step, will approximately evolve the system in time according to the model equations and the uncertainties.

For more details about the implementation, see Evensen (2003).

A more recent, nonlinear, and non-Gaussian alternative to EnKF is the Particle Filter method (Reich and Cotter (2015), Nakamura and Potthast (2015) and van Leeuwen et al. (2015)). The standard particle filter is a bootstrapping technique (see below) operating in the following way:

Algorithm 9 Particle Filter

Assume that data $\mathbf{d}$ at any time are related to the system parameters $\mathbf{x}$ through the equation $\mathbf{d} = \mathbf{h}(\mathbf{x})$. After n time steps, let $\mathbf{x}_1^{(n)}, \mathbf{x}_2^{(n)}, \dots, \mathbf{x}_N^{(n)}$ be N initial realizations (“particles”) of the system parameters, representing the prior probability density

$$p(\mathbf{x}^{(n)}) = \sum_{i=1}^N \frac{1}{N} \delta(\mathbf{x}^{(n)} - \mathbf{x}_i^{(n)}). \quad (16)$$

Assume that the system is propagated forward in time by the generally nonlinear model f :

$$\mathbf{x}^{(n)} = f(\mathbf{x}^{(n-1)}) + \mathbf{e}^{(n)} \quad (17)$$

where $\mathbf{e}^{(n)}$ is modelization noise introduced in the n 'th time step. If the likelihood function for time step n is

$$p(\mathbf{d}^{(n)} | \mathbf{x}^{(n)}) = p_e(\mathbf{d} - \mathbf{h}(\mathbf{x}^{(n)})) \quad (18)$$

We can use Bayes' rule and obtain

$$\begin{aligned} p(\mathbf{x}^{(n)} | \mathbf{d}^{(n)}) &= p(\mathbf{d}^{(n)} | \mathbf{x}^{(n)}) \frac{p(\mathbf{x}^{(n)})}{p(\mathbf{d}^{(n)})} \\ &\approx \sum_{i=1}^N w_i^{(n)} \delta(\mathbf{x}^{(n)} - \mathbf{x}_i^{(n)}) \end{aligned} \quad (19)$$

where

$$w_i^{(n)} = \frac{p(\mathbf{d}^{(n)} | \mathbf{x}_i^{(n)})}{\sum_k p(\mathbf{d}^{(n)} | \mathbf{x}_k^{(n)})}. \quad (20)$$

Iterative application of these rules, where the posterior realizations in one time step are used as prior realizations in the next step, will approximately evolve the system in time according to the model equations and the uncertainties.

Resampling (sampling with replacement from the ensemble) is often used before each time step to obtain more equally weighted, but possibly duplicated particles.

Empirical Monte Carlo

Assume that N realizations from an unknown probability distribution $p(\mathbf{x})$ are available, and that we, for some reason, are unable to obtain more samples. We are interested in inferring information about a parameter of $p(\mathbf{x})$, for example, the distribution of its mean. This can be accomplished by an empirical Monte Carlo technique, for example, the resampling method *Bootstrapping* (Efron 1993):

Algorithm 10 Bootstrapping

To infer information about a parameter η of the distribution $p(\mathbf{x})$ from a set of realizations $A = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, draw N elements from A with replacement (thereby allowing elements to be repeated) and compute η for the N elements. This process is repeated many times, each time obtaining a new value of η (if N is large). The normalized histogram for η of all these resampling experiments will be an approximation to the distribution of η .

Bootstrapping can in this way be used to evaluate biases, variances, confidence intervals, prediction errors, etc.

Monte Carlo Optimization

An important application of Monte Carlo algorithms is in the solution of complex optimization problems, for instance, in the location of the global minimum for an objective function $E(\mathbf{x})$. One of the first examples of a Monte Carlo optimizer was the following modification of the Metropolis-Hastings algorithm (6) (Kirkpatrick et al. 1983):

Simulated Annealing

Algorithm 11 (Simulated Annealing)

Given an objective function $E(\mathbf{x})$, a sequence of random numbers $U_n (n = 1, 2, \dots)$ in $[0, 1]$, a decreasing sequence of positive numbers (“temperature” parameters) $T_n \rightarrow 0$ for $n \rightarrow \infty$, and an iterative random function $V(\mathbf{x})$, sampling a constant probability density:

$$\mathbf{x}^{(n+1)} = V(\mathbf{x}^{(n)}). \quad (21)$$

Then the sample points produced by the iterative random function A :

$$\mathbf{x}^{(n+1)} = A(\mathbf{x}^{(n)}) = \begin{cases} V(\mathbf{x}^{(n)}) & \text{if } U_n \leq \min \left[1, \frac{\exp(-E(V(\mathbf{x}^{(n)}))/T_n)}{\exp(-E(\mathbf{x}^{(n)})/T_n)} \right], \\ \mathbf{x}^{(n)} & \text{otherwise} \end{cases} \quad (22)$$

asymptotically converge to the minimum of $E(x)$ as n goes to infinity.

This algorithm was inspired by the process of chemical annealing, where a crystalline material is slowly cooled ($T \rightarrow 0$) from a high temperature through its melting point, resulting in the formation of highly ordered crystals with low lattice energy E . In each step of the algorithm, thermal fluctuations in the system are simulated by the Monte Carlo algorithm (for applications, see Mosegaard and Vestergaard (1991) and Deutsch and Journal (1994)).

It is seen that, for constant temperature parameter T , the above algorithm is actually a Metropolis-Hastings algorithm designed to sample the Gibbs-Boltzmann distribution (see Mosegaard and Vestergaard 1991)

$$P_B(\mathbf{m}) = \frac{\exp\left(-\frac{E(\mathbf{m})}{T}\right)}{Z(T)} \quad (23)$$

Monte Carlo Optimization-Sampling Hybrids

The use of MCMC algorithms for inference often runs into problems in high-dimensional parameter spaces. When the target probability/objective function has multiple optima, MCMC algorithms show a critical slowing-down often making the analysis computationally intractable. There exist a few methods to alleviate this problem. Common to these methods is to use a “temperature” parameter to artificially increase (or vary) the noise to broaden the sampling density to hinder entrapment in local optima during sampling.

Parallel Tempering

One method is a simple modification of simulated annealing, where $E(\mathbf{x}) = -\ln(p(\mathbf{x}))$, and the decrease of T is stopped at $T = 1$ (see Salamon et al. (2002)). An inherent problem with simulated annealing is, however, the risk of entrapment in local minima for $E(\mathbf{x})$. The *parallel tempering* algorithm (Geyer 1991; Sambridge 2013), which is an improvement of an older algorithm *simulated tempering* (Marinari and Parisi 1992; Sambridge 2013), seeks to avoid the entrapment

problem by skipping the forced decrease of the “temperature” parameter T . The idea is to operate on an ensemble of models, and to include T in the parameters to be sampled:

Algorithm 12 (Parallel Tempering)

Given an ensemble $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_K)$ of K models, and their temperatures $T_1, T_2, \dots, T_K$, all distributed according to the same probability density p over the augmented spaces $(\mathbf{x}_k, T_k)$, $k = 1, 2, \dots, K$. The Metropolis-Hastings algorithm (6), with the special rule for V that any perturbation of T consists of swapping values of T for two random ensemble members, asymptotically samples the joint distribution $\prod_k p(\mathbf{x}_k, T_k)$, and the sample of ensemble members with $T_k = 1$ approaches a sample from the original target distribution $p(\mathbf{x})$.

If positive values of T close to $T^* \approx 0$ are allowed in this algorithm, the ensemble members with $T_k = T^*$ will be located close to the maxima for $p(\mathbf{x})$. In this way, parallel tempering can be used for optimization in a way similar to simulated annealing.

Other Approaches

Biological systems evolve by using a large population to explore many options in parallel, and this is the source of inspiration for evolutionary algorithms. An example of this is *genetic algorithms* (Holland 1975) where an ensemble $\mathbf{x}_1, \dots, \mathbf{x}_K$ of K sample points is assigned a probability $p(\mathbf{x}_k)$ of survival (“fitness”). In a simple implementation, the population is initially generated randomly, but at each iteration it is altered by the action of three operators: selection, crossover, and mutation. *Selection* randomly resamples the ensemble according to $p(\mathbf{x}_k)$, thereby typically duplicating ensemble members with high fitness at the expense of members with low fitness, improving the average fitness of the ensemble. The *crossover* step exchanges parameter values between two random members, and the *mutation* step randomly changes a parameter of one member. Crossover and mutation attempts are only accepted according to predefined probabilities. Iterative application of the above three steps allows the algorithm to asymptotically converge to a population with high fitness values $p(\mathbf{x}_k)$.

Another widely used technique is the neighborhood algorithm (Sambridge 1999a, b) where sample points are iteratively generated from a *neighborhood approximation* to the target distribution $p(\mathbf{x})$. The approximation is a piecewise constant function over Voronoi cells, centered at each of the previous sample points. The approximation to the target distribution is updated in each iteration, concentrating the sampling in multiple high-probability regions. The method will generate a distribution of points biased toward the maxima of $p(\mathbf{x})$.

Neither of the two algorithms above produces samples of the probability distribution p , but if supplemented with appropriate resampling of the output, an approximate sample from p may be produced.

Summary

We have given an overview of the use of Monte Carlo techniques for inference and optimization. A main theme behind the development of many methods has been to reduce the computational workload required to solve large-scale problems. Several ways of dealing with this challenge were discussed, including Markov-Chain Monte Carlo strategies, the use of sparse models, improving numerical forecast schemes through nonparametric representations of probability distributions, and even improving ideas inherited from early works on simulated annealing. Research in Monte Carlo methods is still a prolific area, and there is no doubt that many interesting developments are waiting ahead.

Cross-References

- [Bayes's Theorem](#)
- [Bayesian Inversion in Geoscience](#)
- [Bootstrap](#)
- [Computational Geoscience](#)
- [Ensemble Kalman Filtering](#)
- [Genetic Algorithms](#)
- [Geostatistics](#)
- [Inversion Theory](#)
- [Inversion Theory in Geoscience](#)
- [Markov Chain Monte Carlo](#)
- [Markov Chains: Addition](#)
- [Multiple Point Statistics](#)
- [Optimization in Geosciences](#)
- [Particle Swarm Optimization in Geosciences](#)
- [Realizations](#)
- [Sampling Importance: Resampling Algorithms](#)
- [Sequential Gaussian Simulation](#)
- [Simulated Annealing](#)
- [Statistical Computing](#)
- [Uncertainty Quantification](#)
- [Very Fast Simulated Reannealing](#)

References

- Deutsch CV, Journel AG (1994) Application of simulated annealing to stochastic reservoir modeling. *SPE Adv Technol Ser* 2:222–227
- Efron B (1993) An introduction to the bootstrap. Chapman & Hall/CRC
- Evensen G (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J Geophys Res* 99(C5):10,143–10,162. <https://doi.org/10.1029/94JC00572>
- Evensen G (2003) The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dyn* 53(4):343–367. <https://doi.org/10.1007/s10236-003-0036-9>
- Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell* 6(6):721–741. <https://doi.org/10.1109/TPAMI.1984.4767596>
- Geyer CJ (1991) Markov chain Monte Carlo maximum likelihood. Interface Foundation of North America
- Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82(4):711–732. <https://doi.org/10.1093/biomet/82.4.711>
- Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57(1):97–109
- Holland JH (1975) Adaptation in natural and artificial systems. University of Michigan Press
- Kalman RE (1960) A new approach to linear filtering and prediction problems. *J Basic Eng* 82(1):35–45. <https://doi.org/10.1115/1.3662552>
- Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220(4598):671–680. <https://doi.org/10.1126/science.220.4598.671>
- Marinari E, Parisi G (1992) Simulated tempering: A new Monte Carlo scheme. *Europhys Lett* 19(6):451–458. <https://doi.org/10.1209/0295-5075/19/6/002>
- Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953) Equations of state calculations by fast computing machines. *J Chem Phys* 21(6):1087–1092
- Mosegaard K, Tarantola A (1995) Monte Carlo sampling of solutions to inverse problems. *J Geophys Res* 100:12431–12447
- Mosegaard K, Vestergaard PD (1991) A simulated annealing approach to seismic model optimization with sparse prior information. *Geophys Prosp* 39(05):599–612
- Mrkvička T, Soubeyrand S, Myllymäki M, Grabarnik P, Hahn U (2016) Monte Carlo testing in spatial statistics, with applications to spatial residuals. *Spatial Stat* 18:40–53, *spatial Statistics Avignon: Emerging Patterns*
- Nakamura G, Potthast R (2015) Inverse modeling. IOP Publishing, Bristol
- Reich S, Cotter C (2015) Probabilistic forecasting and Bayesian data assimilation. Cambridge University Press, Cambridge, UK
- Salamon P, Sibani P, Frost R (2002) Facts, conjectures, and improvements for simulated annealing. Society of Industrial and Applied Mathematics
- Sambridge M (1999a) Geophysical inversion with a neighbourhood algorithm – i. searching a parameter space. *Geophys J Int* 138(2): 479–494. <https://doi.org/10.1046/j.1365-246X.1999.00876.x>
- Sambridge M (1999b) Geophysical inversion with a neighbourhood algorithm – ii. appraising the ensemble. *Geophys J Int* 138(3): 727–746. <https://doi.org/10.1046/j.1365-246x.1999.00900.x>
- Sambridge M (2013) A parallel tempering algorithm for probabilistic sampling and multimodal optimization. *Geophys J Int* 196(1): 357–374. <https://doi.org/10.1093/gji/ggt342>
- Sambridge M, Gallagher K, Jackson A, Rickwood P (2006) Trans-dimensional inverse problems, model comparison and the evidence. *Geophys J Int* 167(2):528–542. <https://doi.org/10.1111/j.1365-246X.2006.03155.x>
- Strebel S (2002) Conditional simulation of complex geological structures using multiple-point statistics. *Math Geol* 34:1–21
- van Leeuwen P, Cheng Y, Reich S (2015) Nonlinear data assimilation. Springer, Berlin
- von Neumann J (1951) Various techniques used in connection with random digits. Monte Carlo methods. *Nat Bureau Standards* 12: 36–38

Moran's Index

Giuseppina Giungato¹ and Sabrina Maggio²

¹Department of Economics, Section of Mathematics and Statistics, University of Salento, Lecce, Italy

²Dipartimento di Scienze dell'Economia, Università del Salento, Lecce, Italy

Synonyms

Moran I; Moran index; Moran's coefficient; Moran's I; Moran's ratio

Definition

Moran's index is a measure of spatial autocorrelation.

Introduction to Spatial Correlation Measures

Generally, the grade to which attributes or objects at some location on the earth's area are analogous to other attributes or objects in neighboring places is called spatial autocorrelation. The concept which Tobler has defined as the "first law of geography: everything is related to everything else, but near things are more related than distant things" (Tobler 1970), reflected the existence of the spatial autocorrelation.

As the name indicates, autocorrelation is the correlation between two variables of the same random field. If Z is the attribute observed in the plane, then the term spatial autocorrelation refers to the correlation between the same attribute at two spatial locations (Schabenberger and Gotway 2005).

Spatial analysis treats two quite different types of information. A first type of information is represented by the attributes of spatial characteristics, which comprise both measures (e.g., rainfall or population) and qualitative variables (e.g., name or type of soil). The other type of information is represented by the location of each spatial feature, which can be described by different geographical references or coordinate systems or by its position on a map (Goodchild 1986).

Spatial autocorrelation offers information about a phenomenon distributed in space, which can result essential for a right interpretation of the same phenomenon and which is not obtainable through other types of statistical analysis.

Moreover, in searching for a specific spatial distribution, it often occurs to find a variable which explains a pattern, but not completely. In this case, in order to identify any other variables that could be useful in explaining the remaining variation, the next stage is to examine the spatial model of residuals.

From the eventual absence of spatial autocorrelation in the residuals, it would be possible to deduce that (although the explanation of the model is not perfect) there would be little advantage in looking for further explicative variables, as they would be probably difficult and maybe singular to every sampling location.

As mentioned above, spatial autocorrelation is about comparing two types of information: similarity of location and similarity among attributes. The computation of the spatial proximity depends on the type of objects (points, areas, lines, rasters), while similarity among attributes can be measured in ways which depend on the type of data (interval, ordinal, nominal).

In the continuation, the following notation will be utilized:

- n : number of sampled objects;
- i, j : any two of the objects;
- w_{ij} : the similarity of i 's and j 's objects, where $w_{ii} = 0$ for all i ;
- z_i : the value of the attribute of interest for object i ;
- c_{ij} : the similarity of i 's and j 's attributes.

In general, the measures of spatial autocorrelation compare the set of attribute similarities c_{ij} with the set of locational similarities w_{ij} , combining them into a single index of the form $\sum_{i=1}^n \sum_{j=1}^n w_{ij} c_{ij}$, with properties which lead to quick interpretation.

Various methods have been proposed for estimating spatial proximity in order to create a suitable matrix of w_{ij} . Since the probability that objects are connected is greater for close ones than for distant ones, the existence of a connection between objects is often considered a measure of similarity of location. For example, a simple, indirect binary indicator of spatial proximity is represented by a common boundary between areas.

After obtaining the distances d_{ij} between objects i and j , in order to define the weights w_{ij} some appropriate decreasing function can be used. For example, a negative exponential $w_{ij} = \exp(-bd_{ij})$ or a negative power $w_{ij} = d_{ij}^{-b}$ can be used. In both cases a large b generates a quick decline and a small b a slower one. Indeed b represents a parameter affecting the speed at which weight decreases with spatial distance.

Various modes to assess the similarity of attributes have also been suggested, which are appropriate to the involved types of attributes.

For nominal data the usual approach is to set c_{ij} to 1 if i and j have the same attribute value, and 0 otherwise.

For ordinal data, similarity is usually based on comparing the ranks of i and j , while for interval data both the squared difference $(z_i - z_j)^2$ and the product $(z_i - \bar{z})(z_j - \bar{z})$ are commonly used.

Moran's Index

As previously mentioned, Moran's index (Moran 1948) is a measure of spatial autocorrelation.

In the following, some contexts characterized by different types of data and objects are discussed, as well as the respective possibility to apply spatial autocorrelation measures to them.

Interval Data and Area Objects

In the framework of interval data and area objects, the attribute similarity measure utilized by Moran's index makes it equivalent to a covariance between the values of a pair of objects $c_{ij} = (z_i - \bar{z})(z_j - \bar{z})$, where $\bar{z}$ denotes the mean of the attribute z values and c_{ij} measures the covariance between the value of the variable at one place and its value at another.

The remaining terms in Moran's index are designed to constrain it to a fixed range:

$$I = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} c_{ij}}{s^2 \sum_{i=1}^n \sum_{j=1}^n w_{ij}} \quad (1)$$

where the w_{ij} terms represent the spatial proximity of i and j and s^2 denotes the sample variance $s^2 = \sum_{i=1}^n (z_i - \bar{z})^2 / n$.

If neighboring areas tend to have similar attributes, the Moran index is positive. Conversely, if they tend to have more dissimilar attributes than might be expected, the Moran index is negative. Finally it is approximately null when attribute values are distributed randomly and independently in space.

Moran's index can be used alternatively to Geary's index for the identical application framework (area objects and interval attributes).

For a variable measured on an interval scale, in Geary's index (Geary 1968), attribute similarity c_{ij} is calculated with the squared difference in value $c_{ij} = (z_i - z_j)^2$. In the original paper (Geary 1954) the similarity of location was calculated in a binary way. If i and j had a common border, the weights w_{ij} could be set equal to 1. In the opposite case, the weights w_{ij} could be set equal to 0. The other terms in Geary's index guarantee that extremes happen at definite points:

$$c = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} c_{ij}}{2\sigma^2 \sum_{i=1}^n \sum_{j=1}^n w_{ij}} \quad (2)$$

where σ^2 denotes the variance of the attribute z values, that is, $\sigma^2 = \sum_{i=1}^n (z_i - \bar{z})^2 / (n-1)$. The value of c will be greater when large values of w_{ij} , which coincide with pairs of areas in contact, correspond to large values of c_{ij} , or large differences in attributes. As in Moran's index, the w_{ij} terms measure the spatial proximity of i and j and can be calculated by any suitable method.

In most applications Geary's and Moran's indices are satisfactory in the same way. However, Moran's index has the advantage that its extremes reflect the instinctive concepts of positive and negative correlation. Conversely, the scale used by Geary's index is not very clear as summarized in Table 1.

To explain the sense of the terms, the step-by-step computation of Moran's index can be represented in the following Example 1, using binary weights, as described above.

Example 1 Given the example of data set in Fig. 1, the step-by-step calculation of Moran's I is presented in the following.

$$n = 4 \quad \begin{matrix} z_1 = 1 \\ z_2 = 2 \\ z_3 = 2 \\ z_4 = 3 \end{matrix} \quad w = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

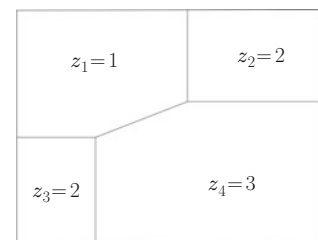
$$\bar{z} = \sum_{i=1}^n z_i / n = 8/4 = 2$$

Moran's Index, Table 1 Conceptual scales of spatial autocorrelation: correspondence between Geary's and Moran's indices

Conceptual scales	c index	I index
Similar	$0 < c < 1$	$I > 0^a$
Independent	$c = 1$	$I \simeq 0^a$
Dissimilar	$c > 1$	$I < 0^a$

^aThe correct expected value is $-1/(n-1)$ rather than 0

Moran's Index, Fig. 1 An example of data set for calculation of Moran's coefficient



$$c_{ij} = (z_i - \bar{z})(z_j - \bar{z}) \quad c = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}$$

$$\sum_{i=1}^n \sum_{j=1}^n w_{ij} c_{ij} = -2$$

$$s^2 = \sum_{i=1}^n (z_i - \bar{z})^2 / n = 2/4 = 0.5$$

$$I = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} c_{ij}}{s^2 \sum_{i=1}^n \sum_{j=1}^n w_{ij}} = -2/0.5 \cdot 10 = -0.4.$$

Interval Data and Other Object Types (Point, Line, and Raster)

In the context of interval attributes, both Moran's index and Geary's index can also be used for other types of objects (point, line, and raster), as long as a proper way can be designed for assessing the geographical closeness of pairs of objects.

First of all it should be noted that one manner of creating a suitable w_{ij} measure for area objects consists in replacing each area by positioned control point and calculating distances. Then, to define the weights some appropriate decreasing function can be used. For example, a negative exponential or a negative power can be used, as said earlier. This indicates a simple method for accommodating both indices to point objects.

Another option would be to replace point objects with areas, using a systemic process; for example, the creation of Thiessen or Dirichlet polygons (Boots 1986). This procedure splits the total survey region into polygons. Each polygon encircles a point and envelopes the surface which is nearer to that point than any other. Relating to this it is suggested to see a powerful efficient way to creating Thiessen polygons from a point set (Brassel and Reif 1979), as well as a case showing the application of this procedure to evaluate spatial autocorrelation for point objects (Griffith 1982). As already mentioned, the weights could be defined on the basis of the existence of the common boundary or its length. In effect, areas and points can be considered replaceable for both Moran's and Geary's indices.

When the application context is characterized by line objects, it is possible to distinguish two different situations.

In the first case, the lines can represent connections between nodes; thus it is necessary to calculate the spatial

autocorrelation existing in some attribute observed at the nodal points, and the c_{ij} are measures of the similarity of the attributes of each pair of nodes, and the w_{ij} are measures of the links between them. For instance, the weight w_{ij} can be set equal to 1 if a direct link exists between i and j and 0 otherwise, or be based on the link capacity or length.

In the other situation, it is necessary to determine the spatial autocorrelation existing in some attribute observed at the links (for instance, probability of motor vehicle accident, or transport cost per unit length). In this case, the weight w_{ij} represents a measure of proximity between two links and might be defined according to the spatial distance between the central points of two links, or on whether or not two links are in direct connection. Therefore Geary's and Moran's indices represent proper measures for the interval attributes, even in the case of line objects, provided appropriate methods can be proposed for defining the spatial proximity measures.

For rasters (lattices) one simple way of defining the weights is to set w_{ij} equal to 1 if i and j have a common border and 0 in the opposite case. In some works two cells are considered to be adjacent, also when they join at a corner. When handling square rasters, the previous two cases are sometimes referred to as the Rook or 4-neighbor case and the Queen or 8-neighbor case, respectively (Lloyd 2010).

Ordinal Data

Several documents have treated the identification of particular spatial autocorrelation indices for ordinal attributes, namely attributes with ordinal scale of measurement. In particular, some experts have dealt with explaining the spatial distribution of the Canadian settlement sizes (Coffey et al. 1982). They wanted to investigate whether big settlements tended to be encircled by little settlements or by other big ones, or whether the size of settlements was casually distributed. In order to make the attribute an ordinal measure restricted to integers from 1 to 5, the settlements have been assigned to one of the five dimensional classes.

The authors proposed two kinds of spatial autocorrelation measures. In the first approach, the ordinal size class data are treated as having range properties, considering that the computation of Geary's and Moran's indices requires the difference between z values. Indeed, both the indices were calculated directly from the size classes by setting z_i equal to an integer between 1 and 5. The w_{ij} terms were set equal to 1 for pairs of settlements sharing a direct road link and 0 in the opposite case.

In the other approach, the set of measures is built on the recorded number of connections between settlements of different dimensional classes, and the ordinal data are treated as if they were merely nominal. For example, n_{23} would represent the count of direct road connections recorded between a

settlement belonging to class 2 and a settlement belonging to class 3. Then, the recorded number of every type of link was confronted with the number predicted, assuming a random distribution for settlement sizes. This type of measure is the issue of the following section.

On the other hand, some works described the index more directly suitable for ordinal data (Royaltey et al. 1975). Each object is given a rank based on its ordinal value, and the c_{ij} are then built on the absolute difference between ranks for each pair. The index was implemented to a set of points, using a procedure which created the matrix of weights called adjacency matrix of the Gabriel graph, obtained setting the weights equal to 1, if no other points were inside a circumference constructed with the pair of points as diameter and 0 alternatively (Gabriel and Sokal 1969). In addition, there was a further generalization of the index (Hubert 1978), and other indices for ordinal data, ever built on ranks, were discussed (Sen and Soot 1977).

Nominal Data

In the context of nominal attributes and for any type of objects (areas, points, lines, or raster) it is appropriate to think of them as a set of objects on the map colored using a limited collection of colors. If k indicates the number of possible attribute classes, the spatial distributions of the nominal data are usually referred to as k -color maps. Thus, a two nominal classes attribute could be imagined as a model of black and white, as good as a three nominal classes attribute could be viewed as a three-color distribution (red, yellow, and blue).

In dealing with nominal data the ways in which they can be compared represent a strong constraint, since no measures of difference are possible. The characteristics of the nominal data permit just two possibilities: two attributes can be the same or different. In constructing indices, then, the c_{ij} can take only one of two values.

In this context, many of the proposed measures are built on join count statistics, by defining w_{ij} in a binary way. Much of early work was in the context of raster data, in which if two cells have a common border, they can be considered as joined. Then the count of times that a color s cell results joined to a color t cell is defined as the join count between color s and color t (indicated with n_{st}). Moreover, in dealing with joins between cells having the same color, it is necessary to avoid double counting. Spatial autocorrelation measures can be based on join count statistics, as these reflected the spatial arrangement of colors. When the distribution exhibits positive autocorrelation the incidence of joins of the same color will be lower than would be expected assuming a casual distribution of colors. In the same way, if the distribution exhibits negative autocorrelation the incidence of joins between different colors will be higher than expected. Join count statistics are an easy

way to measure spatial model. However, they do not represent a summary index, and they are not similar to Geary's or Moran's measures.

The previous ideas can be extended from rasters to other types of objects, whenever analogous binary spatial proximity measures can be developed. Therefore, if two points share a Thiessen border, then they could be treated as united; if two lines share a common node, they could be considered adjacent; if two areas share a common border, they could be considered joined.

Conclusions

Spatial autocorrelation represents the correlation between the same attribute at two spatial locations, and it is about comparing two types of information: similarity among attributes and similarity of location. Moran's index is a measure of spatial autocorrelation for interval attributes and area objects. It uses the covariance between the value of the variable at one place and its value at another, as attribute similarity measure. Moran's index can be used alternatively to Geary's, which uses the squared difference in value, as attribute similarity measure for the same context (area objects and interval attributes). In most applications Geary's and Moran's indices are equally satisfactory. However, Moran's index has the advantage that its extremes reflect the instinctive concepts of positive and negative correlations. Conversely, the scale used by Geary's index is not very clear. In the context of interval attributes, in essence, both Moran's index and Geary's index can also be used for other types of objects (point, line, and raster), provided a proper way can be designed for assessing the geographical closeness of pairs of objects.

Several documents have treated the identification of particular spatial autocorrelation indices for ordinal attributes, and the authors proposed two kinds of spatial autocorrelation measures. In the first approach, the ordinal size class data are treated as having range properties, considering that the computation of Geary's and Moran's indices requires the differences between values. In the second approach, the ordinal data are treated as if they were merely nominal. On the other hand, some works described the index more directly suitable for ordinal data. Finally, for the nominal attributes, join count statistics represent an easy way to measure spatial model. However, they do not provide a summary index, and they are not similar to Geary's or Moran's measures.

Cross-References

- [Database Management System](#)
- [Geostatistics](#)
- [Random Variable](#)

- [Spatial Analysis](#)
- [Spatial Statistics](#)

Bibliography

- Boots BN (1986) Voronoi polygons. CATMOG 45. Geo Books, Norwich
- Brassel K, Reif D (1979) A procedure to generate Thiessen polygons. Geogr Anal 11:289–303
- Coffey W, Goodchild MF, MacLean LC (1982) Randomness and order in the topology of settlement systems. Econ Geogr 58:20–28
- Gabriel KR, Sokal RR (1969) A new statistical approach to geographic variation analysis. Syst Zool 18:259–278
- Geary RC (1954) The contiguity ratio and statistical mapping. Inc Stat 5: 115–141
- Geary RC (1968) The contiguity ratio and statistical mapping. In: Berry BJJ, Marble DF (eds) Spatial analysis: a reader in statistical geography. Prentice-Hall, Englewood Cliffs, pp 461–478
- Goodchild MF (1986) Spatial autocorrelation. Catmog 47. Geo Books, Norwich
- Griffith D (1982) Dynamic characteristics of spatial economic systems. Econ Geogr 58:177–196
- Hubert LJ (1978) Nonparametric tests for patterns in geographic variation: possible generalizations. Geogr Anal 10:86–88
- Lloyd C (2010) Spatial data analysis: an introduction for GIS users. Oxford University Press, Oxford
- Moran PAP (1948) The interpretation of statistical maps. J R Stat Soc Ser B 10:243–251
- Royaltey HH, Astrachan E, Sokal RR (1975) Tests for patterns in geographic variation. Geogr Anal 7:369–395
- Schabenberger O, Gotway CA (2005) Statistical methods for spatial data analysis: texts in statistical science, 1st edn. Chapman & Hall/CRC Press, Boca Raton
- Sen AK, Soot S (1977) Rank tests for spatial correlation. Environ Plan A 9:897–905
- Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. J Econ Geogr 46:234–240

Morphological Closing

Aditya Challa¹, Sravan Danda¹ and B. S. Daya Sagar²

¹Computer Science and Information Science, APPCAIR, BITS, Pilani, India

²Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, Karnataka, India

Definition

Morphological closing is one of the basic operators from the field of Mathematical Morphology (MM). It is also one of the simplest morphological filters which is used for constructing more complex filters. And this operator is dual to morphological opening (chapter-ref). In general, morphological closing is defined as a composition of a dilation followed by an erosion. So, in case of discrete binary images, if X denotes

the binary image and B denotes the structuring element, morphological closing is defined as

$$\phi_B(X) = \varepsilon_B(\delta_B(X)) = (X \oplus B) \ominus B \quad (1)$$

This definition can be reduced to

$$\phi_B(X) = \cap \{B_x | X \subseteq B_x^c\} \quad (2)$$

Equivalently, in case of gray-scale images, if f denotes the gray-scale image and g denotes the structuring element, the closing is defined as

$$\phi_g(f) = \varepsilon_g(\delta_g(f)) \quad (3)$$

Illustrations

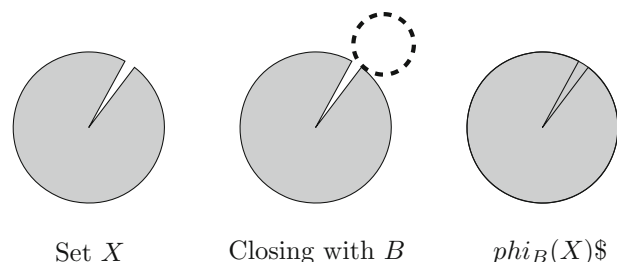
Morphological closing is one of the basic operators from the field of Mathematical Morphology. Its definition can be traced to fundamental works by Jean Serra (1983, 1988) and Georges Matheron (1975). In general, morphological closing is defined as a dilation operator composed with an erosion operator. In this entry, we illustrate the closing operator using binary images and discuss various properties. We also discuss the more general algebraic closings.

In simple words, (2) considers the intersection of all the translates of the structuring element whose complement would contain the original set. This is illustrated in Fig. 1. Here, the structuring element is taken to be a simple disk denoted by a dashed line in Fig. 1. All the points in the complement which do not fit within the structuring element are added to the original dataset. In Fig. 1, this corresponds to the spiked injection of the larger disk which gets filled.

Although not discussed here, similar intuition follows in gray-scale images as well (Dougherty and Lotufo 2003).

Some Important Properties

The properties of the closing operator are explained using the binary images. However, these properties easily extend to the



Morphological Closing, Fig. 1 Illustration of morphological closing. The shaded area refers to the set obtained by closing

gray-scale images as well. First, morphological closing is an *increasing operator*,

$$X \subseteq Y \Rightarrow \phi_B(X) \subseteq \phi_B(Y) \quad (4)$$

Moreover, it is also an *extensive operator*

$$X \subseteq \phi_B(X) \quad (5)$$

Also, the opening operator is idempotent

$$\phi_B(\phi_B(X)) = \phi_B(X) \quad (6)$$

Algebraic Closing

Morphological closing is a subset of a wider class of operators known as *Algebraic Closing*. An algebraic closing is defined as any operator on the lattice which is – (a) increasing, (b) extensive, and (c) idempotent. It is easy to see that morphological closing is a specific example of algebraic closing.

As an example of the algebraic closings which is not a morphological closing, consider the *Convex Closing* – take the smallest convex set which contains the given set. In Matheron (1975), it is shown that any algebraic closing can be defined as infimum of morphological closing.

Application to Filtering

Recall that, in the context of Mathematical Morphology, an operator which is increasing and idempotent is referred to as a *filter*. As stated earlier, morphological closing is one of the simplest filters based on which complex filtering operators such as top-hat transforms, granulometries, and alternating sequential filters are constructed, for instance, *Black top-hat* transform which is defined as

$$\text{BTH}(f) = \phi(f) - f \quad (7)$$

or simply $\text{BTH} = \phi - \mathbf{I}$ where $\mathbf{I}$ is the identity operator.

Summary

In this entry, morphological closing is defined for binary and gray-scale images. The practical utility of morphological closing is then illustrated using a simple binary image. The main properties of morphological closing namely increasing, extensiveness, and idempotence are then mentioned. A generalized notion of morphological closing, i.e., algebraic closing is defined. The entry is concluded by describing construction of more complex filter – black-top-hat transforms using morphological closing.

Cross-References

- [Mathematical Morphology](#)
- [Morphological Filtering](#)
- [Structuring Element](#)

Bibliography

- Dougherty ER, Lotufo RA (2003) Hands-on morphological image processing, vol 59. SPIE Press
- Matheron G (1975) Random sets and integral geometry. Wiley, New York
- Serra J (1983) Image analysis and mathematical morphology. Academic Press
- Serra J (1988) Image analysis and mathematical morphology, volume 2: theoretical advances. Academic Press, New York

Morphological Dilation

Sravan Danda¹, Aditya Challa¹ and B. S. Daya Sagar²

¹Computer Science and Information Systems, APPCAIR, Birla Institute of Technology and Science, Pilani, Goa, India

²Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, Karnataka, India

Definition

Morphological Dilation is one of the basic operators from the field of Mathematical Morphology (MM). Depending on the domain of application the definition varies. In the standard case of discrete binary images, given a binary image X , and a structuring element B the dilation of X is defined as

$$\delta_B(X) = \bigcup_{x \in X} B_x \quad (1)$$

where B_x denotes the structuring element translated by x . It is also a common practice to write $\delta_B(X) = X \oplus B$. Observe that the definition in (1) extends to continuous binary images as well.

In the case of gray-scale images, let f denote the gray-scale image and g denote the structuring function. Then, morphological dilation is defined by

$$\delta_g(f)(x) = \sup_{y \in E} \{f(y) + g(x - y)\} \quad (2)$$

where E denotes the domain of definition.

Both of the above definitions can be obtained from a general lattice-based definition of the dilation operator, as discussed in this entry.

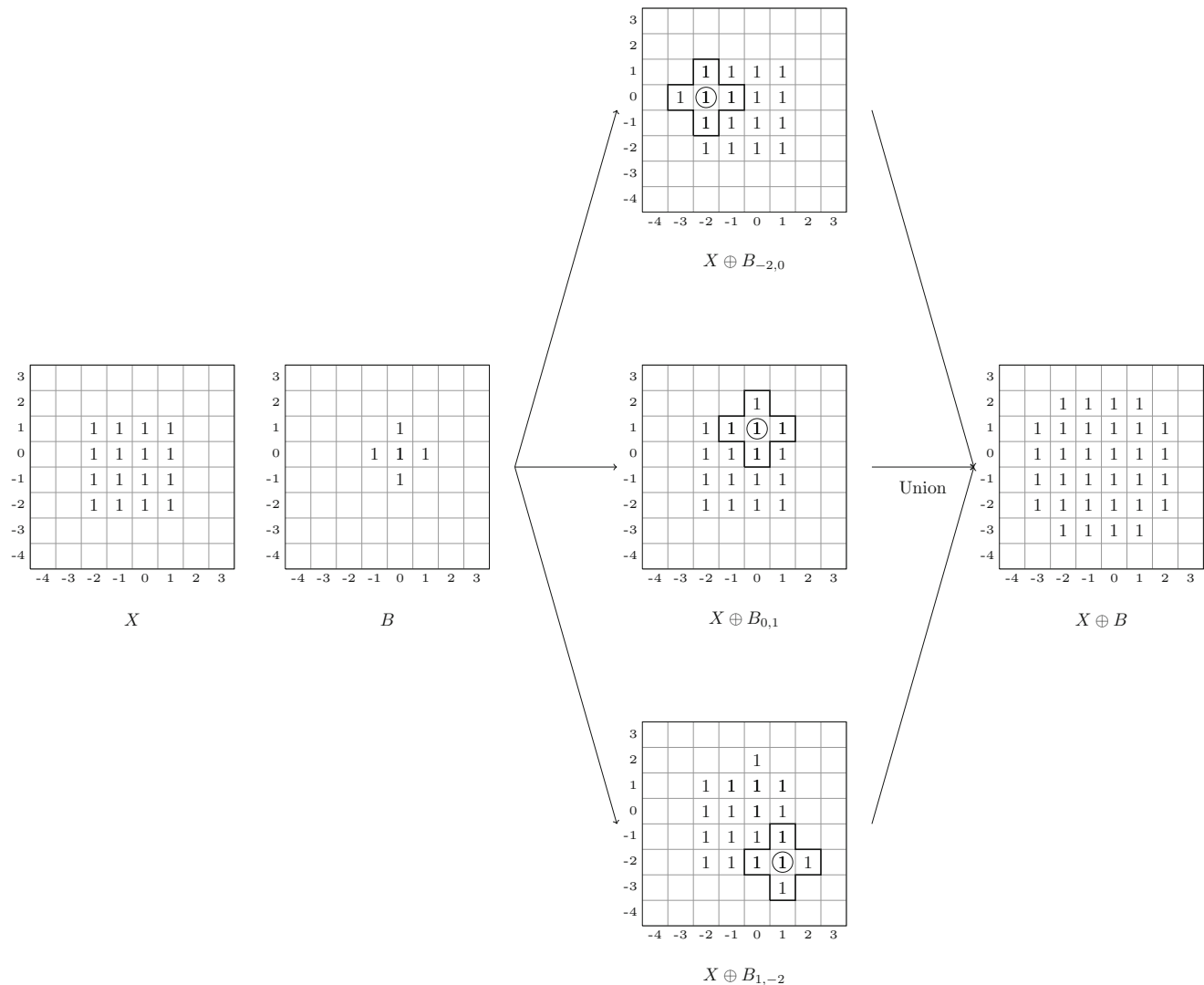
Illustrations

Morphological dilation is one of basic operators from the field of mathematical morphology. Its definition can be traced to fundamental works by Jean Serra (1983, 1988) and Georges Matheron (1975). Depending on the domain the definition of the dilation operator differs (see Dougherty and Lotufo (2003)). However, all these definitions can be obtained from a generic lattice-based definition. In this entry, we discuss the morphological dilation in specific case of discrete binary and gray-scale images. We then provide the generic lattice-based definition and show how this reduces to the definitions in discrete binary and gray-scale images.

Binary Images

The procedure to construct the dilation of a discrete binary image X is illustrated in Fig. 1. 0 values are not indicated in the image. Given the discrete binary image X and the structuring element B , we proceed by translating and placing B at all possible positions of X , as shown in Fig. 1. The final dilated image is obtained by taking the union of all these translations. In mathematical terms, this corresponds to (1).

One can also construct the dilation of a discrete binary image X by translating the image X for all possible values b in B , and then taking the union of these translations. In mathematical terms, this corresponds to



Morphological Dilation, Fig. 1 Illustration of morphological dilation

$$\delta_B(X) = \bigcup_{b \in B} X_b \quad (3) \quad \text{Observe that, in the case of flat structuring element, we have}$$

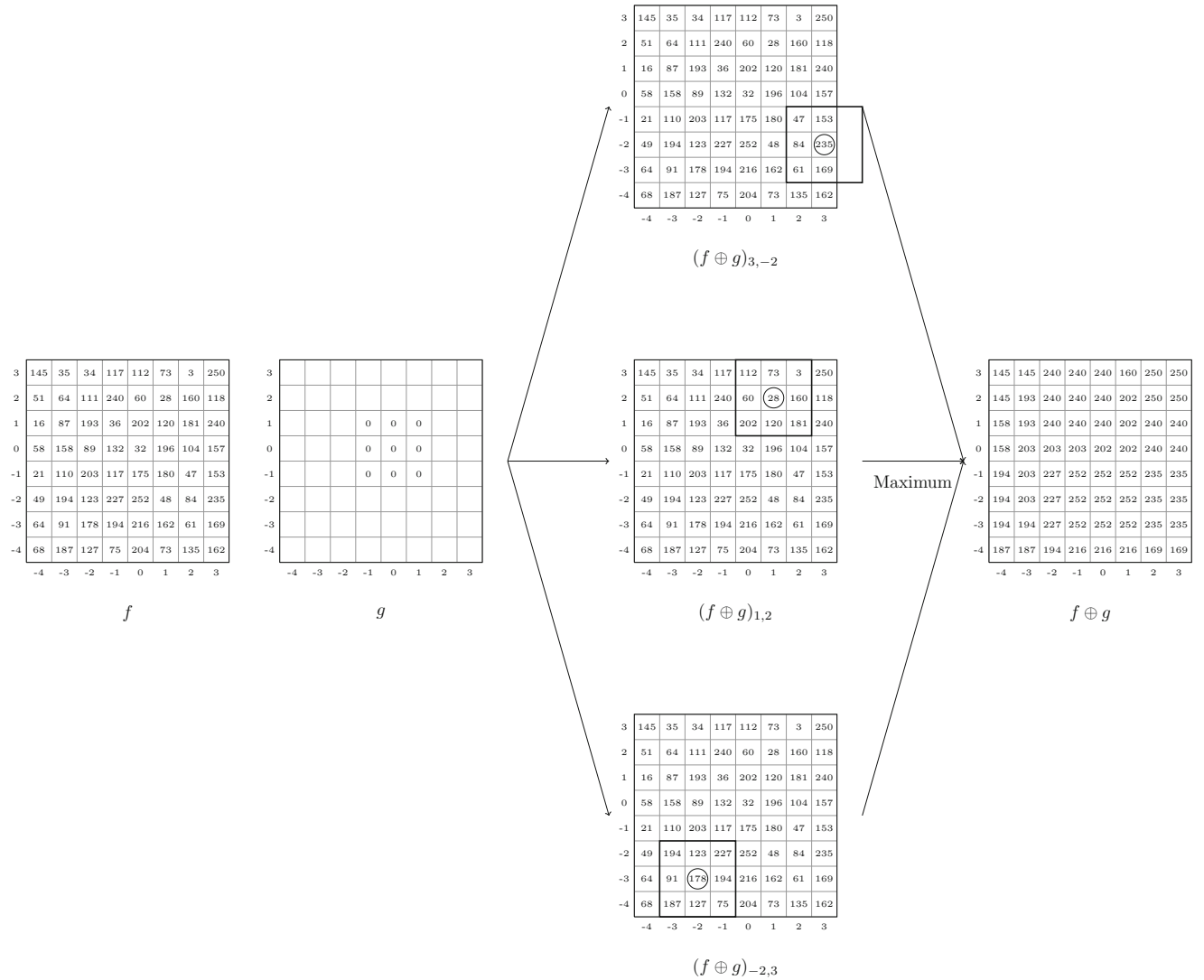
This is equivalent to the definition in (1).

$$\delta_g(f)(x) = \sup_{y \in E} \{f(y) + g(x-y)\} = \sup_{(x-y) \in K} \{f(y)\} \quad (4)$$

Gray-Scale Images

In the case of gray-scale images, structuring element g is a function as well. It is usually the practice that g is only defined on a finite/compact subset K of the domain. If, on the finite/compact subset, g only takes the value 0, it is referred to as the *flat structuring element*, else it is referred to as *non-flat structuring element*.

where K denotes the finite/compact set on which g is defined. That is, the dilated value is taken to be the maximum over a neighborhood. This is illustrated in Fig. 2. We consider a generic gray-scale image f . In gray-scale images, it is a convention that the values in the flat structuring element are taken to be 0 over the domain of definition and empty when it is not defined. As in the case of binary images, the structuring element is translated to each position of the original image. Within each neighborhood,



Morphological Dilation, Fig. 2 Illustration of morphological dilation with gray-scale images

is considered, and the central value is replaced by this maximum value. Finally, the maximum over all these images is taken. At the boundaries, only the values that exist in the domain are considered.

In the case of non-flat structuring element, the values of g are added to the neighborhood values and then the maximum is computed.

Remark: Recall that the gray-scale images can take the values in $\{0, 1, 2, \dots, 255\}$. However, it is possible that in definition (2) the right-hand-side value may be higher than 255. In such cases, we take any value greater than 255 to be 255.

Some Important Properties

The properties of the dilation operator are explained using the binary images. However, these concepts extend to the gray-scale images as well. Firstly, morphological dilation is an *increasing operator*,

$$X \subseteq Y \Rightarrow \delta_B(X) \subseteq \delta_B(Y) \quad (5)$$

Moreover, it is also an *extensive operator* when the structuring element B is symmetric and the center belongs to B .

$$X \subseteq \delta_B(X) \quad (6)$$

Also, dilation can be thought to be *commutative*

$$\delta_B(X) = \delta_X(B) \quad (7)$$

Lattice-Based Definition

Recall that a *partially ordered set* is a set $\mathcal{L}$ with a binary order $\leq$ such that: (i) $x \leq x$ (reflexivity) – (ii) $x \leq y$ and $y \leq x$ implies $x = y$ (anti-symmetry), and (iii) $x \leq y$ and $y \leq z$ implies $x \leq z$ holds true. A partially ordered set is called a *lattice* when any finite subset of $\mathcal{L}$ has an infimum and a supremum. It is called a *complete lattice* if any subset of $\mathcal{L}$ has an infimum and a supremum.

The dilation operator on a complete lattice is defined as an operator $\delta: \mathcal{L} \rightarrow \mathcal{L}$ which

- is increasing, i.e., $x \leq y$ implies $\delta(x) \leq \delta(y)$,
- preserves the supremum, i.e., $\delta(\bigvee_{i \in I} x_i) = \bigvee_{i \in I} (\delta(x_i))$.

In case the index set I is empty, we have $\delta(\mathbf{0}) = \mathbf{0}$, where $\mathbf{0}$ denotes the infimum of $\mathcal{L}$.

How Does This Definition Relate to Definitions in (1) and (2)?

In case of discrete binary images, the complete lattice is taken to be the power set, i.e., $\mathcal{L} = \mathcal{P}(E)$ where E denotes the domain of definition for the binary images. The infimum/supremum of any subset is obtained by intersection/union, respectively. Assume that $\delta(\{0\}) = B$, i.e., the set $\{0\}$ maps a set B in the lattice. Using *invariance w.r.t translation*, we get

$$\delta(\{x\}) = \delta(T_x(\{0\})) = T_x(\delta(\{0\})) = T_x(B) = B_x \quad (8)$$

Where T_x is a translation operator, which translates all elements by x , and B_x denotes the set B , which is translated by x as usual. Now, to obtain the dilation of a generic set X , we have

$$\delta(X) = \delta(\bigvee_{x \in X} x) = \bigvee_{x \in X} (\delta(x)) = \bigvee_{x \in X} B_x = \bigcup_{x \in X} B_x \quad (9)$$

which is identical to the definition of dilation one has in discrete binary images.

In case of discrete gray-scale images, the lattice is obtained by the set of all possible functions $f: E \rightarrow \{0, 1, 2, \dots, 255\}$ where E is the domain of definition for the image. The binary images can be thought “of” as a specific case where $f: E \rightarrow \{0, 1\}$ alone. Define

$$f_1 \leq f_2 \Leftrightarrow f_1(x) \leq f_2(x) \text{ for all } x \in E \quad (10)$$

The infimum of two functions is given by

$$(f_1 \wedge f_2)(x) = (f_1(x) \wedge f_2(x)) \quad (11)$$

and the supremum is given by

$$(f_1 \vee f_2)(x) = (f_1(x) \vee f_2(x)) \quad (12)$$

Now, observe that if E is finite, then the set of all possible functions as above is a complete lattice, $\mathcal{L}$. Consider the set of functions e_i defined as

$$e_{v,z}(x) = \begin{cases} 0 & \text{if } x \neq z \\ v & \text{if } x = z \end{cases} \quad (13)$$

Observe that any function f can be written as

$$f = \bigvee \{e_{v,z} | e_{v,z} \leq f\} \quad (14)$$

So, if one defines the dilation operator on $e_{v,z}$, then from the law of preserving the supremum we obtain the dilation for f .

Moreover, assuming the translation to invariance in the domain, one needs to define the dilations for cases of $e_{v,0}$ where 0 denotes the origin. For each of these functions we assume that the dilation is obtained by

$$\delta(e_{v,0}) = e_{v,0} + g \quad (15)$$

So, we have

$$\delta(f) = \delta(\vee\{e_{v,z} | e_{v,z} \leq f\}) \quad (16)$$

$$= \vee(\{\delta(e_{v,z}) | e_{v,z} \leq f\}) \quad (17)$$

$$= \vee(\{\delta(e_{v,0}) | e_{v,0} \leq T_{-z}(f)\}) \quad (18)$$

$$= \vee(\{e_{v,0} + g | e_{v,0} \leq T_{-z}(f)\}) \quad (19)$$

$$= \vee(\{T_{-z}(f) + g\}) \quad (20)$$

The second equality follows since dilation preserves the supremum. The third equality follows from translation to invariance. That is, if $e_{v,z}(x) < f(x)$, then we have $e_{v,0} < f(x - z)$. Simply substituting the dilation definition gives the fourth equality. The final equality follows from noting that $\vee e_{v,0} \leq T_{-z}(f) = T_{-z}(f)$. Now, the final equation for a specific $x \in E$ can be written as

$$\delta(f)(x) = \vee(\{f(x - z) + g(x)\}) \quad (21)$$

This is identical to the gray-scale definition we have in (2).

Summary

In this entry, morphological dilation is defined in the context of binary and gray-scale images. These definitions are then illustrated on fictitious images. Several types of structuring elements that are intrinsic to the definition of a morphological dilation are described. The main properties of morphological dilation – increasing, extensive, and commutative – are discussed. The definition of morphological dilation in lattices is then described in detail.

Cross-References

- [Mathematical Morphology](#)
- [Morphological Closing](#)
- [Morphological Erosion](#)
- [Morphological Opening](#)
- [Structuring Element](#)

Bibliography

- Dougherty ER, Lotufo RA (2003) Hands-on morphological image processing, vol 59. SPIE Press, Bellingham
- Matheron G (1975) Random sets and integral geometry. AMS, Paris
- Serra J (1983) Image analysis and mathematical morphology. Academic, London
- Serra J (1988) Image analysis and mathematical morphology, Theoretical advances, vol 2. Academic, New York

Morphological Erosion

Aditya Challa¹, Sravan Danda² and B. S. Daya Sagar³

¹Computer Science and Information Systems, APPCAIR, BITS Pilani KK Birla Goa Campus, Pilani, Goa, India

²Computer Science and Information Systems, APPCAIR, Birla Institute of Technology and Science Pilani, Pilani, Goa, India

³Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, Karnataka, India

Definition

Morphological erosion is one of the four basic operators from the field of Mathematical Morphology (MM). This is a dual operator to morphological dilation. Depending on the domain of application, the definition varies. In the standard case of discrete binary images, given a binary image X , and a structuring element B (cross-ref), the erosion of X is defined as

$$\epsilon_B(X) = \bigcap_{b \in \hat{B}} X_b \quad (1)$$

Here $\hat{B}$ denotes the reflection of $B - x \in B \Leftrightarrow -x \in \hat{B}$. X_b denotes the image X translated by b for each b belonging to the $\hat{B}$. It is also a common practice to write $\epsilon_B(X) = X \ominus B$. Observe that the definition in (1) extends to continuous binary images as well.

In the case of gray-scale images, let f denote the gray-scale image and g denote the structuring function. Then, morphological erosion is defined by

$$\epsilon_g(f)(x) = \inf_{y \in E} \{f(y) - g(x - y)\} \quad (2)$$

where E denotes the domain of definition.

These definitions can be derived from the fact that erosion is dual to dilation using the definition of dilation.

Illustrations

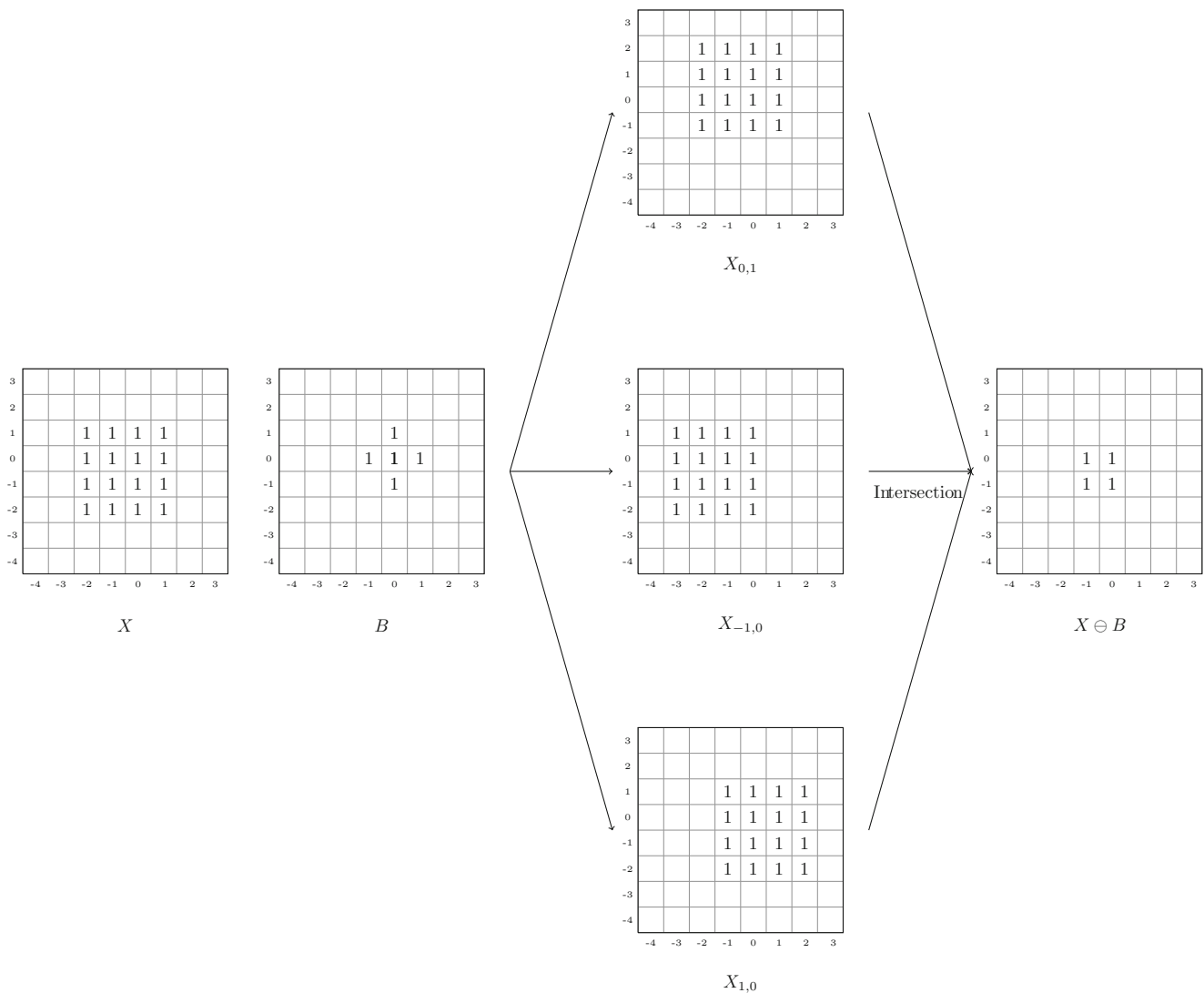
Morphological erosion is one of the basic operators from the field of mathematical morphology. Its definition can be traced to fundamental works by Jean Serra (1983, 1988) and Georges Matheron (1975). Depending on the domain, the definition of the erosion operator differs Dougherty ER, Lotufo RA (2003). However, all these definitions can be obtained from a generic lattice-based definition. In this entry, we discuss morphological erosion in specific case of discrete binary and gray-scale images. We then provide the derivation of these definitions from the lattice-based definitions.

Binary Images

The procedure to construct the erosion of a discrete binary image X is illustrated in Fig. 1. Zero values are not indicated in the image. Given the discrete binary image X and the structuring element B , we proceed by translating X by $-b$ for all possible values in B , as shown in Fig. 1. The final eroded image is obtained by taking the intersection of all these translations. In mathematical terms, this corresponds to (1).

Gray-Scale Images

In the case of gray-scale images, structuring element g is a function as well. It is usually the practice that g is only defined on a finite/compact subset K of the domain. If, on the finite/compact subset, g only takes the value 0, it is referred to as the



Morphological Erosion, Fig. 1 Illustration of morphological erosion

flat structuring element, else it is referred to as nonflat structuring element.

Observe that in the case of flat structuring element, we have

$$\epsilon_g(f)(x) = \inf_{y \in E} \{f(y) - g(x-y)\} = \inf_{(x-y) \in K} \{f(y)\} \quad (3)$$

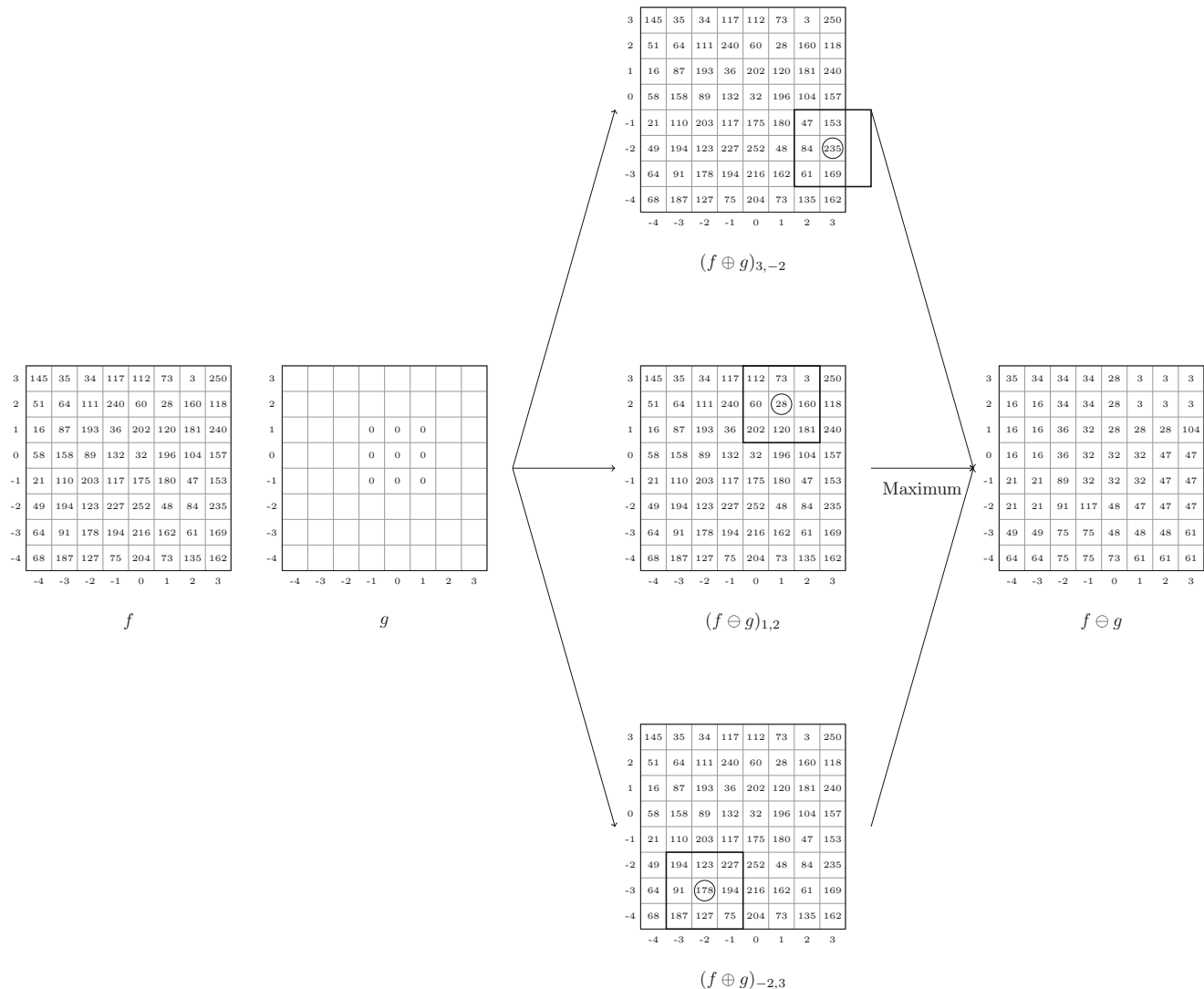
where K denotes the finite/compact set on which g is defined. That is, the eroded value is taken to be the minimum over a neighborhood. This is illustrated in Fig. 2. We consider a generic gray scale image f . In gray-scale images, it is a convention that the values in the flat structuring element are taken to be 0 over the domain of definition and empty when they are not defined. As in the case binary images, the structuring element is translated to each position of the original

image. Within each neighborhood, the maximum is considered, and the central value is replaced by maximum values. Finally, the minimum over all these images is taken. At the boundaries, only the values which exist in the domain are considered.

In the case of the nonflat structuring element, the values of g are added to the neighborhood values and then the minimum is computed.

Some Important Properties

The properties of the erosion operator are explained using the binary images. However, these properties could extend to the gray-scale images as well. Most of these properties can be derived from noting that the erosion operator is dual to the



Morphological Erosion, Fig. 2 Illustration of morphological dilation with gray-scale images

dilation operator. First, morphological erosion is *increasing operator*,

$$X \subseteq Y \Rightarrow \epsilon_B(X) \subseteq \epsilon_B(Y) \quad (4)$$

Moreover, it is also an *antiextensive operator* when the structuring element B is symmetric.

$$X \supseteq \epsilon_B(X) \quad (5)$$

Lattice-Based Definition

Recall that a *partially ordered set* is a set $\mathcal{L}$ with a binary order $\leq$ such that: (i) $x \leq x$ (reflexivity), (ii) $x \leq y$ and $y \leq x$ implies $x = y$ (antisymmetry); and (iii) $x \leq y$ and $y \leq z$ imply $x \leq z$ holds true. A partially ordered set is called a lattice when any finite subset of $\mathcal{L}$ has an infimum and a supremum. It is called a complete lattice if any subset of $\mathcal{L}$ has an infimum and a supremum.

The erosion operator on a complete lattice is defined as an operator $\epsilon : \mathcal{L} \rightarrow \mathcal{L}$ which

- Is increasing, i.e., $x \leq y$ implies $\epsilon(x) \leq \epsilon(y)$
- Preserves the infimum, i.e., $\epsilon(\bigwedge_{i \in I} x_i) = \bigwedge_{i \in I} (\epsilon(x_i))$

In case the index set I is empty, we have $\epsilon(\mathbf{m}) = \mathbf{m}$, where $\mathbf{m}$ denotes the supremum of $\mathcal{L}$.

How Does This Definition Relate to Definitions in (1) and (2)?

The optimal approach to deriving equations (1) and (2) is by using the *duality* property, i.e.,

$$\epsilon(x) = (\delta(x^c))^c \quad (6)$$

where x^c denotes the complement of x in the lattice $\mathcal{L}$. Observe that the definition of the erosion can in fact be derived starting from the duality principle. In the chapter of morphological dilation (chapter-ref), we derived the specific definition from the lattice definitions. Here, we shall use the duality property to arrive at definitions (1) and (2).

In case of binary images, we have the duality by complement operation. It has already known that

$$\delta_B(X) = \bigcup_{x \in X} |B_x = \bigcup_{b \in B} X_b \quad (7)$$

So using duality by complementation, we have

$$\epsilon_B(X) = (\delta_B(X^c))^c = \left(\bigcup_{b \in B} (X^c)_b \right)^c \quad (8)$$

$$= \left(\bigcap_{b \in B} X_b \right) \quad (9)$$

As a convention, if B is the structuring element used for dilation, then $\hat{B}$, the reflection of the structuring element, is used for erosion. Accordingly, a reflection operation is embedded into the definition of the erosion to obtain (1). This is because, only if the structuring element B is reflected, would (δ_B, ϵ_B) form an adjunction, i.e.,

$$\delta_B(X) \leq y \Leftrightarrow x \leq \epsilon_B(y) \quad (10)$$

Adjunction is an important property for defining morphological opening

In the case of gray-scale images, the duality is obtained by inversion

$$\epsilon_g(f) = 255 - \delta_g(255 - f) \quad (11)$$

Then we have,

$$\epsilon_g(f) = 255 - \delta_g(255 - f) \quad (12)$$

$$= 255 - \sup_{y \in E} \{255 - f(y) + g(x - y)\} \quad (13)$$

$$= \inf_{y \in E} \{255 - (255 - f(y) + g(x - y))\} \quad (14)$$

$$= \inf_{y \in E} \{f(y) - g(x - y)\} \quad (15)$$

which is equivalent to definition in (2).

The lattice-based definitions are much more general, and the specific definitions can be derived from here.

Summary

In this chapter, morphological erosion is defined in the context of binary and gray-scale images. These definitions are then illustrated on simple examples. Several types of structuring elements which are intrinsic to the definition of a morphological erosion are described. The main properties of morphological erosion – increasing and antiextensive – are discussed. The definition of morphological erosion in lattices is then described in detail.

Cross-References

- [Mathematical Morphology](#)
- [Morphological Closing](#)
- [Morphological Dilation](#)
- [Morphological Opening](#)
- [Structuring Element](#)

Bibliography

- Dougherty ER, Lotufo RA (2003) Hands-on morphological image processing, vol 59. SPIE Press
- Matheron G (1975) Random sets and integral geometry. Wiley, New York
- Serra J (1983) Image analysis and mathematical morphology. Academic
- Serra J (1988) Image analysis and mathematical morphology, volume 2: Theoretical advances. Academic, New York

Morphological Filtering

Jean Serra

Centre de morphologie mathématique, Ecoles des Mines, Paristech, Paris, France

Definition

An operation which maps a complete lattice into itself is said a *morphological filter* when it is increasing and idempotent. Morphological filtering is mainly used for binary and vector images, and for partitions. When a connection, or a connectivity, is involved, morphological filtering allows to segment images, i.e., to find contours for the objects. Moreover these filters often depend on a size parameter which orders them in a hierarchy and from which one extracts the optimal segmentation.

Morphological Filters

In physics or in chemistry, when one says that one filters a precipitate, one describes in some way an opening. However this meaning is not the only one. In signal processing, one usually means by filter any *linear* operator that is *invariant under translation* and *continuous*. According to a classical result, every filter in the previous sense is expressed as the convolution product $F * \varphi$ of the signal F by a convolution distribution φ , and the transform of $F + F'$ is the sum of the transforms of F and of F' . If you listen on the radio to a duet for piano and violin, it is quite natural that the amplified sound should be the sum of the amplifications from the piano alone and from the violin alone.

To the three axioms of convolution, a fourth one is in fact added: it is in practice very common to consider filters as being of the band-pass type, even if that is not quite exact. One says about a Hi-Fi amplifier that it is “low-pass up to 30 000 Hz”, of a tainted glass that it is monochromatic, etc. In this doing, one attributes implicitly to the operation of filtering the property of not being able to act by iteration: the signal that has lost its frequencies above 30,000 Hz will not be modified further if one submits it to a second amplifier identical to the first. We find again idempotence, which however is not captured by the axioms of convolution.

This approach, based on linear filtering, is very well adapted to telecommunications, where the fundamental problem is the pass band of the channel, which requires to modify consequently the frequency band of the signal. But in geosciences and in image analysis, what matters is not to recognize frequencies, but to detect, localize and measure objects. One must thus be able to express spatial relations between objects, the first one being *inclusion*, which induces the lattice framework, where growth replaces the linearity of vector spaces. Since furthermore complete lattices lend themselves very well to taking into account idempotence, we are led to define as a morphological filter as follows:

Definition *Given a complete lattice $\mathcal{L}$, one calls a morphological filter, or more briefly a filter, any operator $\psi : \mathcal{L} \rightarrow \mathcal{L}$ that is both increasing and idempotent.*

Openings and closings, that we already met in the article ► [“Mathematical Morphology”](#), illustrate this notion, but one can devise many other filters. For example, we can think of these two operations as primitives and combine them serially or in parallel to create new filters. The first mode, by composition products leads to the alternating (sequential) filters, and the second one to the lattice of filters. Remark that translation invariance is never a prerequisite, the filters may vary from place to place, according to the context.

Products of Filters

When ψ and ξ are two ordered filters, that is, $\psi < \xi$, the inequalities

$$\psi\xi \leq \psi\psi\xi \leq \psi\xi\psi\xi \leq \psi\xi\xi\xi \leq \psi\xi$$

show that the product $\psi\xi$ is in turn a filter.

For the sake of concision, let us write $\mathcal{B}_\alpha$ for the invariance domain $\text{Inv}(\alpha)$ of a filter α . One can state the:

Matheron’s Criterion (Matheron 1988)

Let ψ and ξ be two filters such that $\psi \leq \xi$. Then:

1. Their products generate only the four filters $\psi\xi$, $\psi\xi\psi$, $\xi\psi\xi$, and $\xi\psi$ that are partly ordered:

$$\psi \leq \psi \xi \psi \leq \left\{ \begin{matrix} \xi \psi \\ \psi \xi \end{matrix} \right\} \leq \xi \psi \xi \leq \xi.$$

$\xi \psi \xi$ is the least filter above $\xi \psi \vee \psi \xi$, and $\psi \xi \psi$ is the greatest filter below $\xi \psi \vee \psi \xi$.

2. Although the product $\xi \psi$ is not commutative in general, each filter $\xi \psi$ or $\psi \xi$ eliminates positive noise and sharp reliefs, as well as negative noise and narrow hollows. When the two primitive are an opening γ and a closing φ , which is the most popular case, then the two four products $\gamma \varphi$ and $\varphi \gamma$ have the following characteristic property:

$$f \leq g \leq f \vee \gamma \varphi(f) \implies \gamma \varphi(g) = \gamma \varphi(f)$$

and similarly $f \wedge \varphi \gamma(f) \leq g \leq f \implies \varphi \gamma(g) = \varphi \gamma(f)$. The first property ensures that any signal g between f and $f \vee \gamma \varphi(f)$ gives the same filtered version as f itself. The second one is the dual version for $\varphi \gamma$.

Iterations of Over and Underfilters

We just saw how to generate filters by composition products. Is it also possible to obtain filters by suprema or infima? The supremum $\bigvee_{i \in I} \psi_i$ (respectively, the infimum $\bigwedge_{i \in I} \psi_i$) of a family $\{\psi_i \mid i \in I\}$ of filters is visibly overpotent (respectively, underpotent), that is, $\psi^2 \geq \psi$ (respectively $\psi^2 \leq \psi$), which makes it lose the idempotence that the ψ_i had. One can nevertheless speak of a lattice of filters (Matheron 1988).

Put $\bigvee_{i \in I} \psi_i = \alpha$. When $\mathcal{L}$ is finite, we always reach $\alpha^{n+1} = \alpha^n$ for n large enough; it suffices to iterate the overpotent operator α to obtain finally the idempotent limit $\Phi(\alpha) = \alpha^n$ which turns out to be the least filter above all the ψ_i . When $\mathcal{L}$ is infinite, $\alpha, \alpha^2, \dots, \alpha^n$ and $\bigvee_{n \in \mathbb{N}} \alpha^n$ are still overpotent, but $\bigvee_{n \in \mathbb{N}} \alpha^n$ is not necessarily idempotent; it will be if α is $\uparrow$ -continuous. Dually, we put $\bigwedge_{i \in I} \psi_i = \beta$. If the lattice $\mathcal{L}$ is finite, or if β is $\downarrow$ -continuous, then $\bigwedge_{n \in \mathbb{N}} \beta^n$ will be the greatest filter $\Gamma(\beta)$ below all the ψ_i . Let us also remark that in some cases one of the terms, Φ or Γ , is obtained directly, but not the other: for example, when the ψ_i are openings ψ_i , the supremum $\Phi(\bigvee_{i \in I} \gamma_i) = \bigvee_{i \in I} \gamma_i$, but the infimum $\Gamma(\bigwedge_{i \in I} \gamma_i)$ requires iterations.

Hierarchies and Semigroups

By regard for pedagogy, until now we have not parameterized the filters. To make them depend upon a parameter $\lambda > 0$ of scale or size, leads us to ponder the structure of the obtained family $\{\Psi_\lambda \mid \lambda \geq 0\}$. For example, is the product $\Psi_\mu \Psi_\lambda$ of two of the filters in the family? If the increasing λ correspond to stronger and stronger simplifications, can one start from any intermediate level $\Psi_\mu(A)$ to reach $\Psi_\lambda(A)$, when $\mu < \lambda$? For answering these questions, we can firstly examine the openings γ and the alternating filters $\gamma \varphi$.

Granulometries

The case of openings is the simplest. Following G. Matheron, we will say that the family $\{\gamma_\lambda \mid \lambda \geq 0\}$ with a positive parameter generates a *granulometry* when:

1. For every $\lambda \geq 0$, the operator γ_λ is an opening, with $\gamma_0 = \text{id}$
2. In the composition product, the most severe opening imposes its law:

$$\gamma_\mu \gamma_\lambda = \gamma_{\max\{\lambda, \mu\}}. \quad (1)$$

In terms of sifting, the grains refused by a sieve will be *a fortiori* by sieves with smaller holes. This second axiom amounts to saying that the openings decrease when the parameter increases, that is, $\lambda \geq \mu \geq 0 \implies \gamma_\lambda \leq \gamma_\mu$, or as well that their invariance domains decrease, that is, $\lambda \geq \mu \geq 0 \implies \mathcal{B}_\lambda \subseteq \mathcal{B}_\mu$. By duality, one constructs in a similar way the anti-granulometry $\{\varphi_\lambda \mid \lambda \geq 0\}$.

The two axioms of Matheron model the sieving technique, where solid particles are classified according to a series of sieves with decreasing meshes. They describe, similarly, the sizes of the individuals in a population. But they also apply to procedures which are not based on individual particles, like the Purcell method used in the petroleum industry, where mercury is injected into specimen of porous rocks under various pressures. Or again they model the linear intercept distributions in image analysis, the disc openings, etc.

The relation (1) defines a commutative semigroup, with the identity **id** as neutral element, called the Matheron semigroup, whose use extends well beyond the granulometric case of openings.

In particular, in the Euclidean case with invariance under translation, this relation weaves a link between granulometry by adjunction, similarity, and convexity. Indeed, every λB homothetic to the compact B is open by adjunction by μB for every $\mu \geq \lambda$ if and only if B is convex, and consequently the family $\{\gamma_{\lambda B} \mid \lambda \geq 0\}$ of openings by the convex sets $\{\lambda B \mid \lambda > 0\}$ is granulometric. Here, the hypotheses of convexity or similarity are equivalent. On the other hand, if one does not impose similarity, then the $B(\lambda)$ do not need to be convex, nor even connected.

Alternating Sequential Filters

Let us proceed to the products of the type $\varphi \gamma$, or $\gamma \varphi$, of an opening by a closing. To parameterize directly the two primitives leads to a rather coarse result, that one refines by replacing the primitives by a granulometry $\{\gamma_\lambda \mid \lambda \geq 0\}$ and an anti-granulometry $\{\varphi_\lambda \mid \lambda \geq 0\}$. Let us suppose for the moment λ a positive integer, and set:

$$\varpi_n = \varphi_n \gamma_n \cdots \varphi_2 \gamma_2 \varphi_1 \gamma_1.$$

The operator ϖ_n is a filter, designated as *alternating sequential* (or *ASF*), and due to S.R. Sternberg. Although it



Morphological Filtering, Fig. 1 Left: initial mosaic; other views, from left to right: connected ASF of sizes $\lambda = 1, 4, 6$

is constructed in view of the semigroup, it does not satisfy Eq. (1), but only the *absorption law*.

$$p \geq n \Rightarrow \varpi_p \varpi_n = \varpi_p,$$

which is sufficient to construct hierarchies of more and more severe ASF. Nevertheless, when the γ_λ and φ_γ , $\lambda \geq 0$, are *grain operators*, that is, families of connected openings (respectively, closings) that process each grain (respectively, pore) independently from the others, then the associated ASF ϖ_n form a Matheron semigroup Serra (2000). An example is given in Fig. 1. The γ_i are connected openings by reconstruction from the eroded initial image by a disc of radius λ , and the φ_i are the dual operators. Each contour is preserved or suppressed, but never deformed: the initial partition increases under the successive filters, which form a Matheron semigroup.

When λ is a positive real, one often subdivides the segment $[0, \lambda]$ into $2, 4, \dots, 2^k, \dots$ sections of filters $\varpi_k(\lambda) = \varphi_\lambda \gamma_\lambda \cdots \varphi_{i2^{-k}\lambda} \gamma_{i2^{-k}\lambda} \cdots \varphi_0 \gamma_0$, with $0 \leq i \leq 2^k$. When k increases, the $\varpi_k(\lambda)$ decrease, and $\bigwedge_k \varpi_k(\lambda) = \varpi(\lambda)$ still remains a filter Serra (1988) p.205. The preceding properties extend to that continuous version. Finally, what has been said here for the sequences of primitives $\varphi_\lambda \gamma_\lambda$ remains true if one starts with from $\gamma_\lambda \varphi_\lambda$.

Connections

In image processing, the fundamental operation associated with connectivity consists in directing a point towards a set and extracting the marked particle. The result depends on the choice of the said connectivity (e.g., 4- or 8-connectivities in square grid), but in all cases, the particles of a set A pointed at x and at y are either identical, or disjoint, for all x and y of the space. Moreover, the operation which goes from A to its connected component in x is obviously an opening. Besides, connected zones that intersect are included in a same connected component. These characteristics can be

summarized in the following axiomatics (Serra 1988) p.51. Let E be an arbitrary space:

- A *connection* on $\mathcal{P}(E)$ is a family $\mathcal{C} \subseteq \mathcal{P}(E)$ that satisfies the following three conditions:
 1. $\emptyset \in \mathcal{C}$.
 2. $\forall p \in E, \{p\} \in \mathcal{C}$.
 3. $\forall \{C_i | i \in I\} \subseteq \mathcal{C}, \bigcap_{i \in I} C_i \neq \emptyset \Rightarrow \bigcup_{i \in I} C_i \in \mathcal{C}$.
 An element of $\mathcal{C}$ is called *connected*.
- A *system of connection openings* on $\mathcal{P}(E)$ associates with each point $p \in E$ an opening γ_p on $\mathcal{P}(E)$ that satisfies the following three conditions, $\forall p, q \in E$:
 1. $\gamma_p(\{p\}) = \{p\}$.
 2. $\forall X \in \mathcal{P}(E), \gamma_p(X) \cap \gamma_q(X) \neq \emptyset \Rightarrow \gamma_p(X) = \gamma_q(X)$.
 3. $\forall X \in \mathcal{P}(E), p \notin X \Rightarrow \gamma_p(X) = \emptyset$.

For $p \in X$, $\gamma_p(X)$ is referred to as the *connected component* of X marked by p .

The two notions are indeed equivalent (Serra 1988) p.52, as shown by the following theorem:

Theorem

There exists a bijection between the connections on $\mathcal{P}(E)$ and the systems of connection openings on $\mathcal{P}(E)$:

- with a connection $\mathcal{C}$ one associates the system of connection openings $(\gamma_p, p \in E)$ defined for all $X \in \mathcal{P}(E)$ by:

$$\gamma_p(X) = \bigcup \{C \in \mathcal{C} | p \in C \subseteq X\}; \quad (2)$$

- with a system of connection openings $(\gamma_p, p \in E)$ one associates the connection $\mathcal{C}$ defined by:

$$\mathcal{C} = \{\gamma_p(X) | p \in E, X \in \mathcal{P}(E)\}. \quad (3)$$

Classically in topology, a set is connected when it cannot be partitioned into two non-empty closed (or open) regions. In topology, one also speaks of arcwise connectivity, according

to which a set A is connected when for each pair of points $a, b \in A$, one can find a continuous map ψ from $[0, 1]$ into A such that $\psi(0) = a$ and $\psi(1) = b$. This last connectivity is more restrictive than the first one, though in R^n both notions coincide on the open sets. In discrete geometry, the digital connectivities are particular cases of graph connectivity, which transposes to arcs the Euclidean arcwise definition. In Z^2 one then finds, among others, the classical 4- and 8-connectivities of the square grid and the 6-connectivity of the hexagonal grid; and in Z^3 the 6- and 26-connectivities of the cube, or that of the cube-octahedron.

All these topological or digital connectivities satisfy the axiomatics of connections. But the latter does not presuppose any topology, or the distinction between continuous and discrete approaches. It is thus more comprehensive, and also better adapted to image processing, as it starts from one of its basic operations. The axiomatic of a connection has been relaxed by C. Ronse, by suppressing the axiom about the points (Ronse 2008), and by M. Wilkinson by extending it to ultra-connections which allow some covering (Wilkinson 2006).

Examples of Connections

We now describe two examples of connections taken among many other ones (Ronse 2008; Serra 1988). Both are often used in practice and cannot be reduced to usual connectivities.

Connection by dilation Start from a set $\mathcal{P}(E)$ which is already provided with a connection C and consider an extensive dilation $\delta : \mathcal{P}(E) \rightarrow \mathcal{P}(E)$ which preserves C , i.e., $\delta(C) \subseteq C$; equivalently, it is required that $\forall p \in E, p \in \delta(p) \in C$. Then the inverse image $C' = \delta^{-1}(C)$ of C by δ constitutes a second connection, richer than C , i.e., $C' \supseteq C$. For all $A \in \mathcal{P}(E)$, the C -components of $\delta(A)$ are exactly the images by δ of the C' -components of A . If γ_x stands for the opening associated with C , and v_x for that associated with C' , we have:

$$\delta v_x(A) = \gamma_x \delta(A) \cap A \quad \text{if } x \in A, \quad v_x(A) = \emptyset \quad \text{otherwise.}$$

If, in the Euclidean or digital plane, we take for example for δ the dilation by a disc of radius r , then the openings γ_x

characterize the clusters of objects whose *infimum* of the distances between points is $\leq 2r$ (Fig. 2, right). *A contrario*, the same approach extracts also the isolated connected components in a set A , since they specifically satisfy the equality $v_x(A) = \gamma_x(A)$.

Connection induced by a partition Consider a given partition D , and a point $x \in E$. The operation which associates, with each $A \subseteq E$, the transform

$$\gamma_x(A) = D(x) \cap A \quad \text{if } x \in A, \quad \gamma_x(A) = \emptyset \quad \text{otherwise.}$$

is clearly an opening, and as x varies, the $\gamma_x(A)$ and $\gamma_y(A)$ are identical or disjoint because they correspond to classes of partitions. The class:

$$C = \{\gamma_x(A) | x \in E, A \in \mathcal{P}(E)\}$$

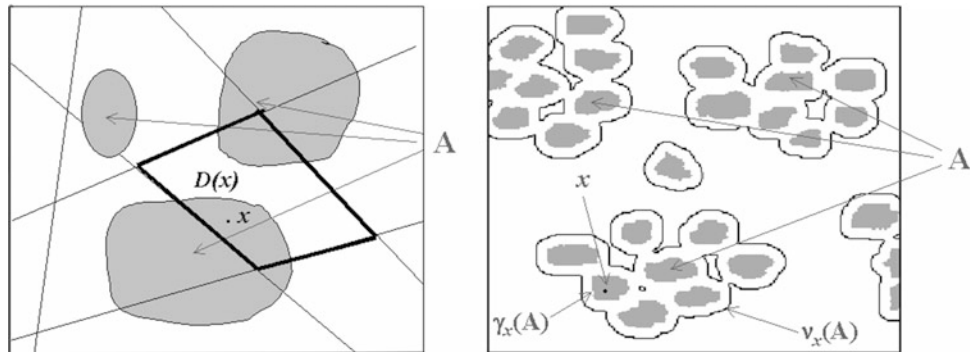
is thus a connection. One sees in Fig. 2 (left) that C breaks the usual connected components, and puts together their pieces when they lie in a same class $D(x)$. If E has been provided with a prior connection C' , like the usual arcwise one, then the elements of $C \cap C'$ are the connected components in the sense of C' of the intersections $A \cap D(x)$.

A few properties of connections are as follows:

1. *Lattice*: The set of all connections on E is a complete lattice, where the *infimum* of the family $\{C_i | i \in I\}$ is the intersection $\bigcap_{i \in I} C_i$, and the *supremum* is the least connection containing $\bigcup_{i \in I} C_i$.
2. *Maximum partitioning*: The openings γ_x of connection C partition every set $A \subseteq E$ into the set $\{\gamma_x(A) | x \in A\}$ of its connected components; it is the coarsest partition whose all classes belong to C , and this partition increases with A : if $A \subseteq B$, then every connected component of A is included in a unique connected component of B ;
3. *Increasingness*: If C and C' are two connections on $\mathcal{P}(E)$, with $C \subseteq C'$, then the partition of A into its C' -components is coarser than that into its C -components;

Morphological Filtering,

Fig. 2 Left: Connection by partition: the connected component of A at point x is the union of the two pieces of particles of $A \cap D(x)$; right: connection by dilation the particles of each cluster generate a second connection



4. *Arcs*: The set A is connected for C if and only if for all $x, x' \in A$, one can find a component $X \in C$ that contains x and x' .

Connected Filters

In what follows, E is any space and $\mathcal{F}$ designates a lattice of functions $E \rightarrow \bar{R}$ (or $\bar{Z}$). We have already met a connected filter, since the datum of a connection C on E is equivalent to that of the connected point openings $\{\gamma_x | x \in E\}$. We shall extend a property of the latter by stating that an operator $\psi : \mathcal{F} \rightarrow \mathcal{F}$ is *connected* relatively to the connective criterion σ when for every $F \in \mathcal{F}$ and every $A \subseteq E$ it satisfies the implication:

$$\sigma[F, A] = 1 \Rightarrow \sigma[\psi(F), A] = 1. \quad (4)$$

In other words, the connection generated by the pair $(\sigma, \psi(F))$ contains that of (σ, F) . When, moreover, the operator ψ is a filter, one speaks of a *connected filter*. Since the partition of E into zones where $\psi(F)$ is homogeneous according to σ is coarser than that of E for (σ, F) , one proceeds from the second one to the first one by removing frontiers only. On the other hand, an operator that coarsens partitions is not necessarily connected. Some examples of connected filters are as follows:

1. E is provided with a connection C , and $\sigma(F, A) = 1$ for A connected and F constant over A . Every filter ψ that increases the flat connected zones, that is, those where F is constant, is connected. This corresponds to the usual meaning of the expression *connected filter* (Salember and Serra 1995).
2. E is still provided with the connection C , and $\mathcal{F}$ is the lattice of binary functions $E \rightarrow \{0, 1\}$. Since this lattice is isomorphic to $\mathcal{P}(E)$, we develop the example in the set-theoretical framework. The opening γ_x of the flat zones criterion applied to the set A gives simply the connected component of A that contains the point x . It is a connected filter, as well as, more generally, the opening $\gamma_M = \bigvee \{\gamma_x | x \in M\}$, called *of marker M* , where $\gamma_M(A)$, called *reconstruction opening*, is the union of connected components, or *grains* of A , that meet M .
3. Reconstruction openings and closings extend to numerical functions through flat operators (see an example in the article ► “Mathematical Morphology”). The set A (respectively, M) becomes the section of threshold t of the function F (respectively, of the marker function), one finds back criteria of the “flat zones” type. The alternating sequential filters of primitive connected openings and closings are themselves connected (see Fig. 1 above)

Connective Segmentation

In image processing, a numerical or multivalued function F is said to be segmented when its space of definition E has been partitioned into homogeneous regions, in the sense of some given criterion. The operation will be meaningful if the regions are as large as possible. But do all criteria lend themselves to such a maximum cut-out? Suppose for instance that we want to partition E into (connected or not) zones, where the numerical function f under study is Lipschitz of unity parameter. We run the risk of finding three disjoint zones A, B , and C such that the criterion will be satisfied on $A \cup B$ and on $A \cup C$, but not on $B \cup C$. In this case, there is no *largest region* containing the points of A and where the criterion is realized, so that the segmentation requires a non-deterministic choice between the partitions $\{A \cup B, C\}$ and $\{A \cup C, B\}$. This type of problem occurs typically in the “split and merge” approaches which have been used for more than 30 years in image segmentation.

How to sort out the criteria in order to be sure that they generate segmentation? First, we will make precise a few notions that are needed. Let E and T two arbitrary sets, and $\mathcal{F}$ a family of functions from $E \rightarrow T$. Then:

- A criterion $\sigma : \mathcal{F} \times \mathcal{P}(E) \rightarrow \{0, 1\}$ is a binary function such that for all $A \in \mathcal{P}(E)$ and all $F \in \mathcal{F}$ we have $\sigma[F, A] = 1$ when σ is satisfied by F on A and $\sigma[F, A] = 0$ when not. It is supposed that for all $F \in \mathcal{F}$ the criterion is always satisfied on the empty set: $\sigma[F, \emptyset] = 1$.
- A criterion σ is *connective* when:
 1. it is satisfied for the singletons:

$$\forall F \in \mathcal{F}, \forall x \in E, \sigma[F, \{x\}] = 1; \quad (5)$$

2. for all $F \in \mathcal{F}$ and all families $\{A_i | i \in I\}$ in $\mathcal{P}(E)$, we have:

$$\begin{aligned} \bigcap_{i \in I} A_i \neq \emptyset \text{ and } \bigwedge_{i \in I} \sigma[F, A_i] = 1 \\ \Rightarrow \sigma[F, \bigcup_{i \in I} A_i] = 1. \end{aligned} \quad (6)$$

- Finally, a criterion σ *segments* the functions of $\mathcal{F}$ when for all $F \in \mathcal{F}$, the family $\mathcal{D}(E, F, \sigma)$ of partitions of E whose all classes A satisfy $\sigma[F, A] = 1$, is non-empty and closed under *supremum* (including the empty one, $\mathcal{D}(E, F, \sigma)$ comprises the identity partition D_0). Then the partition $\bigvee \mathcal{D}(E, F, \sigma)$ defines the *segmentation of F according to σ* .

We saw that each connection on E partitioned every set $A \subseteq E$ into its maximum components. Therefore, if a criterion allows us to generate a connection associated to F , then we have good reason to think that it will segment the function. This is exactly what the following result states (Serra 2006):

Theorem of Segmentation *The following three statements are equivalent:*

1. *The criterion σ is connective*
2. *For each function $F \in \mathcal{F}$, the class of sets A where $\sigma[F, A] = 1$ is a connection*
3. *Criterion σ segments all functions of $\mathcal{F}$.*

Thus, a difficult problem (how to determine whether a class of partitions is closed under *supremum*?) comes down to the simpler question of checking whether a criterion is connective. Since this theorem turns out to be an alternative to the variational methods in segmentation, a short comparison is instructive.

1. In the connective approach, no differential operators intervene, such as Lagrange multipliers, for example. No assumption on the continuous or discrete nature of the space E is necessary: both E and T are arbitrary.
2. Moreover, not only the definition domain E of F is segmented, but even all the subsets of this domain. Therefore, if Y and Z are two masks in E , if $x \in Y \cap Z$, and if both segmentation classes $D_x(Y)$ and $D_x(Z)$ in x are included in $Y \cap Z$, then $D_x(Y) = D_x(Z)$ (in a traveling, the contours are preserved). The approaches based on global optimization of an integral in E are not able to obtain such a regional result.
3. The connective criteria form a complete lattice, where the *infimum* of the family $\{\sigma_i \mid i \in I\}$ is given by the Boolean minimum, what allows to parallelize the conditions. Note that if D_1 and D_2 are the segmentations associated with the two connective criteria σ_1 and σ_2 , then we have $\sigma_1 \leq \sigma_2$ if and only if $D_1 \leq D_2$, but the segmentation D relative to the $\inf \sigma = \sigma_1 \wedge \sigma_2$ satisfies only $D \subseteq D_1 \wedge D_2$. The petrographic example of Fig. 3 demonstrates an infimum of criteria.

Examples of Connective Segmentations

In what follows, we omit to repeat, for each criterion σ , that it is satisfied by the singletons and by the empty set (alternatively we could stop demanding that the singletons satisfy σ , and that would lead to a partially connective criterion (Ronse 2008)). Besides, the starting space E is supposed to be provided with an initial connection C_0 , which may intervene or not in the definition of the connective criterion under study, and the arrival space T is $\bar{R}$ or $\bar{Z}$.

The various connective segmentations can be classified into two categories, according to the presence, or the absence, of particular points, namely, the seeds. We begin with the second category, that we call the *simple* connective criteria.

Smooth Connection The criterion a for *smooth connection*, given by $\text{par } \sigma[F, A] = 1$ if and only if:

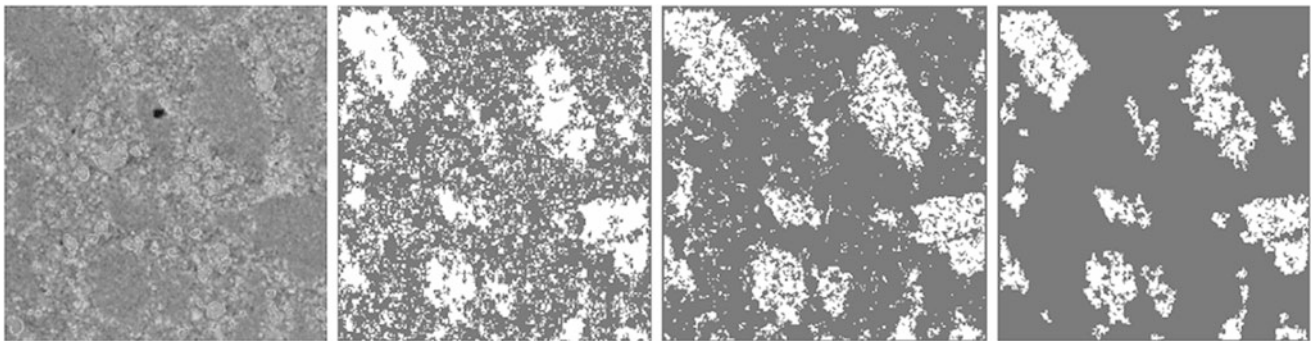
$$\forall x \in A, \exists a(x) > 0, \overset{\circ}{B}_{a(x)}(x) \subseteq A$$

$$\text{and } F \text{ is } k\text{-Lipschitz in } \overset{\circ}{B}_{a(x)}(x)$$

is obviously connective and induces the connection C_1 .

When C_0 is the arwise connection, the criterion $C = C_0 \cap C_1$ means that function F is k -Lipschitz along all paths included in the interior A° of A . In $\mathbb{Z}^2$, for example, where the smallest value of a is 1, it suffices, for segmenting, to erode the functions F and $-F$ by the cone $H(k, 1)$ whose base is the unit square (or hexagon), height is k , and summit the origin o , and then to take the intersection of the two sets where F (resp. $-F$) is equal to its eroded.

In the example of the micrograph of Fig. 3, one obtains for the smooth connection of slope 6 the white zones of Fig. 3(c), where the black ones indicate the isolated singletons. This connection differentiates the granular zones from those which are smoother, even when they both exhibit the same average gray value, as it often happens in electron microscopy.



Morphological Filtering, Fig. 3 From left to right: Electron micrograph of concrete, segmentation for a jump connection of value 12, segmentation by smooth connection of slope 6, infimum of the jump and smooth criteria

Connections by Clustering on Seeds Many segmentation processes work by binding together points of the space around an initial family $G_0 \subseteq E$ of seeds, which may possibly move, or whose number may vary when the process progresses during several iterations. All these processes satisfy the following property:

Theorem of Seeds *Given a function $F \in \mathcal{F}$, an initial family G_0 of seeds, and an aggregation process that yields the final seeds G , the criterion σ obtained by:*

$$\sigma[F, A] = 1$$

if all points of A are allocated to a same final seed $g \in G$, and $\sigma[F, A] = 0$ otherwise, is connective.

Watershed lines: We now interpret the numerical function F as a relief in R^n or Z^n . All points of the relief whose deepest descent line ends to a same minimum form an arcwise connected catchment basin, and the singletons which do not belong to any catchment basin define the watershed (Lantuéjoul and Beucher 1981). There exist several method to obtain a watershed. One can notice that for binary images, the watershed is nothing but the SKIZ of Lantuejoul, or skeleton by zones of influence. It extends to numerical functions by level set approach, the portion of watershed of level i being the geodesic SKIZ of level i inside level $i - 1$ (Meyer and Beucher 1990; Najman et al. 2012) (see also [Serra 1982] in the article ► “Mathematical Morphology”). Alternatively, one can determine the steepest descents of the points towards minima (Cousty et al. 2009; Bertrand 2005). The reference [Najman and Talbot 2010] in the article ► “Mathematical Morphology” comprises several chapters on the subject.

Whatever the exact way, the watershed line has been introduced, the theorem of seeds shows that the involved criterion is connective. Now, among the singleton components we find of course the crest line, but also all points x of the intermediate flat zones, like stairs, stuck between an upstream and a downstream, and where F is constant on an

open set surrounding x . We then take for second connective criterion that “each point x of an intermediate flat zone goes to the same catchment basin as the point of the downstream frontier closest to x .” This criterion is applied to the set of singletons; by repeating the process as many times as needed for going up from stair to stair, one finally reduces the singleton zones to the true crest lines.

Remark that the watershed of function F gives the zones of influence of the minima of F , as can be shown in Fig. 4 (first two left images), and not the contours of the objects. The latter appear as the watershed lines of the $|\text{gradient}(F)|$ image, as one can see in Fig. 4 (last two left images). Moreover, for the sake of robustness, the gradient image is stamped here by the desired minima, which are both minima and watershed lines of F .

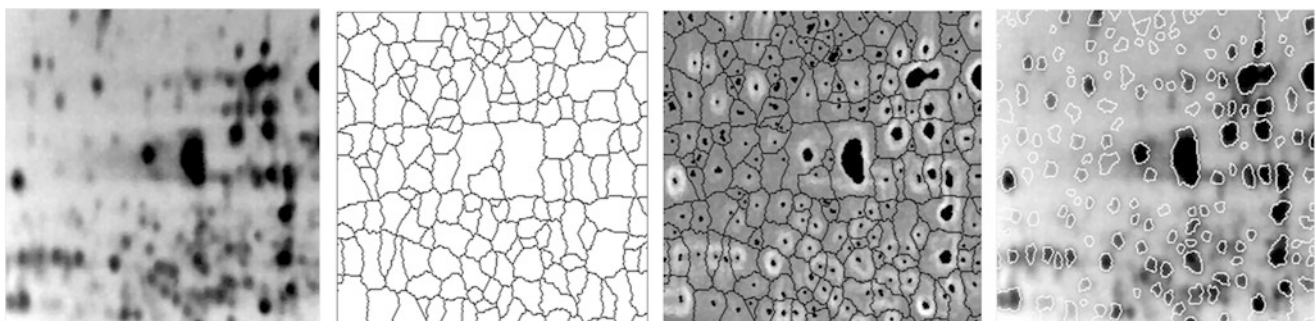
Jump connection: It is a kind of alternative to watershed, which often leads to excellent segmentation. The jump connection works by aggregating seeds, each one being a connected basin of given height around a minimum. The space is supposed to be equipped with an initial connection C_0 , and a jump value $k > 0$ is fixed. Around the supports M of the minima, we build the largest connected sets $S_1(M)$ where the level goes up from 0 to $k - 1$.

By binding the $S_1(M)$ one obtains the components associated to the first application of the jump criterion. A second application of the criterion, this time on the residual, gives new components $S_2(M)$, etc. until nk exceeds the dynamic of values of the function (Fig. 5). The family $\{S_i(M)\}$ is the set of classes of the final jump partition. Alternatively, one can jump up from the minima, and down from maxima, in a symmetrical manner (Serra 2006).

Returning to the micrograph of Fig. 3, we observe that the smooth connection and the jump connection taken alone are not excellent, but that the infimum of the two criteria suppresses the noise, and gives a satisfactory result.

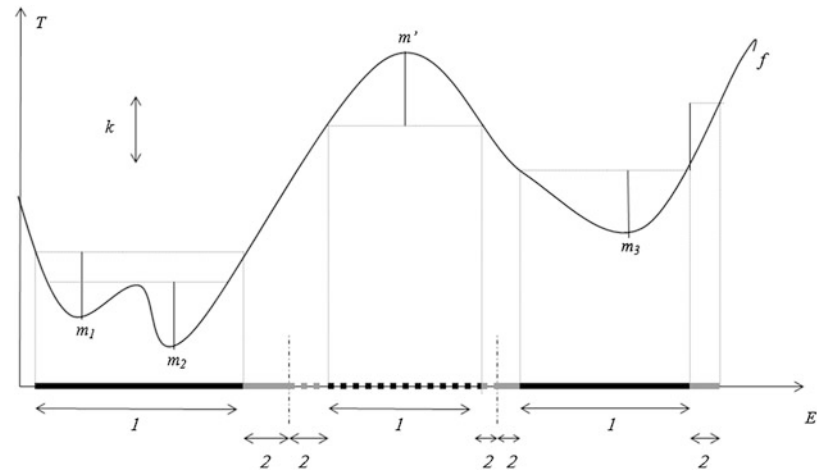
Compound Segmentation

Instead of parallelizing criteria, one can also take them into account successively, and segment differently various regions



Morphological Filtering, Fig. 4 Watershed of an electrophoresis and of its derivative

Morphological Filtering,
Fig. 5 The first two steps of a
 jump connection



Morphological Filtering, Fig. 6 From left to right: initial photo, segmentation of the face (color criterion), marker for the bust, final segmented silhouette. (By courtesy of C. Gomila)

of an image. For this purpose, one relaxes the first axiom of connective criteria, namely, relation (5). This condition aims at guaranteeing that every point of the space belongs to a class of the performed segmentation. By breaking away from relation, one thus renounce the covering of a set by the classes of a partition or by its connected components. This approach, formalized by C. Ronse (2008), is based on partial connections and leads to more flexible processing.

In the example of Fig. 6, one tries to segment silhouettes of the person in Fig. 6 (left). A first partial segmentation, based on the hue of the skin and the hair, leads to the head contour. A second segmentation that deals only with the complement of the head, extracts next the shoulder from the marker formed by three superposed rectangles (minus the already segmented points). The union of the two segmentations leads to the mask in Fig. 6 (right).

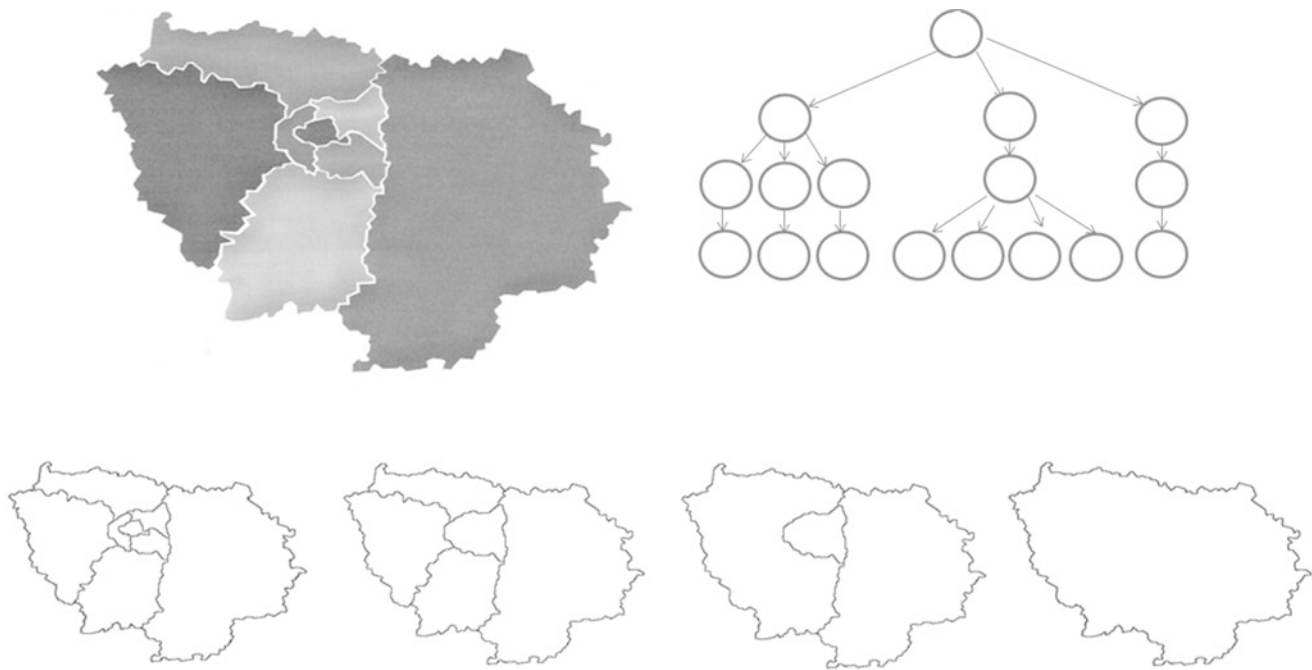
Hierarchies of Partitions

From now on we concentrate upon the morphological filtering and segmentation defined on finite hierarchies of partitions in R^2 or Z^2 , for whatever origin. They can be due to anterior connected filters, or not. Figure 7 (left), for example,

represents administrative divisions around the city of Paris, which generate a hierarchy of four partitions according to some parameter. The hierarchies of same base, same top, and same number of levels form a complete lattice for the product ordering of their partition levels.

A hierarchy can be summarized in two ways, depending on whether we focus on the edges or on the classes. Both ways are equivalent but serve differently. Firstly, one can weight the edges by the level when they disappear. That condenses the hierarchy into a unique map of the so-called *saliencies*. The saliency maps lend themselves to morphological filtering on hierarchies. For example, the operation which suppresses all edges of saliency $\leq \lambda$, or shorter than μ , and leaves unchanged the others, is a closing and it generates an anti-granuometry as λ increases (see Ch.9 in the reference [Najman and Talbot 2010] of the article ► “Mathematical Morphology”).

Alternatively one can replace the hierarchy by its *dendrogram*, as depicted in Fig. 7 (right), i.e., by a tree whose levels are those of the hierarchy. The top level is called the root and the lowest level the leaves. The classes, or nodes, are indicated by small discs and their ordering is described by arrows. The parameters associated with the classes are usually written



Morphological Filtering, Fig. 7 Displays of a hierarchy

in the small discs. The dendrogram highlights the “sons” of each class S , i.e., the partial partition just below S . The relation between classes and sons directly intervenes when looking for minimal cuts.

Formally speaking, a partition D is a map $E \rightarrow \mathcal{P}(E) : x \mapsto D(x)$ that satisfies the following two conditions:

- for all $x \in E$, $x \in D(x)$;
- for all $x, y \in E$, $D(x) \cap D(y) \neq \emptyset \Rightarrow D(x) = D(y)$.

$D(x)$ is called the *class* of the partition in x . Partition D is *finer* than partition D' when each class of D is included in a class of D' . One writes $D \leq D'$. This relationship defines the *refinement* ordering, which provide partitions with the structure of a complete lattice (see article ► “[Mathematical Morphology](#)”). A hierarchy is an ordered sequence of partitions (e.g., Fig. 7 bottom). In the following, we start from a unique hierarchy H and consider the family $\mathcal{H}$ of all the hierarchies whose classes are taken among those of H . Similarly, we name *cut through hierarchy* H any partition D whose all classes are taken among the classes of the hierarchy and we denote by $\mathcal{D}$ the set of all cuts of H . Finally, when the partitioning is restricted to a class S of the hierarchy H , one speaks of *partial partition* of support, or nod, S .

Optimal Cuts

Suppose we want to summarize the hierarchy into a significant partition. The information conveyed by the parameters on the classes may be pertinent for some classes at a certain level and for others at another level. In order to express this

pertinence, one classically associates an energy ω with each partial partition of H . Three questions arise here, namely:

1. Most of the segmentations involve several features (color, shape, size, etc.). How to combine them in unique energy?
2. How to generate, from the set of all partial partitions, a partition that minimizes ω , and which can be determined easily?
3. When one energy ω depends on an integer λ , i.e., $\omega = \omega_\lambda$ how to generate a sequence of optimal partitions that increase with λ , which therefore should form an optimal hierarchy?

These questions have been taken up by several authors, over many years, and by various methods. Morphological approach unifies them and allows to go further. We will successively answer the three above questions.

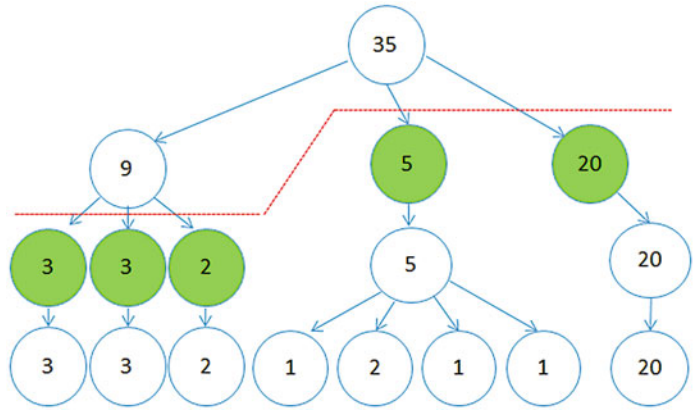
Dynamic Programming

The most popular energies ω for hierarchical partitions derive from that of Mumford and Shah (Arbelaez et al. 2011; Guigues et al. 2006; Salembier and Garrido 2000). It is defined in each class as the sum of a fidelity term (e.g., the variance of the values) and a boundary regularization term (e.g., the length of the edges). The energy of a partial partition is the sum of those of its classes. The optimization turns out to be a trade-off between fidelity and regularization.

In 1984, L. Breiman (Breiman et al. 1984) introduced a dynamic programming method to determine the cut of minimum energy in a hierarchy. The hierarchy is spanned only

Morphological Filtering,

Fig. 8 Breiman's minimal cut over the hierarchy: the energy ω at node S is compared to the sum of the energies of its sons T_S . When $\omega(S) \leq \sum \omega(T_S)$ one keeps S , when not, one replaces S by its sons



once, from the leaves to the root. The energy of each node S is compared to the sum of the energies of its children, and one keeps the partial partition which has the least energy. In case of equality one keeps S . Fig. 8 depicts a dendrogram with the energies of the classes and the minimal cut, i.e., the partition with the least total energy.

However, one can wonder whether additivity is the very underlying cause of the simplifications via a unique spanning, since Soille's constrained connectivity (Soille 2008), where the addition is replaced by the supremum, satisfies similar properties. Finally, one finds in literature a third type of energy, which holds on nodes only, and no longer on partial partitions. It appears in labeling methods (Arbelaez et al. 2011), and again, it yields optimal cuts. Is there a common denominator to all these approaches, more comprehensive than just additivity, and which explains why they always lead to unique optima?

Energetic Lattices

Consider a given hierarchy H , and the family $\mathcal{D}$ of all its cuts. An energy ω is assigned to all partial partitions of H whose support S is a class of H . Let $D_1, D_2 \in \mathcal{D}$, be two cuts. At point x , either the class $D_1(x)$ is the node $S_1(x)$ support of a partial partition $D_2(x)$ of D_2 , or we have the inverse. We then say that cut D_1 is ω -smaller than D_2 when

$$\omega[D_1(x)] = \omega[D_2(x)] \text{ and } D_2(x) = S_1(x).$$

These two conditions define an order relation over the set of the cuts $\mathcal{D}$ of hierarchy H , which in turn induces a complete lattice on the cuts $\mathcal{D}$. Both ordering and lattice are said *energetic* (Kiran and Serra 2014). The structure of the cuts as an energetic lattice guarantees us that there is one and only one minimal cut, which is not useless (see Fig. 1 in (Kiran and Serra 2015)).

 h -Increasingness and Partition Closing

The common property of all energies which make use of dynamic programming is that they all are h -increasing

(Kiran and Serra 2014). Let D_1 and D_2 be two partial partitions of the same support S , and D_0 a partial partition of support S_0 disjoint of S . An energy ω over the partial partitions is said to be h -increasing when:

$$\omega(D_1) \leq \omega(D_2) \Rightarrow \omega(D_1 \sqcup D_0) \leq \omega(D_2 \sqcup D_0)$$

where the symbol " $\sqcup$ " stands for the concatenation of two partial partitions.

When the energy ω is h -increasing, and only then, each step of the dynamic programming is an increasing and extensive operation on the family $\mathcal{H}$ of hierarchies, so that at the end, when the root is reached, we obtain the *closing* of H w.r.t. energy ω . This closing is a hierarchy whose all levels are identical, and equal to the unique minimal cut.

h -increasing energies include, of course, the classical additive energies, but also the composition laws by infimum, harmonic sum, number of classes, quadratic sum, and supremum, among others.

Families of Minimal Cuts

An energy ω_δ over the partial partitions is said *inf-modular* when for each partial partition D of support S we have

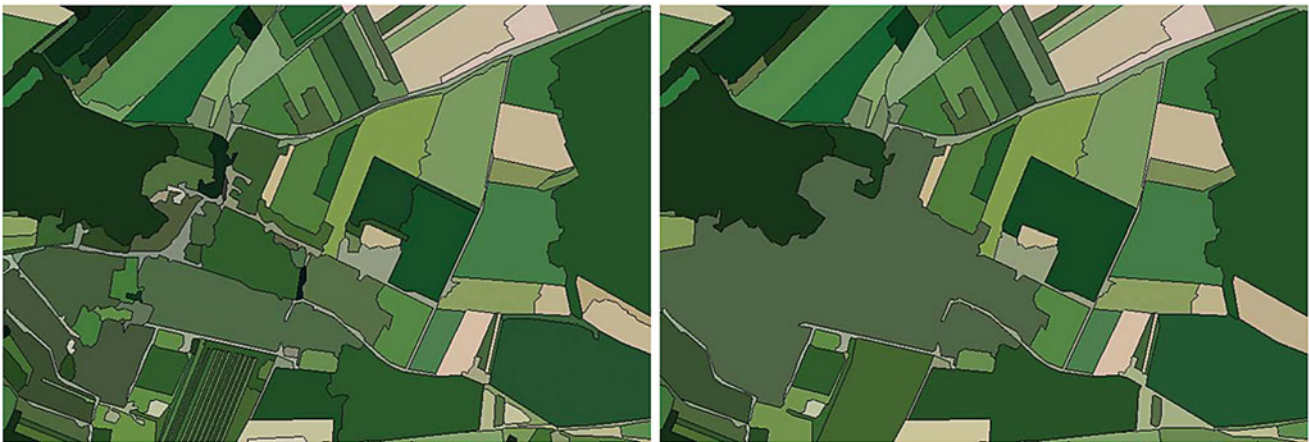
$$\omega_\delta(\{S\}) \leq \wedge \omega_\delta[D(x), x \in S].$$

The Lagrange formalism proposes a convenient way to order a family of minimal cuts (Guigues et al. 2006; Soille 2008). Introduce a scalar Lagrange family of energies by $\{\omega(\lambda) = \omega_\varphi + \lambda \omega_\delta, \lambda > 0\}$ where $\omega(\lambda)$, ω_φ , and ω_δ are h -increasing, and further ω_δ is inf-modular.

These energies are defined over partial partitions, and we suppose $\lambda > 0$. Given hierarchy H , the three energies $\omega(\lambda)$, ω_φ , and ω_δ induces the three energetic lattices over the set $\mathcal{D}$ of all cuts of hierarchy H , namely, $\mathcal{D}_\lambda$, $\mathcal{D}_\varphi$, $\mathcal{D}_\delta$. Just as for the constrained minimization for functions, the energy $\omega(\lambda)$ stands for the Lagrangian of objective term ω_φ and constraint term ω_δ . The constrained Lagrangian minimization results then in the minimal cut of H w.r.t. energy $\omega(\lambda)$. This minimal



Morphological Filtering, Fig. 9 Left: initial view; right: minimal cut for an additive energy. (By courtesy of L. Guigues)



Morphological Filtering, Fig. 10 Two cuts of the minimal hierarchy

cut increases with λ for the refinement order, so that it produces a new hierarchy H^* (Kiran and Serra 2015). Remark that one does not change the H^* by iterating the process.

Figs. 9 and 10 give an example of the optimal hierarchy H^* . It depicts the evolution of the optimal cuts when ω_φ and ω_δ are the variance of the values and the length of the edges in a partial partition, respectively. The parameter λ increases when going from left to right, and the minimal cut of Fig. 9 (right) corresponds to a transitional λ . We see that for the greatest λ , the fields are well rendered, but the region of the village, which has a high length of edges, is swept out. Here a better technique would probably be the Ronse compound segmentation (Ronse 2008), as demonstrated in Figs. 6 and 10.

Summary

The contribution of Morphological Filtering to geosciences and more generally to image processing can be contemplated

from the two points of view of theory and practice. Several notions used in qualitative description receive a precise meaning in mathematical morphology, owing to adapted axioms. Morphological filtering: turns out to be a sort of trade-off between its two axioms of increasingness, which simplifies the structures under study, and idempotence, which blocks the simplification. The axioms of a connection directly apply to the objects which are connected in the usual sense, but also to those which may seem disconnected at first glance, like the dotted lines of a trajectory. When a connection is added to the filters, then they come to segmentation methods. They extract, from numerical functions, partitions of the space into regions where these functions are homogeneous in the sense of a given connective criterion. In image processing this process is called a segmentation. When it depends on a positive parameter, segmentation results in a sequence of increasing partitions called a hierarchy. Then a closing filter on the hierarchy allows us to extract its “best” segmentation.

Some of the above examples belong to petrography or to remote sensing, thus come within the domain of geosciences.

But the others, like district grouping, or zones of influence of spots, can easily be transposed to geosciences questions. Morphological Filtering intervenes, indeed, at all scales and includes hydrology, generation of maps, fractal landscape description, etc. (Sagar 2013).

Cross-References

- [Mathematical Morphology](#)
- [Matheron, Georges](#)

Bibliography

- Arbelaez P, Maire M, Fowlkes C, Malik J (2011) Contour detection and hierarchical image segmentation. *IEEE Trans PAMI* 33(5):898–916
- Bertrand G (2005) On topological watersheds. *J Math Imag Vis* 22(2–3):217–230
- Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey
- Cousty J, Bertrand G, Najman L, Couprie M (2009) Watershed cuts: minimum spanning forests and the drop of water principle. *IEEE Trans Pattern Analysis Machine Intel* 31(8):1362–1374
- Guigues L, Cocquerez JP, Men HL (2006) Scale-sets image analysis. *Int J Comput Vis* 68(3):289–317
- Kiran BR, Serra J (2014) Global-local optimizations by hierarchical cuts and climbing energies. *Pattern Recogn* 47(1):12–24
- Kiran BR, Serra J (2015) Braids of partitions. In: LNCS 9082 mathematical morphology and its applications to signal and image processing. Springer, Berlin Heidelberg, pp 217–228
- Lantuéjoul C, Beucher S (1981) On the use of the geodesic metric in image analysis. *J Microsc* 121:39–49
- Matheron G (1988) Chapter 6, Filters and lattices. In: Serra J (ed) Image analysis and mathematical morphology. volume 2: theoretical advances. Academic Press, London, pp 115–140
- Meyer F, Beucher S (1990) Morphological segmentation. *J Vis Commun Image Represent* 1(1):21–46
- Najman L, Talbot H (2010) Mathematical morphology: from theory to applications. Wiley, New York
- Najman L, Barrera J, Sagar BSD, Maragos P, Schonfeld D (eds) (2012) Filtering and segmentation with mathematical morphology. *IEEE J Sel Topics Signal Process* 6(7):737–738
- Ronse C (2008) Partial partitions, partial connections and connective segmentation. *J Math Imag Vis* 32(2):97–125. <https://doi.org/10.1007/s10851-008-0090-5>
- Sagar BSD (2013) Mathematical morphology in geomorphology and gisci. CRC Press, Taylor and Francis Group, A Chapman and Hall Book, London
- Salembier P, Garrido L (2000) Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval. *IEEE Trans Image Process* 9(4):561–576
- Salembier P, Serra J (1995) Flat zones filtering, connected operators, and filters by reconstruction. *IEEE Trans Image Process* 4(8):1153–1160
- Serra J (1982) Image analysis and mathematical morphology. Academic Press, London
- Serra J (ed) (1988) Image analysis and mathematical morphology. volume 2: theoretical advances. Academic Press, London
- Serra J (2000) Connections for sets and functions. *Fundamenta Informaticae* 41(1/2):147–186
- Serra J (2006) A lattice approach to image segmentation. *J Math Imag Vis* 24(1):83–130. <https://doi.org/10.1007/s10851-005-3616-0>
- Soille P (2008) Constrained connectivity for hierarchical image partitioning and simplification. *IEEE Trans Pattern Anal Mach Intell* 30(7):1132–1145
- Wilkinson MH (2006) Attribute-space connectivity and connected liters. *IVC* 25(4):426–435

Morphological Opening

Sravan Danda¹, Aditya Challa¹ and B. S. Daya Sagar²

¹Computer Science and Information Systems, APPCAIR, Birla Institute of Technology and Science, Pilani, Goa, India

²Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, Karnataka, India

Definition

Morphological opening is one of the basic operators from the field of Mathematical Morphology (MM). It is also one of the simplest morphological filters used for constructing more complex filters. This operator is dual to Morphological Closing. In general, morphological opening is defined as a composition of an erosion followed by a dilation. So, in the case of discrete binary images, if X denotes the binary image and B denotes the structuring element, morphological opening is defined as

$$\gamma_B(X) = \delta_B(\varepsilon_B(X)) = (X \ominus B) \oplus B \quad (1)$$

This definition can be reduced to

$$\gamma_B(X) = \cup\{B_x | B_x \subseteq X\} \quad (2)$$

Equivalently, in the case of gray-scale images, if f denotes the gray-scale image and g denotes the structuring element, opening is defined as

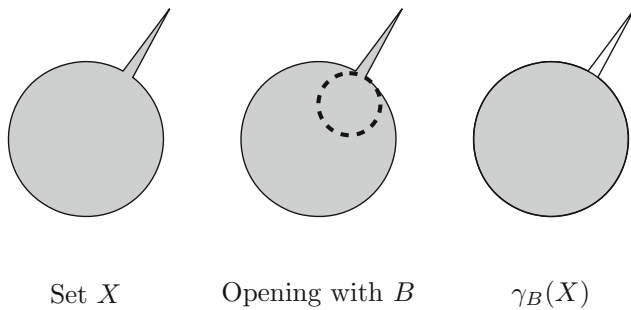
$$\gamma_g(f) = \delta_g(\varepsilon_g(f)) \quad (3)$$

Illustrations

Morphological opening is one of the basic operators from the field of Mathematical Morphology. Its definition can be traced to fundamental works by Jean Serra (1983, 1988) and Georges Matheron (1975). In general, morphological opening is defined as an erosion operator composed with a dilation operator. In this entry, we illustrate the opening operator using binary images, and discuss various properties. We also discuss the more general algebraic openings.

In simple words, (2) considers the union of all the translations of the structuring element which fits into the original set. This is illustrated in Fig. 1. Here, the structuring element is taken to be a simple disk denoted by a dashed line in Fig. 1. All the points that do not fit within the structuring element are removed. In Fig. 1, this corresponds to the spiked projection of the larger disk.

Although not discussed here, similar intuition follows in gray-scale images as well.



Morphological Opening, Fig. 1 Illustration of Morphological Opening. The shaded area refers to the set obtained by opening

Some Important Properties

The properties of the opening operator are explained using the binary images. However, these extend to the gray-scale images as well (see Dougherty and Lotufo (2003)). Firstly, morphological opening is an *increasing operator*,

$$X \subseteq Y \Rightarrow \gamma_B(X) \subseteq \gamma_B(Y) \quad (4)$$

Moreover it is also an *anti-extensive operator*

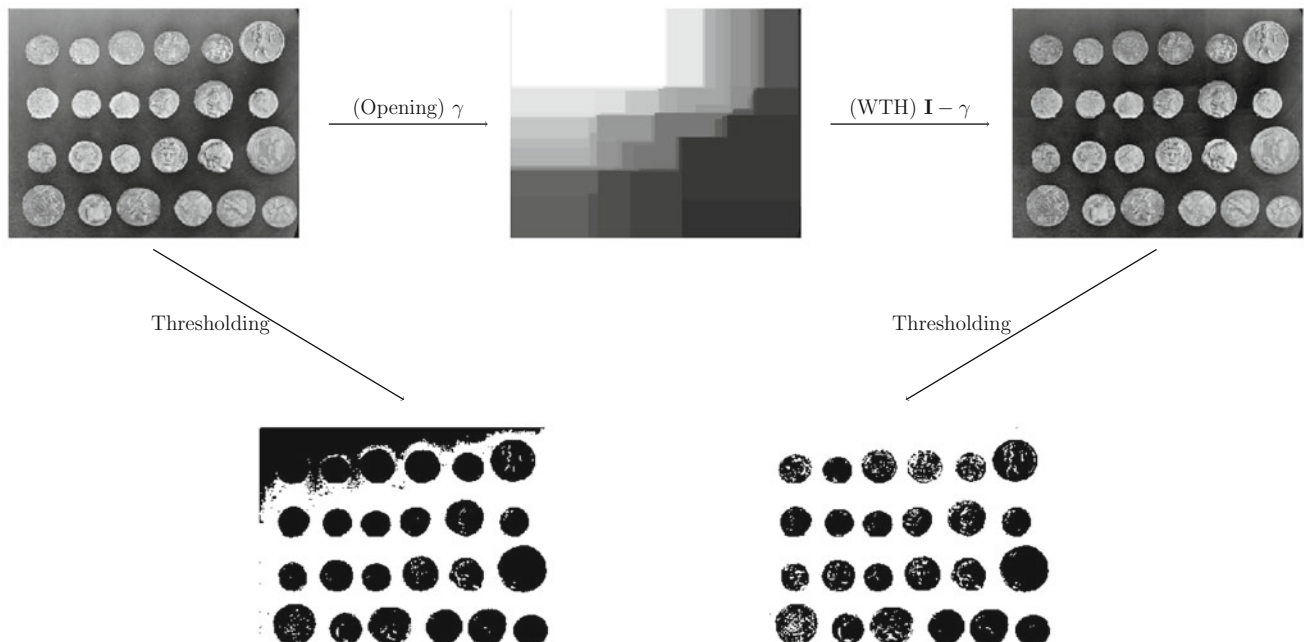
$$X \supseteq \gamma_B(X) \quad (5)$$

Also, the opening operator is idempotent

$$\gamma_B(\gamma_B(X)) = \gamma_B(X) \quad (6)$$

Algebraic Openings

Morphological openings are a subset of a wider class of operators known as *Algebraic Openings*. An algebraic opening is defined as any operator on the lattice which is: (a) increasing (b) anti-extensive, and (c) idempotent. It is easy to see that morphological opening is a specific example of algebraic openings.



Morphological Opening, Fig. 2 Application of Morphological Opening. Using the white top-hat transform one can remove the illumination effects. https://scikit-image.org/docs/dev/auto_examples/color_

[exposure/plot_regional_maxima.html#sphxglr-download-auto-examples-color-exposure-plot-regional-maxima-py](https://scikit-image.org/docs/dev/auto_examples/color_exposure/plot_regional_maxima.html#sphxglr-download-auto-examples-color-exposure-plot-regional-maxima-py). Copyright: 2009–2022 the scikit-image team. License: BSD-3-Clause

An example of algebraic opening, which is not a morphological opening, is *Area Opening* – remove all connected components of the image that has an area lesser than the parameter λ . (Recall that in a binary image, two adjacent pixels x, y are said to be connected if they have the same value. A path between two pixels x, y is a sequence of pixels $x = x_0, x_1, \dots, x_k = y$ such that x_i and x_{i+1} is connected. A subset is said to be connected if for any two pixels in the subset there exists a path between them. A connected component is a maximal subset which is connected.)

In Matheron (1975) it is shown that any algebraic opening can be defined as a supremum of morphological opening.

Application to Filtering

Recall that, in the context of mathematical morphology, an operator which is increasing and idempotent is referred to as a *filter*. As stated earlier, morphological opening is one of the simplest filters based on which complex filtering operators such as top-hat transforms, granulometries, and alternating sequential filters are constructed. Here we discuss the example of *White top-hat* transform which is defined as

$$\text{WTH}(f) = f - \gamma(f) \quad (7)$$

or simply $\text{WTH} = \mathbf{I} - \gamma$ where $\mathbf{I}$ is the identity operator. The effect of this operator is illustrated in Fig. 2. A simple thresholding of the original image results in dark patches. This is due to the bad illumination of the image. The dark patches can be removed by opening with a large structuring element and subtracting the result from the original image, as White-top-hat transform. By thresholding the filtered image, the illumination effects can be seen filtered out.

Summary

In this entry, morphological opening is defined for binary and gray-scale images. The practical utility of morphological opening is then illustrated using a fictitious binary image. The main properties of morphological opening, namely, increasing, anti-extensiveness and idempotence are then mentioned. A generalized notion of morphological opening, i.e., algebraic opening is defined. The entry concludes by describing the construction of more complex filter: White-top-hat transforms using morphological opening, along with illustrations on real images.

Cross-References

- [Mathematical Morphology](#)
- [Morphological Filtering](#)
- [Structuring Element](#)

Bibliography

- Dougherty ER, Lotufo RA (2003) Hands-on morphological image processing, vol 59. SPIE Press, Bellingham
- Matheron G (1975) Random sets and integral geometry. AMS, Paris
- Serra J (1983) Image analysis and mathematical morphology. Academic, London
- Serra J (1988) Image analysis and mathematical morphology, Theoretical advances, vol 2. Academic, New York

Morphological Pruning

Sin Liang Lim¹ and B. S. Daya Sagar²

¹Faculty of Engineering, Multimedia University, Cyberjaya, Selangor, Malaysia

²Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, Karnataka, India

Definition

In digital image processing, skeletonization and thinning algorithms are procedures that tend to leave parasitic components, or “spurs” behind. As a result, pruning is a technique that could be adopted as an essential procedure to eliminate these unwanted spurs. The spurs, in this case, refer to unwanted branches of line which should be removed since, by removing them, the overall shape of the line is not affected. These parasitic components are the undesirable result generated from edge-detection algorithms (e.g., character recognition, car plate detection) or digitization. Very often, these spurs can be treated as “noise” and thus they should be cleaned-up before proceeding to the subsequent procedure.

Introduction

Pruning is an important post-processing operator that complements skeletonization and thinning algorithms because these operations tend to leave spurs that need to be eliminated (Gonzalez and Woods 2008; Shaked and Bruckstein 1998). Skeletons are concise descriptors of objects in images (Duan et al. 2008; Maragos and Schafer 1986). Skeletonization, on

the other hand, often provides a compact yet effective representation of two-dimensional and three-dimensional objects. Skeletonization is commonly applied in many low-level and high-level image-processing applications which include object representation, tracking, recognition, and compression. Furthermore, skeletonization also facilitates efficient description of geometry, topology, scale, and other local properties related to the object (Saha et al. 2017; Baja 2006). Particularly, skeletonization is used in automated character recognition including handwritten characters recognition to extract the skeleton of each character (Gonzalez and Woods 2008). However, skeletonization will often result in not only the skeleton but also the parasitic components, or “spurs.” Spurs are created as a result of erosion due to the non-uniformities in the strokes that are possessed by the characters.

Methodology

In Gonzalez and Woods (2008), a morphological-based pruning framework was developed by assuming that the length of a parasitic component does not exceed a certain number of points. In other words, the standard pruning algorithm will eliminate all branches shorter than a specific number of pixels. For example, if a parasitic branch is shorter than three points and the thinning algorithm is repeated for three times, then the parasitic branch will be removed. The framework comprises four steps: (i) thinning, (ii) finding end points, (iii) dilating end points, and (iv) union. It is highlighted that step (ii) is performed to ensure that the main branch of each line are preserved and not shortened by the procedure. The algorithms and the example in Fig. 1 below are adapted with permission from Digital Image Processing, fourth Edition by Gonzales and Woods, Pearson 2018.

Step 1: Thinning

$$X_1 = A \otimes \{B\} \quad (1)$$

Thinning of set A with a sequence of structuring element $\{B\}$ is performed, as shown in Eq. 1.

As an example, Fig. 1a shows the skeleton of a hand-printed letter “a” and it is denoted as set A . It is observed that the leftmost part of the character (three pixels in vertical) represents the “spur” which should be removed. This leads us to assume that any branch with three or less pixels should be removed.

Figure 1b represents the ordered list of 8 structuring elements which is rotated 90° and each of which contains two pixels. The “ $\times$ ” symbol denotes a “don’t care” condition. In order to remove any branch with n or less pixels, Step 1 has to

be repeated for n times. In this case, since the spur is determined to be three pixels, thus Step 1 is repeated three times to generate set X_1 in Fig. 1c.

Step 2: Find End Points

$$X_2 = \bigcup_{k=1}^8 (X_1 \ominus B^k) \quad (2)$$

In this step, a set X_2 containing all end points in X_1 is constructed. In Eq. 2, B^k denotes end-point detectors which are used to determine the end points in X_1 using the hit-and-miss operation of B^k with X_1 . End points are defined as the points where the origin of the structuring elements are satisfied. Fig. 1d shows the resultant X_2 which consists of two end points.

Step 3: Dilate End Points

$$X_3 = (X_2 \oplus H) \cap A \quad (3)$$

The next step is to perform dilation of the end points as conditioned on A , as given in Eq.3. In particular, the end points is dilated using a 3×3 matrix (H) which contains all 1’s and intersects with A . This step is executed n times in all directions for each endpoint.

In this example, as no new spurs are created by inspection, the end points are dilated three times with reference to A as delimiter. This is the same number as the thinning process in Step 1. Fig. 1e represents the resultant X_3 .

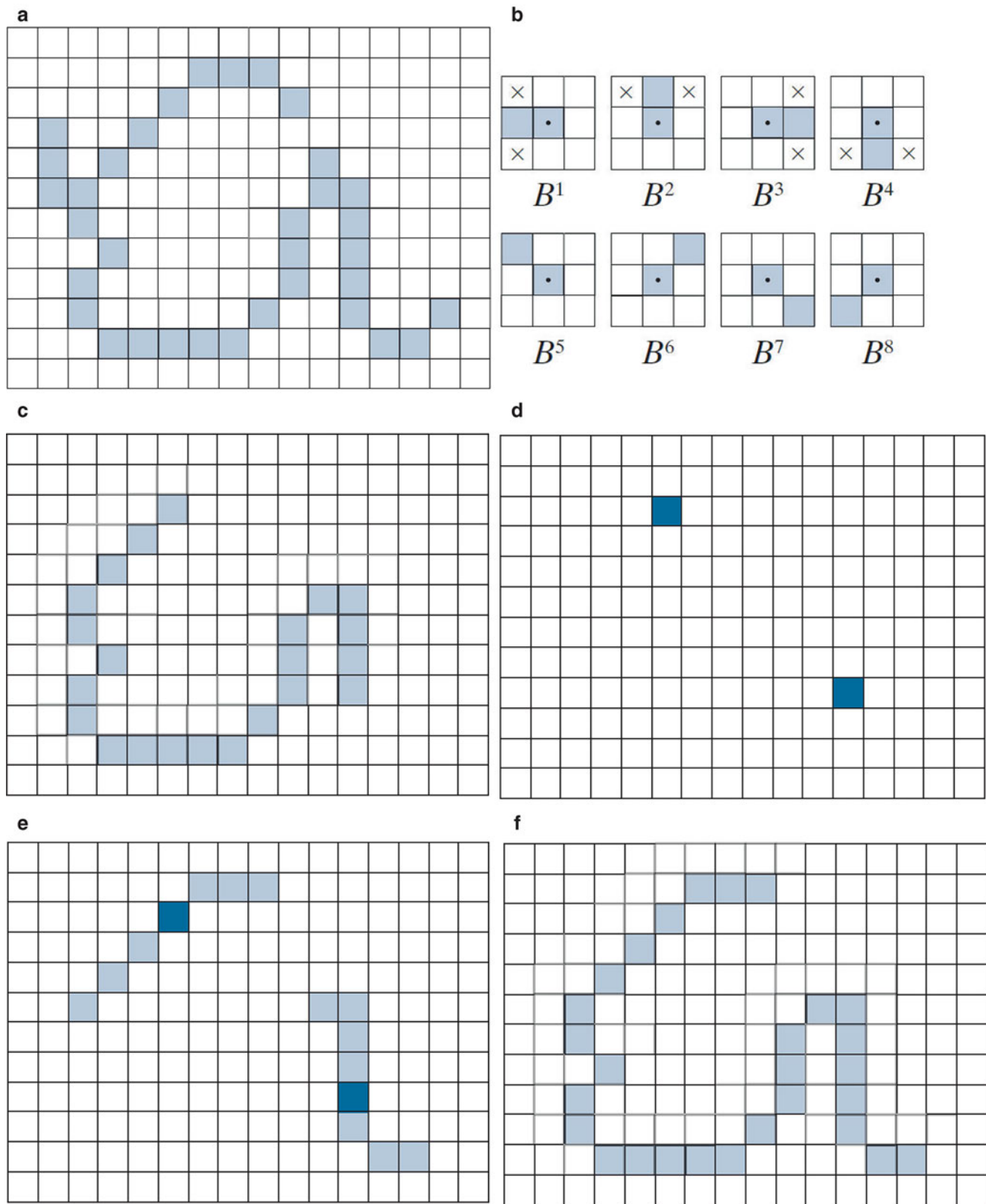
Step 4: Union of X_1 and X_3

$$X_4 = X_1 \cup X_3 \quad (4)$$

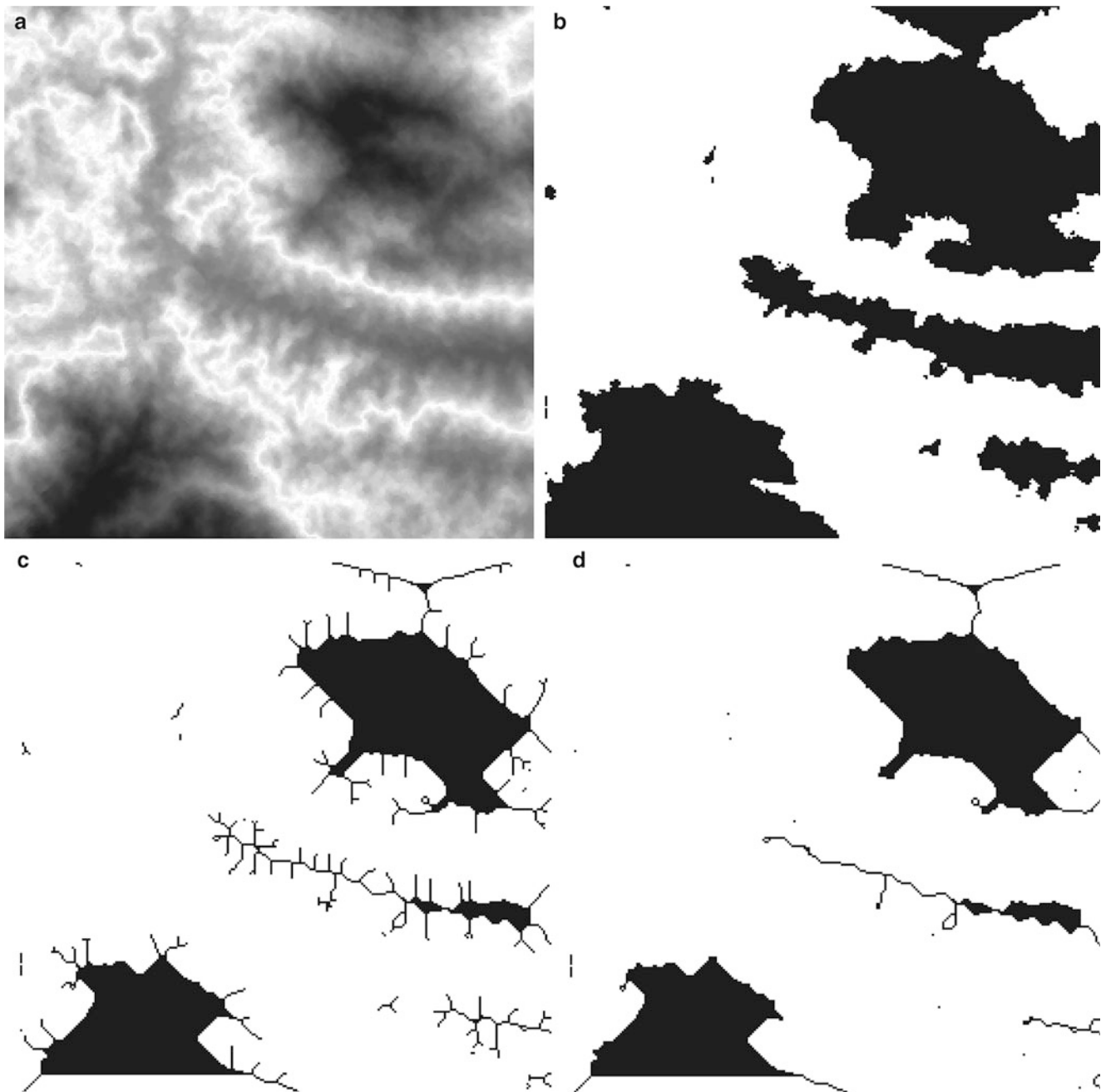
The purpose of the last step is to restore the character to its original form but with the spurs removed. In essence, the union of X_1 and X_3 will produce the final pruned image X_4 , as shown in Fig. 1f.

Case Studies

For simple illustration, the ASTER Global Digital Elevation Model (ASTER GDEM) of the mountainous area in Cameron Highlands, Pahang in Malaysia is taken as the study area for extracting the pruned image. With elevation up to 1600 meters above sea level, the study area covers the tea plantation, vegetable farm, and various other agricultural activities. The study area encompasses 400 km^2 with coordinates of $4^\circ 22' 57'' \text{ N}$ to $4^\circ 33' 50'' \text{ N}$ and $101^\circ 22' 4'' \text{ E}$ to $101^\circ 32' 50'' \text{ E}$ (Lim et al. 2016; Kalimuthu et al. 2016).



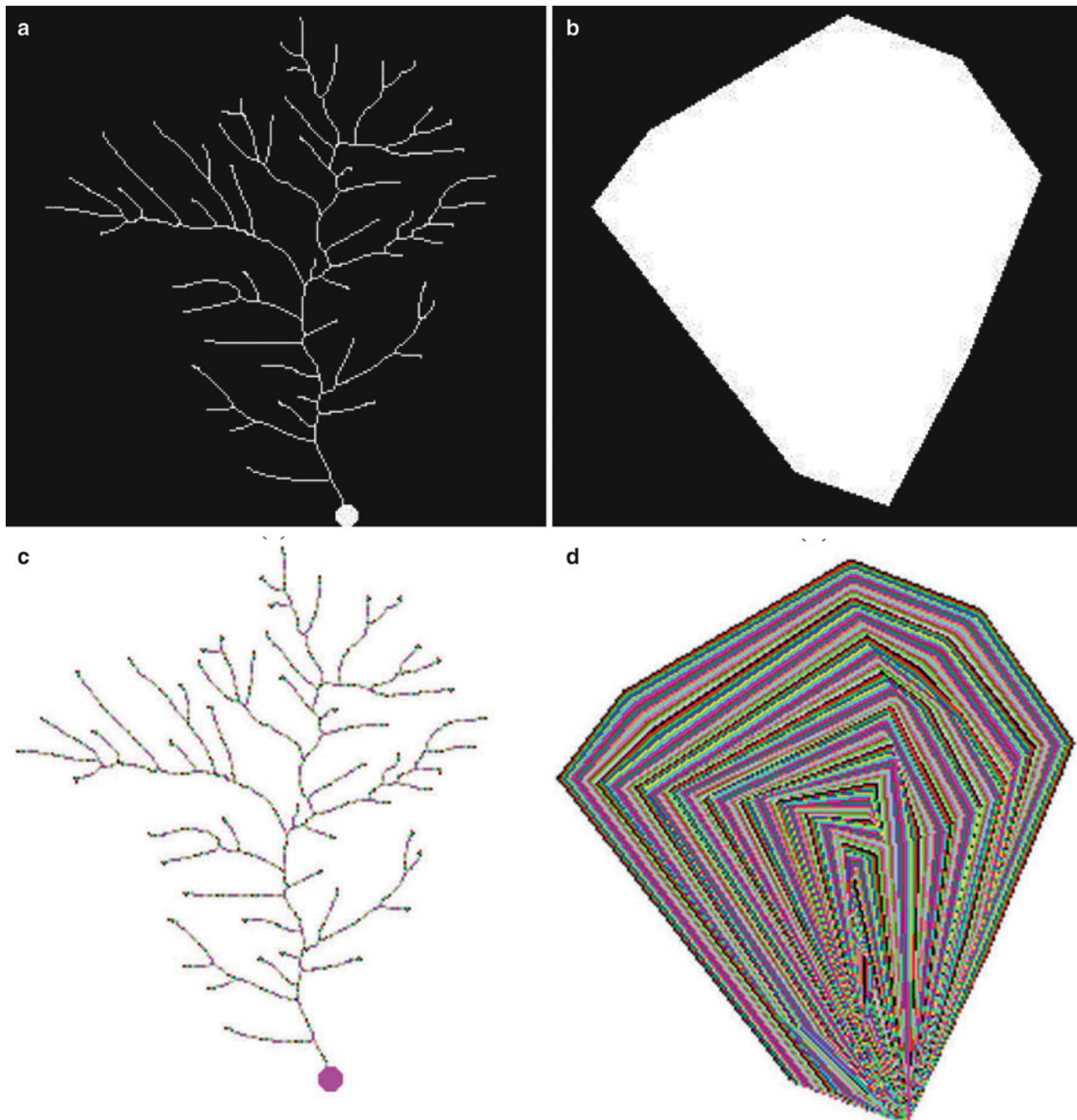
Morphological Pruning, Fig. 1 (a) Set A , (b) sequence of structuring elements B used for detecting end points, (c) X_1 - after three cycles of thinning, (d) X_2 - end points detected, (e) X_3 - dilation of X_2 based on (a), (f) X_4 - Final pruned image of (a)



Morphological Pruning, Fig. 2 (a) DEM of the study area in Cameron Highlands, Malaysia, (b) binary image of (a), (c) skeleton image of (b), and (d) pruned image of (c)

Figure 2a shows the DEM of the study region in Cameron Highlands with resolution of 257×258 pixels. Figure 2b is the resultant image after binarization is performed on Fig. 2a. Figure 2c shows the skeleton image of Fig. 2b. It is observed that there are some parasitic branches (spurs) that appear in the end points of the objects. Hence, after pruning is applied, Fig. 2d shows the final pruned image of Fig. 2a where most of the spurs have significantly been removed resulting in a cleaner skeleton image.

Morphological pruning has been applied in many applications related to geoscience. For example, in Tay et al. (2006), the traveltime channel networks were pruned downward from the tips of the branches to the stationary outlet. This mimics realistic phenomenon such as burning of fire from all the extremities of river source, and propagate progressively toward the outlet. The corresponding convex hulls of the traveltime pruned networks were also computed, which leads to the derivation of convexity measure. Based on the



Morphological Pruning, Fig. 3 (a) An example of (nonconvex set) channel network where the source point is represented as a round dot at the bottom of the tree, (b) convex hull of (a), (c) Traveltime pruned

network until it reaches the outlet, (d) union of corresponding convex hulls of pruned networks in (c). This figure is adopted from (Tay et al. 2006)

convexity measures obtained, new power law relations could be further derived to characterize the scaling structure of the networks. The proposed methodology has been explained with the use of a simple tree-like non convex set, shown in Fig. 3a. Its corresponding convex hull is as shown in Fig. 3b.

The traveltime network of Fig 3a are pruned recursively towards the outlet and color-coded in Fig. 3c, while Fig. 3d illustrates the convex hulls of the traveltime channel network. Furthermore, the framework has been applied on seven sub-basins of Cameron Highlands, Malaysia and the convexity

measures acquired from the dynamically shrinking traveltime basins which move toward their outlets offers geophysicists and geomorphologists a better visualization and understanding of the channelization process.

Summary or Conclusions

Morphological pruning is a useful technique for post-processing to eliminate unwanted parasitic components which appear following skeletonization and thinning operations. We have demonstrated the use of morphological pruning to eliminate unwanted parasitic components in the digital elevation model (DEM). From the results, it is obvious that morphological pruning is able to remove the unwanted spurs and helps to generate a cleaner output with significantly lesser noise. This framework can be applied to extract the ridge of the mountains or to determine the contour lines of drainage basins from the DEMs. Furthermore, it is proven that pruning can be adopted to applications related to geoscience which helps in deriving statistically significant metrics.

Cross-References

- [Digital Elevation Model](#)
- [Morphological Dilation](#)

Bibliography

- Baja GS (2006) Skeletonization of digital objects. *Progress in Pattern Recognition, Image Analysis and Applications*. pp 1–13
- Duan H, Wang J, Liu X, Liu H (2008) A scheme for morphological skeleton pruning. *Proceedings of 2008 IEEE International Symposium on IT in Medicine and Education*
- Gonzalez RC, Woods RE (2008) *Digital image processing*, 3rd edn. Prentice Hall, Upper Saddle River. ISBN 978-0131687288
- Kalimuthu H, Tan WN, Lim SL, Fauzi MFA (2016) Interpolation of low resolution Digital Elevation Models: A comparison. 2016 8th Computer Science and Electronic Engineering Engineering Conference (CEECE), 28–30 September 2016
- Lim SL, Tee SH, Lim ST (2016) Morphological interpolations of low-resolution DEMs vs High-resolution DEMs: A comparative study. *Progress in Computer Sciences and Information Technology International Conference, Procsit*, 20–22 December 2016
- Maragos P, Schafer R (1986) Morphological skeleton representation and coding of binary images. *IEEE Trans Acoust Speech Signal Process* 34(5):1228–1244
- Saha PK, Borgfors G, Baja GS (2017) *Skeletonization theory, methods, and applications*. Academic Press. ISBN 978-0-08-101291-8
- Shaked D, Bruckstein AM (1998) Pruning Medial Axes. *Computer Vision and Image Understanding* 69(2):156–169
- Tay LT, Sagar BSD, Chuah HT (2006) Allometric relationships between Traveltime Channel networks, convex hulls, and convexity measures. *Water Resour Res* 42(W06502)

Morphometry

Sebastiano Trevisani¹ and Igor V. Florinsky²

¹University IUAV of Venice, Venice, Italy

²Institute of Mathematical Problems of Biology, Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, Pushchino, Moscow Region, Russia

Synonyms

Geomorphometry; Quantitative morphology; Surface analysis; Terrain analysis

Definition

Morphometry is usually reckoned as the quantitative analysis of solid earth surface morphology. It is mainly focused on the analysis of topography and bathymetry; however, its related concepts and methods can be applied to the analysis of any surface, being it natural (e.g., planetary geomorphology) or anthropic (e.g., surface metrology), and at any spatial scale. The morphometric analysis is most efficiently conducted by means of geocomputational tools applied to digital representations (1D, 2D, etc.) of earth surface. The computational algorithms are implemented both in standalone specialized software for terrain analysis as well as additional tools in Geographical Information Systems (GIS). The morphometric analysis can be conducted also in the field by means of field equipment, e.g., for the characterization of local characteristics and for validation purposes. In a broader sense, morphometry also includes the quantitative spatial analysis of morphometric features, such as the drainage network and structural lineaments; in this context, the analysis can be conducted with geocomputational tools applied to remote sensing imagery.

Introduction

Topography is one of the main factors controlling processes taking place in the near-surface layer of the planet and it is expression of multiple geomorphic processes and factors. Accordingly, it is not surprising that the qualitative and quantitative study of surface morphology has a long record of research in the geosciences. For example, the work of Strahler on hypsometry (see ► [“Hypsometry”](#)) is emblematic of the use of quantitative methods for the description of surface morphology. Moreover, the heterogeneity and complexity of topographic surfaces and features have always attracted scientists; in this perspective, the development of fractal theory

(Mandelbrot 1967) is emblematic. Before the 1990s, topographic maps were the main source of quantitative information on topography. These maps were analyzed using geomorphometric techniques to calculate manually morphometric variables and produce morphometric maps. In the mid-1950s, a new research field – digital terrain modeling – emerged in photogrammetry. Within its framework, digital elevation models (DEMs), two-dimensional discrete functions of elevation, became the main source of information on topography. DEMs were used to calculate digital terrain models (DTMs), two-dimensional discrete functions of morphometric variables. Initially, digital terrain modeling was mainly applied to produce raised-relief maps using computer-controlled milling machines, and to design highways and railways. However, digital terrain modelling became widely applicable many years later, thanks to technological innovations in computing and topographic data collection.

Relevant developments in the physical and mathematical theory of the topographic surface in the gravity field accompanied the technological innovations. As a result, geomorphometry evolved into the science of quantitative modeling and analysis of the topographic surface and of the study of its relationships with the other natural and artificial components of geosystems. Currently, geomorphometry is widely adopted to solve various multiscale problems of geomorphology, hydrology, remote sensing, soil science, geology, geophysics, geobotany, glaciology, oceanology, climatology, planetology, and other disciplines (e.g., Moore et al. 1991; Pike 2000; Hengl and Reuter 2009; Florinsky 2016; Wilson 2018). Geomorphometric analysis has deep interconnections with many geocomputational, statistical, and mathematical approaches; these include geostatistics, fractal analysis, spectral methods (e.g., wavelets, Fourier analysis, etc.), mathematical morphology (Sagar 2013), and many others. Moreover, considering the raster representation of elevation, morphometry has strong methodological interconnections with image analysis and pattern recognition algorithms, with various examples of applications of advanced pattern recognition approaches such as the Local Binary Pattern or the classical gray level co-occurrence matrix (e.g., Haralick et al. 1973). In this context, the links between topographic surface roughness (or, as synonym, surface texture) and image texture analysis are evident (Haralick et al. 1973; Trevisani and Rocca 2015).

The current popularity of morphometry is strongly related to recent technological innovations in multiple fields, including techniques for the derivation of digital terrain models, computational power, and software development. Cloud storing services and parallel computing will play a key role in the next future, given the increasing need of computational resources for storing and analyzing digital elevation data. In

fact, the quantity of topographic-related data is continuously increasing for multiple reasons. First, new sensing technologies, e.g., LiDAR (Light Detection and Ranging) and SAR (Synthetic Aperture Radar), permit the efficient collection of topographic data for wide areas and at high resolution. Second, the increase in the geosphere-anthroposphere interactions and the need to deal with multiple geoenvironmental issues (e.g., natural hazards) require an accurate, detailed, and, often, multi-temporal knowledge of surface morphology. Third, the morphometric analysis can be conducted by means of a variety of algorithms, permitting to derive a wide set of morphometric variables and local statistical indices (e.g., Pike 2000; Florinsky 2017); moreover, the various morphometric indices can be computed at multiple scales.

Geomorphometric methods provide a multiscale quantitative description of solid-earth surface morphology by means of multiple derivatives and local statistical indices. These can be used directly for a visual analysis and a description of landscape, for comparing the morphometric characteristics of different study sites and for studying the connections between morphology and geomorphic processes and factors. Some morphometric indices can be also used as parameters in physical-based computational models; for example, local roughness indices can be used as impedance factor in surface-flow models. Often, the calculated geomorphometric features are further processed by means of supervised (e.g., landslide susceptibility models, soil properties mapping, etc.) or unsupervised learning approaches (e.g., landscape classification). In this context new, approaches based on machine learning are becoming increasingly popular in geomorphometric analysis.

Morphometry is inherently related to the field of remote sensing. First, frequently the digital elevation models are derived by means of remote sensing technologies, using active (e.g., LiDAR, SAR, etc.) or passive (e.g., imagery) sensors, mounted on a wide set of platforms (terrestrial, satellite, aerial, etc.). Second, many of the techniques adopted to analyze digital elevation models can be applied to the analysis of remote sensing imagery and vice versa. Third, relevant morphological features (e.g., structural lineaments) that can be analyzed quantitatively can be derived from remote sensing imagery. Fourth, in many applications (e.g., for studying soil properties, for mapping landslide susceptibility, or for ecological considerations), the quantitative analysis can be performed by means of an integrated use of morphometric derivatives and remote sensing-derived variables. Fifth, digital elevation models and local morphometric characteristics (e.g., slope, aspect, roughness, etc.) are important factors for the processing of remote sensing products (e.g., orthorectification of imagery, post-processing of SAR products, etc.).

Morphometrics

The digital representation of surface morphology is crucial for modern geomorphometry. The most popular and easy implementation is based on a 2.5 D representation of topography by means of a raster coding: i.e., the topography is represented by a single band image, in which the value of the pixel is the elevation. According to this representation, topographic surface is uniquely defined by a continuous, single-valued bivariate function

$$z = f(x, y), \quad (1)$$

where z is elevation, and x and y are the Cartesian coordinates. This means that caves, grottos, and similar landforms are excluded. However, alternative approaches for digital representation of topography can be adopted, including, for example, true 3D representations (i.e., caves can be represented), 3D digital globes, triangulated irregular networks, etc.

Even if there is not a consensus, many researchers differentiate morphometrics (i.e., indices describing some aspects of surface morphology) between morphometric variables and local statistical indices. Morphometric variables (Fig. 1) generally imply specific mathematical requirements for the representation of topographic surface (e.g., differentiability, see Florinsky 2017) and are defined through their relation with the gravity field (see section “[Local Morphometric Variables](#)” difference between flow and form attributes).

Local statistical indices do not require particular mathematical properties in the topographic surface data to be analyzed and are not necessarily related to the gravity field. For example, an important set of local statistical indices is adopted to describe surface roughness (or its synonym “surface texture”), a fundamental property of surface morphology. Surface roughness is a complex and multiscale characteristic that cannot be described by a single mathematical index (Fig. 2). In this regard, there is a quite ample literature on the use and development of surface texture indices (e.g., Haralick et al. 1973; Grohmann et al. 2011; Trevisani and Rocca 2015).

Given that morphometric variables represent the core of geomorphometry, we furnish a basic review of the fundamental ones. A morphometric (or topographic) variable (or attribute) is a single-valued bivariate function describing properties of the topographic surface. Here, we review fundamental topographic variables associated with the theory of the topographic surface and the concept of general geomorphometry, which is defined as “the measurement and analysis of those characteristics of landform which are applicable to any continuous rough surface. ... General geomorphometry as a whole provides a basis for the quantitative comparison ... of qualitatively different landscapes ...” (Evans 1972, p. 18). There are several classifications

of morphometric variables based on their intrinsic (mathematical) properties (Shary et al. 2002; Florinsky 2016). Here, the classification of Florinsky (2017) is adopted.

Morphometric variables can be divided into four main classes: (1) local variables; (2) nonlocal variables; (3) two-field specific variables; and (4) combined variables. The terms “local” and “nonlocal” are used regardless of the study scale or model resolution. They are associated with the mathematical sense of a variable (c.f. the definitions of a local and a nonlocal variable). Being a morphometric variable, elevation does not belong to any class listed, but all topographic attributes are derived from DEMs.

Local Morphometric Variables

A local morphometric variable is a single-valued bivariate function describing the geometry of the topographic surface in the vicinity of a given point of the surface, along directions determined by one of the two pairs of mutually perpendicular normal sections (Fig. 3).

A normal section is a curve formed by the intersection of a surface with a plane containing the normal to the surface at a given point. At each point of the topographic surface, an infinite number of normal sections can be constructed, but only two pairs of them are important for geomorphometry.

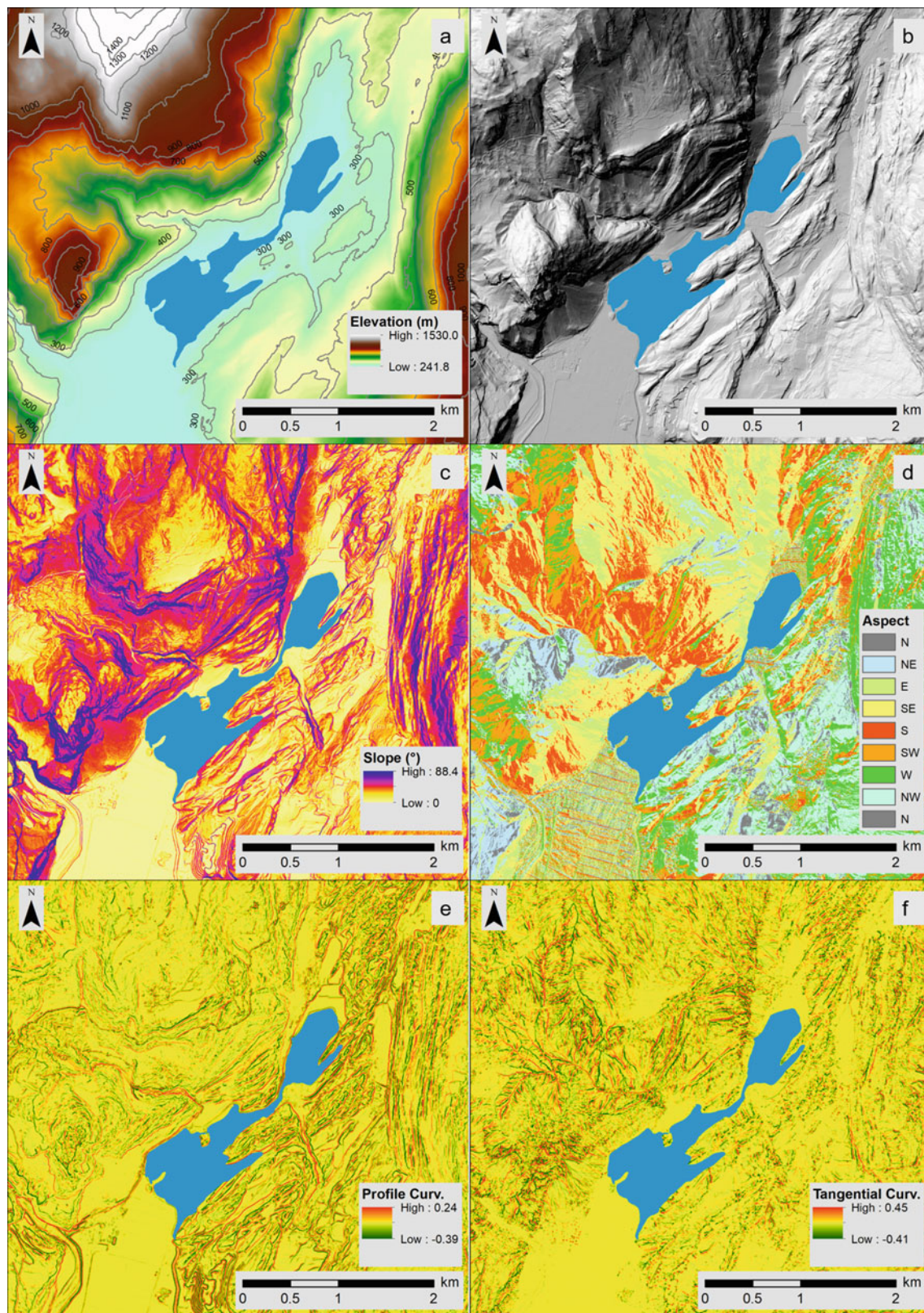
The first pair of mutually perpendicular normal sections includes two principal sections (Fig. 3a) well known from differential geometry. These are normal sections with extreme – maximal and minimal – bending at a given point of the surface.

The second pair of mutually perpendicular includes two normal sections (Fig. 3b) dictated by gravity. One of these two sections includes the gravitational acceleration vector and has a common tangent line with a slope line at a given point of the topographic surface. The other section is perpendicular to the first one and tangential to a contour line at a given point of the topographic surface.

Local variables are divided into two types – form and flow attributes – which are related to the two pairs of normal sections (Shary et al. 2002).

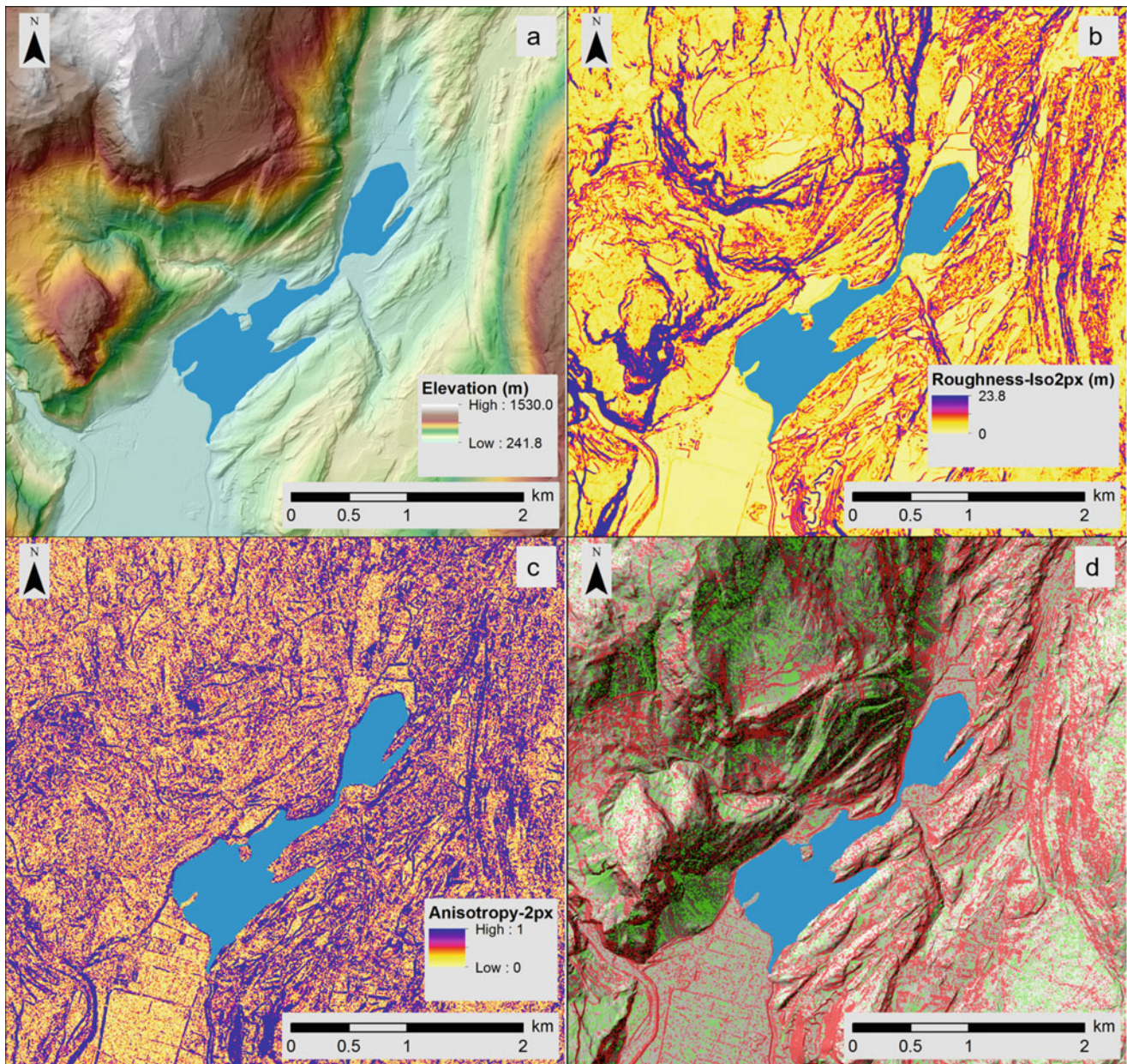
Form attributes are associated with two principal sections. These attributes are gravity field invariants. This means that they do not depend on the direction of the gravitational acceleration vector. Among these are minimal curvature ($k_{\min}$), maximal curvature ($k_{\max}$), mean curvature (H), the Gaussian curvature (K), unsphericity curvature (M), Laplacian (∇^2), shape index (IS), curvedness (C), and some others.

Flow attributes are associated with two sections dictated by gravity. These attributes are gravity field-specific variables. Among these are slope (G), aspect (A), northwardness (AN), eastwardness (AE), horizontal (or, tangential) curvature (k_h), vertical (or profile) curvature (k_v), difference curvature (E), horizontal excess curvature (k_{he}), vertical excess



Morphometry, Fig. 1 Example of basic morphometric variables for an alpine area (Trentino, Italy). (a) digital terrain model (pixel size 2 m); (b) shaded relief; (c) slope in degrees; (d) aspect classified in eight directions

(i.e., classes 45° degree wide); (e) profile curvature; (f) tangential curvature



Morphometry, Fig. 2 Examples of roughness indices (see Trevisani and Rocca 2015), based on a lag of 2 pixels and a circular moving window with a radius of 3 pixels, calculated for an alpine terrain (Trentino, Italy). (a) DTM and shaded relief (pixel size 2 m); (b)

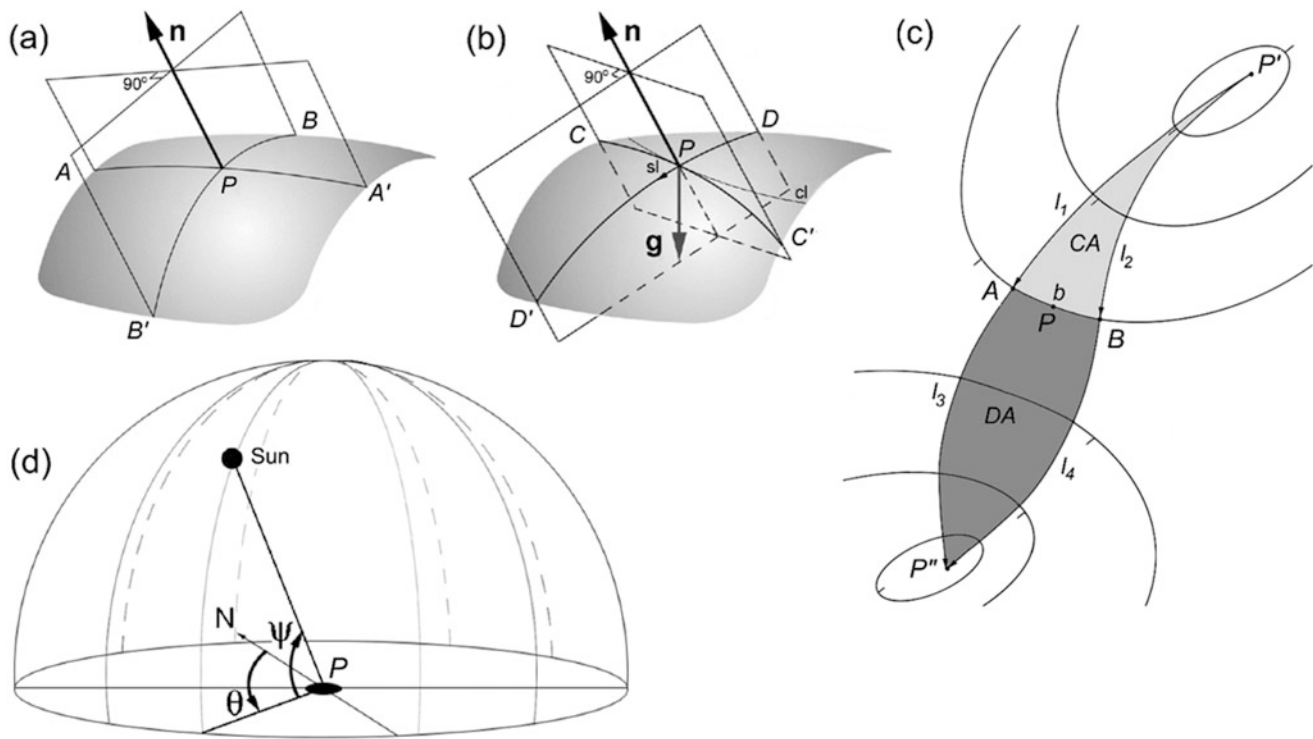
isotropic roughness; (c) anisotropy in roughness (1 maximum anisotropy); (d) landscape classified according the ratio of isotropic roughness versus flow directional roughness: in green morphologies elongated in the direction of flow; in red morphologies elongated along contour lines

curvature (k_{ve}), accumulation curvature (K_a), ring curvature (K_r), rotor (rot), horizontal curvature deflection (D_{kh}), vertical curvature deflection (D_{kv}), and some others.

Local topographic variables are functions of the partial derivatives of elevation. In this regard, local variables can be divided into three groups: (1) first-order variables – G , A , AN , and AE – are functions of only the first derivatives; (2) second-order variables – k_{min} , k_{max} , H , K , M , ∇^2 , IS , C , k_h , k_v , E , k_{he} , k_{ve} , K_a , K_r , and rot – are functions of both the first and second derivatives; and (3) third-order variables –

D_{kh} and D_{kv} – are functions of the first, second, and third derivatives. Equations can be found elsewhere (Florinsky 2016, 2017).

The partial derivatives of elevation (and thus local morphometric variables) can be estimated from DEMs by (1) several finite-difference methods using 3×3 or 5×5 moving windows and (2) analytical computations based on DEM interpolation by local splines or global approximation of a DEM by high-order orthogonal polynomials.



Morphometry, Fig. 3 Schemes for the definitions of local, nonlocal, and two-field specific variables. (a) and (b) display two pairs of mutually perpendicular normal sections at a point P of the topographic surface: (a) Principal sections APA' and BPB'. (b) Sections CPC' and DPD' allocated by gravity. n is the external normal; g is the gravitational acceleration vector; cl is the contour line; sl is the slope line. (c) Catchment and

dispersive areas, CA and DA, are areas of figures P'AB (light gray) and P''AB (dark gray), correspondingly; b is the length of a contour line segment AB; l_1, l_2, l_3 , and l_4 are the lengths of slope lines P'A, P'B, AP'', and BP'', correspondingly. (d) The position of the Sun in the sky: θ is solar azimuth angle, ψ is solar elevation angle, and N is north direction. (From Florinsky (2017, Fig. 1))

Nonlocal Morphometric Variables

A nonlocal (or regional) morphometric variable is a single-valued bivariate function describing a relative position of a given point on the topographic surface.

Among nonlocal topographic variables are catchment area (CA) and dispersive area (DA). To determine nonlocal morphometric attributes, one should analyze a relatively large territory with boundaries located far away from a given point (e.g., an entire upslope portion of a watershed) (Fig. 3c).

Flow routing algorithms are usually applied to estimate nonlocal variables. These algorithms determine a route, along which a flow is distributed from a given point of the topographic surface to downslope points. There are several flow routing algorithms grouped into two types: (1) eight-node single-flow direction (D8) algorithms using one of the eight possible directions separated by 45° to model a flow from a given point and (2) multiple-flow direction (MFD) algorithms using the flow partitioning. There are some methods combining D8 and MFD principles.

Two-Field Specific Morphometric Variables

A two-field specific morphometric variable is a single-valued bivariate function describing relations between the topographic surface (located in gravity field) and other fields, in particular, solar irradiation and wind flow.

Among two-field specific morphometric variables are reflectance (R) and insolation (I). These variables are functions of the first partial derivatives of elevation and angles describing the position of the Sun in the sky (Fig. 3d). Reflectance and insolation can be derived from DEMs using methods for the calculation of local variables.

Combined Morphometric Variables

Morphometric variables can be composed from local and nonlocal variables. Such attributes consider both the local geometry of the topographic surface and a relative position of a point on the surface.

Among combined morphometric variables are topographic index (TI), stream power index (SI), and some others.

Combined variables are derived from DEMs by the sequential application of methods for nonlocal and local variables, followed by a combination of the results.

Conclusions

Geomorphometry is a discipline still in evolution and of growing interest in many applied and research contexts. Geomorphometry is a fundamental tool for studying geosphere-anthroposphere interlinked dynamics in the critical zone, especially when performing multitemporal morphometric analysis. Moreover, geomorphometric analysis permits to derive relevant geoenvironmental and hydrological parameters, useful for the numerical modelling of different processes such as floods, landslides, wildfires and ecological processes. Geomorphometric indices can be also exploited as secondary variables in various prediction approaches based on geostatistics or machine learning, for example for mapping soil properties. Geomorphometry is also adopted for the characterization of landscape and of geodiversity. Future developments in mathematical/computational approaches as well as in technology will likely amplify morphometry potentials.

Cross-References

- [Digital Elevation Model](#)
- [Geostatistics](#)
- [Horton, Robert Elmer](#)
- [Hypsometry](#)
- [LiDAR](#)
- [Machine Learning](#)
- [Mathematical Morphology](#)
- [Pattern Analysis](#)
- [Quantitative Geomorphology](#)
- [Remote Sensing](#)
- [Virtual Globe](#)

Bibliography

- Evans IS (1972) General geomorphometry, derivations of altitude, and descriptive statistics. In: Chorley RJ (ed) *Spatial analysis in geomorphology*. Methuen, London, pp 17–90
- Florinsky IV (2016) *Digital terrain analysis in soil science and geology*, 2nd edn. Academic, Amsterdam
- Florinsky IV (2017) An illustrated introduction to general geomorphometry. *Prog Phys Geogr* 41(6):723–752
- Grohmann CH, Smith MJ, Riccomini C (2011) Multiscale analysis of topographic surface roughness in the Midland Valley, Scotland. *IEEE Trans Geosci Remote Sens* 49(4):1200–1213
- Haralick RM, Shanmugam K, Dinstein I (1973) Textural features for image classification. *IEEE Trans Syst Man Cybern* SMC-3(6):610–621
- Hengl T, Reuter HI (2009) *Geomorphometry: concepts, software, applications*. Elsevier, Amsterdam
- Mandelbrot B (1967) How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science* 156(3775):636–638
- Moore ID, Grayson RB, Ladson AR (1991) Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. *Hydrol Process* 5(1):3–30
- Pike RJ (2000) Geomorphometry - diversity in quantitative surface analysis. *Prog Phys Geogr* 24(1):1–20
- Sagar BSD (2013) *Mathematical morphology in geomorphology and GISci*. Chapman and Hall/CRC
- Shary PA, Sharaya LS, Mitusov AV (2002) Fundamental quantitative methods of land surface analysis. *Geoderma* 107(1–2):1–32
- Trevisani S, Rocca M (2015) MAD: robust image texture analysis for applications in high resolution geomorphometry. *Comput Geosci* 81:78–92
- Wilson JP (2018) *Environmental applications of digital terrain modeling*. Wiley-Blackwell, Chichester

Moving Average

Frits Agterberg
Central Canada Division, Geomathematics Section,
Geological Survey of Canada, Ottawa, ON, Canada

Definition

A moving average is the mean or average of all values of a variable within a given domain that is moved across the space of study. Domain and space can be one-, two-, or three-dimensional. In Euclidian space, every moving average value is located at the center of its domain. For time series, the moving average can apply to values before the point in time at which it is located. In weighted moving average analysis, the values are assigned different weights usually decreasing with distance away from the point of location of each estimated moving average value.

Introduction

Along with “probability,” the mean or average is probably the best-known statistical tool. The first average on record was taken by William Borough in 1581 for a number of successive compass readings (*cf.* Eisenhart 1963; also see Hacking 2006). The procedure of averaging numbers was regarded with suspicion for a long time. Thomas Simpson (1755) advocated the approach in a paper entitled: “On the advantage of taking the mean of a number of observations in practical astronomy,” stating: “It is well-known that the method practiced by astronomers to diminish the errors arising from the imperfections of instrument and the organs of sense by taking the mean of several observations has not so generally been received but that some persons of note have publicly

maintained that one single observation, taken with due care, was as much to be relied on, as the mean of a great number.”

The simple, unweighted moving average of n values x_i within a given domain can be written as $\bar{x}_n = \frac{\sum_{i=1}^n x_i}{n}$. Different domains can have different values of n . An example is shown in Fig. 1. It is for a long series of 2796 copper (XRF) concentration values for cutting samples taken at 2 m intervals along the so-called Main KTB borehole of the German Continental Deep Drilling Program (abbreviated to KTB). These data are in the public domain (citation: KTB, WG Geochemistry). Depths of the first and last cuttings used for this series are 8 and 5596 m, respectively. Locally, in the database, results are reported for a 1-m sampling interval; then, alternate copper values at the standard 2 m interval were included in the series used, for example. Most values are shown in Fig. 1 together with a 101-point average representing consecutive 202-m log segments of drill-core. Locally, the original copper values deviate strongly from the moving average. It can be assumed that these deviations are largely random and that the moving average represents a reproducible copper concentration value contrary to the original XRF values that are subject to significant measurement errors.

Weighted Moving Average Methods

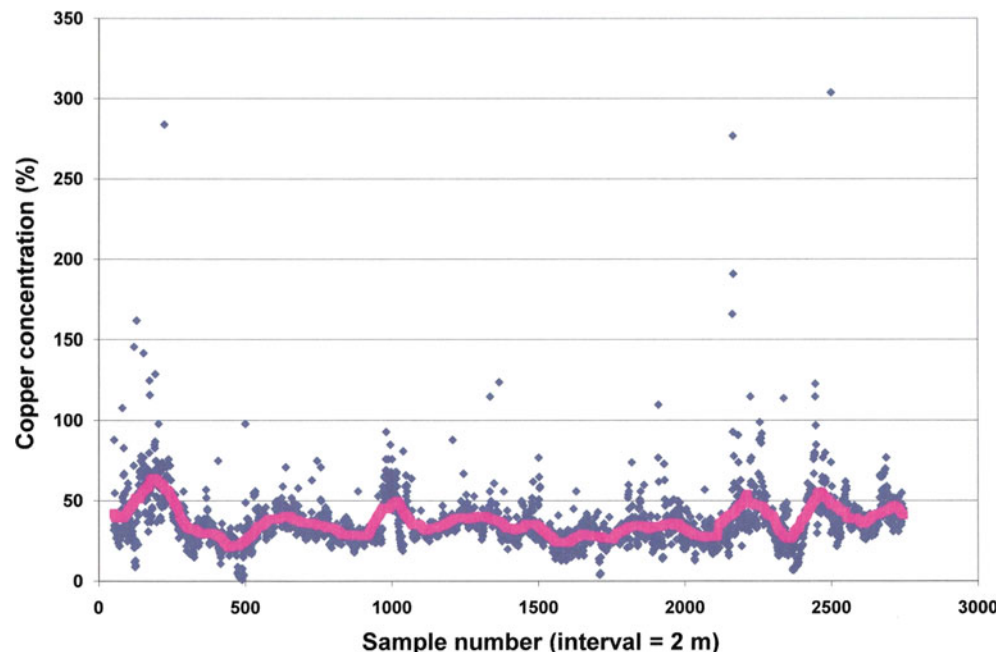
In most applications of the moving average method, different weights are assigned to the observed values depending on the distance between the point at which the moving average value

is to be calculated and the point at which a measurement is to be used. In time series analysis, more weight is commonly assigned to values at points in the more immediate past. A frequently used method in time series analysis is the exponentially weighted moving average (EWMA) method summarized by Hunter (1986) as follows: The predicted value at time $i + 1$ in a time series with observed values y_i can be written as $\hat{y}_{i+1} = \hat{y}_i + \lambda \epsilon_i$ where $0 < \lambda < 1$ is a constant and $\epsilon_i = y_i - \hat{y}_i$ is the observed random error at time i .

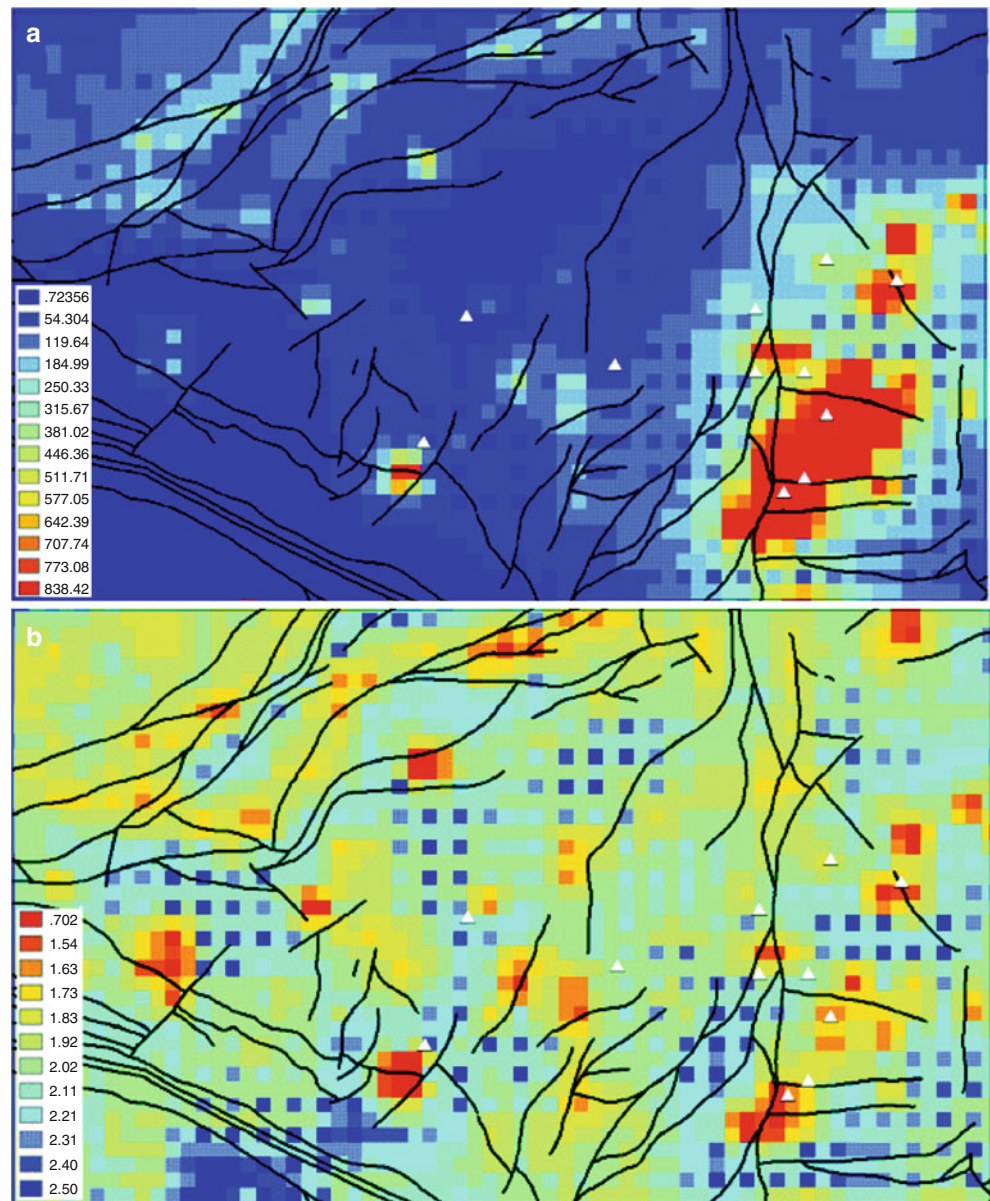
For applications in Euclidian space, more weight is assigned to values at points that are closer to the points at which the values are to be estimated. A well-known method in two-dimensional space consists of assigning weights to the observations that are inversely proportional to the squared distances between locations of the known values and location at which the value is to be estimated. An example from Cheng and Agterberg (2009) based on arsenic concentration values in about 1000 stream sediment samples in the Gejiu area in China is shown in Fig. 2a. This area contains large tin deposits in its southeastern part that have a long history of mining causing extensive pollution in the environment. The stream sediment sample locations were equally spaced at approximately 2 km in the north-south and east-west directions. Every sample represented a composite of materials from the drainage basin upstream of the collection site. In this application, each square on Fig. 2a represents the moving average for a square window measuring 26 km on a side with influence of samples decreasing according to the square of distance. For comparison, Fig. 2b also shows results of local singularity analysis applied to the same data set. This is a

Moving Average,

Fig. 1 Copper concentration (ppm) values from Main KTB bore-hole together with mean values for 101 m long segments of drill-core. Locally, the original data deviate strongly from the moving average. (Source: Agterberg 2014, Fig. 6.22)



Moving Average, Fig. 2 Map patterns derived from arsenic concentration values in stream sediment samples; triangles represent large tin mines rich in other elements including arsenic as well: (a) weighted moving average with weights inversely proportional to squared distance from point at which value is plotted; (b) local singularity based on small square cells with half-widths set equal to 1, 3, 5, ..., 13 km; for further explanation. (See Cheng and Agterberg 2009)

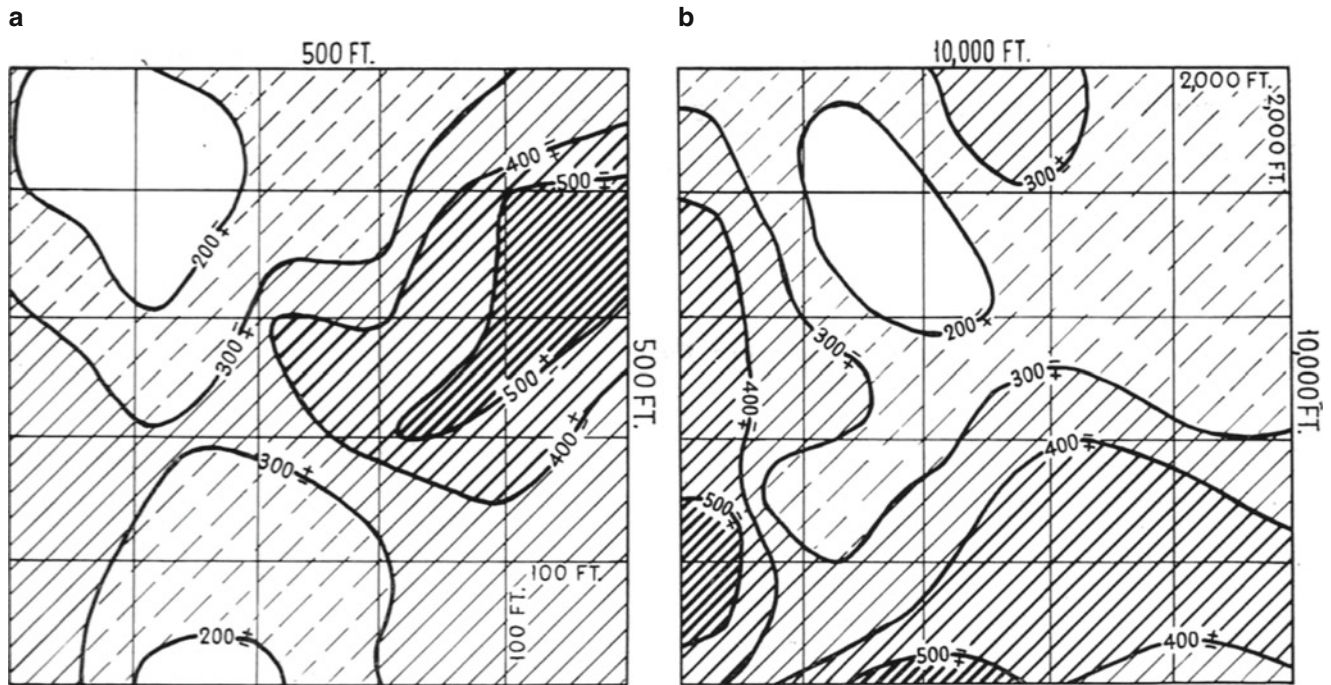


different contouring technique that places more emphasis on the strictly local environments of the points at which the average is to be computed (for details, see Cheng and Agterberg 2009). Obviously, the pattern of Fig. 2b is more advantageous when the objective is to find undiscovered mineral deposits in relatively unexplored target areas, because the areas used for the moving averaging are much larger in Fig. 2a.

One of the first applications of a weighted moving average technique in the geosciences is described in detail by Krige (1968) for gold assay values in South African gold mines. In this application, the weights were determined by the method of sampling that was applied to the gold deposits with more weight assigned to nearby sampling locations. Figure 3 (after

Krige 1966) shows two kinds of results for one particular gold mine although the same moving average method was applied. The area covered in Fig. 3b is 400 times as large as the area in Fig. 3a indicating scale independence (self-similarity) of the gold distribution pattern.

Other weighted moving average methods include hot spot analysis (see Getis and Ord 1992). It is widely applied by geographers to enhance two-dimensional patterns of random variables that exhibit spatial clustering. Typical examples are counts (x_{ij}) of occurrences (e.g., cases of a specific disease or accidents) for small areas (e.g., counties). Originally, the technique was based on Moran's I statistic for spatial correlation (Moran, 1968). It led to the Getis $G_i^*(d)$ statistic that satisfies:



Moving Average, Fig. 3 Typical gold inch-dwt weighted moving average trend surfaces in the Klerksdorp goldfield obtained on the basis of two-dimensional moving averages for two areas with similar average gold grades; value plotted is inch-pennyweight (1 unit = 3.95 cm-

g). (a) moving averages of 100×100 ft. areas within a mined-out section of 500×500 ft; (b) moving averages of 2000×2000 ft. areas within mined-out section of $10,000 \times 10,000$ ft. (Source: Krige 1966, Fig. 3)

$$G_i(d) = \frac{\sum_{j=1}^n w_{ij}(d)x_j}{\sum_{j=1}^n x_j}, j \text{ not equal to } i,$$

where $\{w_{ij}\}$ is an asymmetric one/zero spatial weight matrix with ones for all links defined as being within distance d of a given i , all other links being zero (cf. Getis and Ord 1992, p. 190). Setting $W_i = \sum_{j=1}^n w_{ij}(d)$, it can be shown that the expected value of $G_i(d)$ and its variance satisfy:

$$E[G_i(d)] = \frac{W_i}{n-1}; \sigma^2[G_i(d)] = \frac{W_i(n-1-W_i)Y_{i2}}{(n-1)^2(n-2)Y_{i1}^2}$$

where E denotes mathematical expectation and σ^2 is variance,

$$Y_{i1} = \frac{\sum_{j=1}^n x_j}{n-1}; \text{ and } Y_{i2} = \frac{\sum_{j=1}^n x_j^2}{n-1} - Y_{i1}^2.$$

For a one-dimensional application of hot-spot analysis, see Agterberg et al. (2020).

Summary and Conclusions

A moving average is the mean or average of all values of a variable within a given domain that is moved across its

Euclidian space of study. For time series, the moving average can apply to values before the point in time at which it is located. In weighted moving average analysis, the input values are assigned different weights, usually decreasing with distance away from the point of location of each estimated moving average value. By means of practical examples, it was shown in this entry that using the moving average can help to eliminate “noise” from the data. Also, applying weights to the observed values that decrease with distance away from each point, at which the moving average is estimated, often improves results. Further improvements may be obtainable when estimated average values are restricted to strictly local environments. The three examples (Figs. 1, 2 and 3) are concerned with element concentration values across study areas that exhibit scale-independence, which is documented in more detail in the original publications.

Cross-References

- [Local Singularity Analysis](#)
- [Scaling and Scale Invariance](#)
- [Time Series Analysis](#)
- [Trend Surface Analysis](#)

Bibliography

- Agterberg FP (2014) Geomathematics: theoretical foundations, applications and future developments, Quantitative geology and geostatistics 18. Springer, Heidelberg, 553 pp
- Agterberg FP, Da Silva A-C, Gradstein FM (2020) Geomathematical and statistical procedures. In: Gradstein FM, Ogg JG, Schmitz MB, Ogg GB (eds) Geologic time scale 2020, vol 1. Elsevier, Amsterdam, pp 402–425
- Cheng Q, Agterberg FP (2009) Singularity analysis of ore-mineral and toxic trace elements in stream sediments. *Comput Geosci* 35: 234–244
- Eisenhart MA (1963) The background and evolution of the method of least squares. In: Proceedings of the 34th Sess. Int. Statist Inst, Ottawa (preprint)
- Getis A, Ord JK (1992) The analysis of spatial association by use of distance statistics. *Geogr Anal* 24:189–206
- Hacking I (2006) The emergence of probability, 2nd edn. Cambridge University Press, 209 pp
- Hunter JS (1986) The exponentially weighted moving average. *J Qual Technol* 18:203–210
- Krige DG (1966) A study of gold and uranium distribution pattern in the Klerksdorp goldfield. *Geoexploration* 4:43–53
- Krige DG (1968) Two-dimensional weighted moving average trend surfaces in ore evaluation. In: Proceedings of symposium on mathematical statistics and computer applications. Southern African Institute of Mining and Metallurgy, Johannesburg, pp 13–38
- Moran PAP (1968) An introduction to probability theory. Oxford University Press, 542 pp

Multidimensional Scaling

Francky Fouedjio

Department of Geological Sciences, Stanford University,
Stanford, CA, USA

Definition

Multidimensional scaling (MDS) is a family of methods used for representing (dis)similarity measurements among objects (entities) as distances between points in a low-dimensional space, each point corresponding to one object (entity). MDS is concerned with the problem of constructing a configuration of n points in a m -dimensional space using information about the (dis)similarities between n objects, in such a way that inter-point distances reflect, in some sense, the (dis)similarities. Thus, similar objects are mapped close to each other, whereas dissimilar objects are located far apart.

Introduction

Multidimensional scaling (MDS) is frequently used in geosciences for exploring (dis)similarities among a set of

objects (entities) such as signals, images, simulations, predictions, or samples. The term “similarity” depicts the degree of likeness between two objects, while the term “dissimilarity” indicates the degree of unlikeness. Some objects are more dissimilar (or similar) to each other than others. For instance, colors red and pink are more similar (less dissimilar) to each other than colors red and green. Equivalently, colors red and green are more dissimilar (less similar) than colors red and pink. For dissimilarity data, a higher value indicates more dissimilar objects, while for similarity data, a higher value indicates more similar objects. Dissimilarities are distance-like quantities, while similarities are inversely related to distances. Dissimilarity or similarity information about a set of objects can arise in many different ways. For instance, if one considers an ensemble of continuous geostatistical realizations, the dissimilarity between any two realizations can be defined as the square root sum of square differences between realizations, which is calculated cell-by-cell.

Multidimensional scaling (MDS) aims to map objects’ relative location using data that show how these objects are dissimilar or similar. The basic idea consists of representing the objects in a low-dimensional space (usually Euclidean) so that the inter-point distances best agree with the observed (dis)similarities between the objects. In other words, MDS depicts interobject (dis)similarities by inter-point distances, each point corresponding to one object. In this way, MDS provides interpretable graphical displays (i.e., maps of objects), in which dissimilar objects are located far apart, and similar objects are mapped close to each other. In the color example, the points representing red and pink will be located closer in the space than the points representing red and green. MDS’s map facilitates our intuitive understanding of the relationships among the objects represented in the map. It helps to understand and possibly explain any structure or pattern in the data.

Multidimensional scaling’s map (or spatial representation) of a set of n objects consists of a set of n m -dimensional coordinates, each one of which represents one of the n objects. The required coordinates are generally found by minimizing some measure of goodness-of-fit between the inter-point distances implied by the coordinates and the observed (dis)similarities. MDS seeks to find the best-fitting set of coordinates and the appropriate value of the number of dimensions m needed to adequately represent the observed (dis)similarities. The hope is that the number of dimensions, m , will be small, ideally two or three, so that the derived configuration of points can be easily plotted. It is important to note that the configuration of points produced by any MDS method is indeterminate with respect to translation, rotation, and reflection. There is no unique set of coordinate values that give rise to a set of distances since the distances are unchanged by shifting the whole configuration of points

from one place to another or by a rotation or a reflection of the configuration. In other words, one cannot uniquely determine either the location or the orientation of the configuration of points.

There are mainly two types of MDS, depending on the scale of the (dis)similarity data: metric multidimensional scaling and non-metric multidimensional scaling. The metric multidimensional scaling is adequate when dissimilarities are also distances. In contrast, non-metric multidimensional scaling is appropriate when dissimilarities are not distances. The metric MDS uses the actual values of the dissimilarities, while the non-metric MDS effectively uses only the rank order (orderings) of the dissimilarities. These two types of MDS also differ in how the agreement between fitted inter-point distances and observed dissimilarities is assessed. A complete description of the theoretical and practical aspects of MDS can be found in the following reference books: Cox and Cox (2000), Borg and Groenen (2005), and Borg et al. (2018).

(Dis)Similarity Measure

Multidimensional scaling (MDS) is based on similarity or dissimilarity information between pairs of objects. The similarity between two objects is a numerical measure of the degree to which the two objects are alike. Thus, similarities are higher for pairs of objects that are more alike. In contrast, the dissimilarity between two objects is the numerical measure of the degree to which the two objects are different. Dissimilarity is lower for more similar pairs of objects. The term proximity is often used as an umbrella that encompasses both dissimilarity and similarity.

Let O be the set of objects. A similarity measure s on O is defined as $s: O \times O \rightarrow \mathbb{R}$ such that: (1) $s(\mathbf{x}, \mathbf{y}) \leq s_0$; (2) $s(\mathbf{x}, \mathbf{x}) = s_0$; and (3) $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$; if in addition, (4) $s(\mathbf{x}, \mathbf{y})s(\mathbf{y}, \mathbf{z}) \leq [s(\mathbf{x}, \mathbf{y}) + s(\mathbf{y}, \mathbf{z})]s(\mathbf{x}, \mathbf{z})$ (triangle inequality), s is called a metric similarity measure.

A dissimilarity measure is a function $d: O \times O \rightarrow \mathbb{R}$ such that: (1) $d(\mathbf{x}, \mathbf{y}) \geq 0$; (2) $d(\mathbf{x}, \mathbf{x}) = 0$; and (3) $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$; if in addition, (4) $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ (triangle inequality), d is called a metric dissimilarity measure (distance measure).

A (dis)similarity matrix is a matrix such that their elements (representing (dis)similarities between pairs of objects) fulfill the conditions depicted above. The input data for MDS is in the form of a dissimilarity matrix representing the dissimilarities between pairs of objects. When the similarities are available, MDS requires first to convert the similarities into dissimilarities. There are many ways of transforming similarity measures into dissimilarity measures. One way is to take $d(\mathbf{x}, \mathbf{y}) = s_0 - s(\mathbf{x}, \mathbf{y})$, $\forall (\mathbf{x}, \mathbf{y}) \in O \times O$.

Metric MDS

Assume that there are n objects with dissimilarities δ_{ij} between them, for $i, j = 1, \dots, n$ (one dissimilarity for each pair of objects). Let $\mathbf{\Delta} = [\delta_{ij}]$ be the $n \times n$ dissimilarity matrix. The metric MDS attempts to find the coordinates of the n objects in a m -dimensional Euclidean space, in the form of an $n \times m$ matrix of coordinates values $\mathbf{X} = [\mathbf{x}_{kl}]$, where each point represents one of the objects, and the inter-point distances $d_{ij}(\mathbf{X})$ are such that:

$$d_{ij}(\mathbf{X}) \approx f(\delta_{ij}), \quad (1)$$

where $f(\cdot)$ is a continuous parametric monotonic function, such that $f(\delta_{ij}) = \alpha + \beta\delta_{ij}$. The function $f(\cdot)$ can either be the identity function or a function that attempts to transform the dissimilarities to a distance-like form. The two main metric MDS methods are classical scaling and least squares scaling.

Classical Scaling

Classical scaling transforms a dissimilarity matrix into a set of coordinates such that the (Euclidean) distances derived from these coordinates approximate as closely as possible the original dissimilarities. The basic idea of classical scaling is to transform the dissimilarity matrix into a cross-product matrix and then perform its spectral decomposition, which gives a principal component analysis (PCA). Given the dissimilarity matrix $\mathbf{\Delta} = [\delta_{ij}]$, the classical MDS carried out the following steps:

1. Construct the matrix $\mathbf{A} = \left[-\frac{1}{2}\delta_{ij}^2\right]$.
2. Form the matrix $\mathbf{B} = \mathbf{H}\mathbf{A}\mathbf{H}$, where $\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^T$ is the centering matrix of size n and $\mathbf{1}$ is a column-vector of n ones.
3. Perform the eigen-decomposition of $\mathbf{B} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T$, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_n)$ is the diagonal matrix formed from the eigenvalues of $\mathbf{B}$, and $\mathbf{\Gamma} = [\mathbf{\Gamma}_1, \dots, \mathbf{\Gamma}_n]$ is the corresponding matrix of eigenvectors; the eigenvalues are assumed to be labeled such that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$.
4. Take the first m eigenvalues greater than 0 ($\mathbf{\Lambda}_+$) and the corresponding first m columns of $\mathbf{\Gamma}$ ($\mathbf{\Gamma}_+$). The solution of classical MDS is $\mathbf{X} = \mathbf{\Gamma}_+\mathbf{\Lambda}_+^{1/2}$.

If the dissimilarity matrix $\mathbf{\Delta}$ is Euclidean, i.e., the matrix $\mathbf{B}$ is positive semi-definite, an exact representation of dissimilarities, that is, one for which $d_{ij} = \delta_{ij}$, can be found in a q -dimensional Euclidean space. The coordinate matrix of classical MDS is then given by $\mathbf{X} = \mathbf{\Gamma}_q\mathbf{\Lambda}_q^{1/2}$, where $\mathbf{\Lambda}_q$ is

the matrix of the q non-zero eigenvalues and Γ_q is the matrix of the corresponding q eigenvectors. Using all q -dimensions will lead to complete recovery of the original Euclidean dissimilarity matrix. The best-fitting m -dimensional representation is given by the m eigenvectors of $\mathbf{B}$ corresponding to the m largest eigenvalues. The adequacy of the m -dimensional representation can be judged based on one of the following two criteria:

$$P_m^{(1)} = \frac{\sum_{i=1}^m |\lambda_i|}{\sum_{i=1}^n |\lambda_i|}, \quad P_m^{(2)} = \frac{\sum_{i=1}^m \lambda_i^2}{\sum_{i=1}^n \lambda_i^2}. \quad (2)$$

It is important to highlight that when the dissimilarity matrix contains Euclidean distances calculated from an $n \times q$ data matrix $\mathbf{X}$, classical scaling can be shown to be equivalent to principal component analysis (PCA), with the required coordinate values corresponding to the scores on the principal component extracted from the covariance matrix of the data. One result of this duality is that classical multidimensional scaling is also referred to as principal coordinates analysis (PCoA). Classical scaling can also be formulated in terms of the optimization of a particular goodness-of-fit function. The classical scaling solution is optimal in the least squares sense. The m -dimensional principal components solution ($m < q$) is “best” in the sense that it minimizes the following goodness-of-fit:

$$\text{Strain}(\mathbf{X}) = \left(\sum_{i,j} (d_{ij}(\mathbf{X}) - \delta_{ij})^2 / \sum_{i,j} d_{ij}^2(\mathbf{X}) \right)^{1/2} \quad (3)$$

where δ_{ij} is the Euclidean distance between objects i and j based on their original q variable values and $d_{ij}(\mathbf{X})$ is the corresponding distance calculated from the m principal component scores.

If the dissimilarity matrix Δ is not Euclidean, the matrix $\mathbf{B}$ is not positive semi-definite. In this circumstance, some of the eigenvalues of $\mathbf{B}$ are negative; correspondingly, some coordinate values are complex. If, however, $\mathbf{B}$ has only a small number of small negative eigenvalues, a useful representation of the dissimilarity matrix may still be possible using the eigenvectors associated with the m largest positive eigenvalues. If, however, the matrix $\mathbf{B}$ has a considerable number of large negative eigenvalues, classical scaling may be inadvisable, and some other methods of scaling, for example, non-metric scaling, might be used instead.

Least Squares Scaling

Metric least squares scaling finds a configuration of points matching the inter-point distances $d_{ij}(\mathbf{X})$ to observed

dissimilarities δ_{ij} by minimizing a loss function, with possibly a continuous monotonic transformation of the dissimilarities, $f(\delta_{ij})$. To assess the match between inter-point distances and dissimilarities, one needs to define an objective function which takes the value zero if the inter-point distances and dissimilarities are a perfect match and increases in value as the fit becomes less good. An obvious candidate for such a function is a least squares type fit criterion, for instance:

$$\text{Stress}(\mathbf{X}) = \left(\sum_{i < j} (d_{ij}(\mathbf{X}) - f(\delta_{ij}))^2 / \sum_{i < j} d_{ij}^2(\mathbf{X}) \right)^{1/2} \quad (4)$$

The denominator in Eq. (4) is a normalizing term making the goodness-of-fit criterion scale-free. This normalization has the advantage that it makes the goodness-of-fit criterion values do not depend on the size of the configuration of points. A configuration of points that minimizes the goodness-of-fit criterion might be found by an appropriate optimization technique such as the steepest descent approach or iterative majorization method. Unlike classical scaling, least squares scaling is an optimization process minimizing a loss function through iterative algorithms. As one can note, the classical scaling algorithm is algebraic (closed-form solution) and not iterative.

Non-metric MDS

It is often not the actual values of dissimilarities that are important or meaningful but their values in relation to the distances between other pairs of objects. In many situations, the dissimilarities may contain little reliable information beyond that implied by their rank order. In other words, the dissimilarities do not have strict numerical significance, only an ordinal significance. For example, in comparing a range of colors, one might be able to specify that one was brighter than another without being able to attach any value to the exact numerical significance between them. In such circumstances, the dissimilarities can be interpreted only in an ordinal sense, and metric MDS will no longer be appropriate.

Non-metric MDS assumes that only the ranks or orderings of the dissimilarities are known. Hence, this method produces a map that tries to reproduce these ranks and not the observed or actual dissimilarities. Thus, only the ordering of the dissimilarities is relevant to the method. Consequently, the technique is invariant under monotonic transformations of the dissimilarities.

Given the $(n \times n)$ dissimilarity matrix Δ , non-metric MDS seeks to find a configuration of points $\mathbf{X} = [\mathbf{x}_{ki}] \in \mathbb{R}^q$ in a q -dimensional space, such that the following relation is satisfied as much as possible:

$$d_{ij}(\mathbf{X}) \approx h(\delta_{ij}), \quad (5)$$

where $h(\cdot)$ is a monotonic function that preserves the rank order of the dissimilarities: $\delta_{ij} < \delta_{kl} \Leftrightarrow h(\delta_{ij}) \leq h(\delta_{kl})$. Unlike metric MDS, here $h(\cdot)$ is much general and is only implicitly defined.

Thus, one seeks a configuration of points in a given dimension space so that the rank order of the inter-point distances best agrees with the dissimilarities' rank order. The configuration of points is determined such that the inter-point distances minimize a loss function defined by:

$$\text{Stress}(\mathbf{X}) = \min_h \left[\frac{\sum_{i < j} [d_{ij}(\mathbf{X}) - h(\delta_{ij})]^2}{\sum_{i < j} d_{ij}^2(\mathbf{X})} \right]^{\frac{1}{2}}, \quad (6)$$

where the minimization is performed over all monotonic functions $h(\cdot)$. Here $h(\delta_{ij})$ represents the monotone least squares regression of the $\{d_{ij}(\mathbf{X})\}$ on the $\{\delta_{ij}\}$. $h(\delta_{ij})$ is called disparities. Note that the original dissimilarities are only used in checking the monotonic condition. In fact, only the order $\delta_{ij} < \delta_{kl} < \dots < \delta_{mf}$ among dissimilarities is needed.

The loss function defined in Eq. (6) provides a measure of the degree to which the rank order of the configuration inter-point distances concurs with the rank order of the dissimilarities to be mapped. It is invariant by translation, rotation, or rescaling (uniform stretching or shrinking) of the configuration. So the non-metric MDS solution is only known up to one of these transformations. The loss function is minimized with respect to $d_{ij}(\mathbf{X})$, i.e., with respect to $\mathbf{X} = [\mathbf{x}_{kl}]$, the coordinates of the configuration, and also with respect to $h(\cdot)$ using isotonic regression.

The minimization of the loss function is performed using the Shepard-Kruskal iterative algorithm as described, for example, in Borg and Groenen (2005). It is a two-phase optimization algorithm. In each phase, one set of parameters (distances or disparities, respectively) is taken as fixed values, while the other set of arguments is modified in such a way that *stress* is reduced. One starts from an initial configuration $\mathbf{X}^{(0)}$. Then, one seeks $h(\delta_{ij})$ such that $\sum_{i < j} [d_{ij}(\mathbf{X}^{(0)}) - h(\delta_{ij})]^2$ is minimum. This problem admits a unique solution: isotonic regression. The value of the *stress* is thus deduced. The configuration of points is then modified by means of small displacement of points using the steepest descent method to reduce the *stress*. Then, one goes back to the isotonic regression phase and so on until convergence is reached. The choice of the starting configuration is essential to finding the global rather than a local minimum. In practice, the solution of the metric MDS is taken as the starting configuration. There are several other methods for choosing a starting configuration – for example, generating n points according to a Poisson process in a region of $\mathbb{R}^q$.

Goodness-of-Fit and Number of Dimensions

There are several ways of assessing the goodness of an MDS solution. The quality of an MDS solution can be assessed informally by plotting the original dissimilarities versus the corresponding inter-point distances in MDS space. This plot is known as a Shepard diagram. This latter shows how well inter-point distances within the configuration of points agree with the original dissimilarities. In an ideal situation, the points fall on the bisecting line. The Shepard diagram is highly informative. It shows the number of points that have to be fitted, the optimal regression line, the size of the deviations, possible outliers, and systematic deviations from the requested regression line.

Different indices can measure the formal goodness of an MDS solution. These indices measure how well the obtained configuration of points represents the set of pairwise dissimilarities. The common way to do this is by computing the MDS solution's goodness-of-fit criterion, such as the *stress*. This latter is a loss function: it is zero when the solution is perfect; otherwise, it is greater than zero. *Stress* aggregates into one measure the deviations of the points from the regression line in a data-versus-distances plot (Shepard diagram). A perfect MDS solution has *stress* = 0. If this is true, then the distances of the MDS configuration represent the data precisely (in the desired sense).

For each value of the number of dimensions, m , the configuration of points that has the smallest goodness-of-fit criterion is called the best-fitting configuration in m dimensions. A rule of thumb for judging the quality of an MDS solution has been provided by Kruskal based on his own experience with experimental and synthetic data: *stress* = 20% (poor), *stress* = 10% (fair), *stress* = 5% (good), *stress* = 2.5% (excellent), and *stress* = 0 (perfect). More recent articles caution against using a table like this since acceptable values of *stress* depends on the quality of the distance matrix and the number of objects in that matrix.

In the practical application of MDS, a critical question concerns the appropriate dimensionality of the resulting configuration of points needed to give an adequate representation of the observed dissimilarities. The goal is to keep the number of dimensions as small as possible, i.e., an MDS representation with considerably fewer dimensions. For illustrative or visualization purposes, the preferred number of dimensions to be chosen is two, at most, three dimensions. Using a two- or three-dimensional MDS solution makes more straightforward the interpretation of the results. Although a less well-fitted configuration in two dimensions may be preferable to one in several dimensions (where only projections of points can be graphically displayed), it is also essential to keep in mind that scaling with too few dimensions may distort the true MDS structure due to over-compression.

To choose an appropriate number of dimensions, it is advocated that several values of the number of dimensions are tried, and the goodness-of-fit criterion (such as the *stress*) of the final configuration of points is plotted against the number of dimensions. *Stress* always decreases as the number of dimensions increases; the greater the m , the smaller the expected *stress* (because higher-dimensional spaces offer more freedom for optimal positioning of points). What one looks at in this scree plot is an elbow in this curve, a point where the decrements in *stress* begin to be less pronounced, i.e., for which further increase in m does not significantly reduce *stress*. That point corresponds to the dimensionality that should be chosen. The rationale of this choice is that the elbow marks the point where MDS uses additional dimensions to essentially only scale the noise in the data after having succeeded in representing the systematic structure in the given dimensionality m .

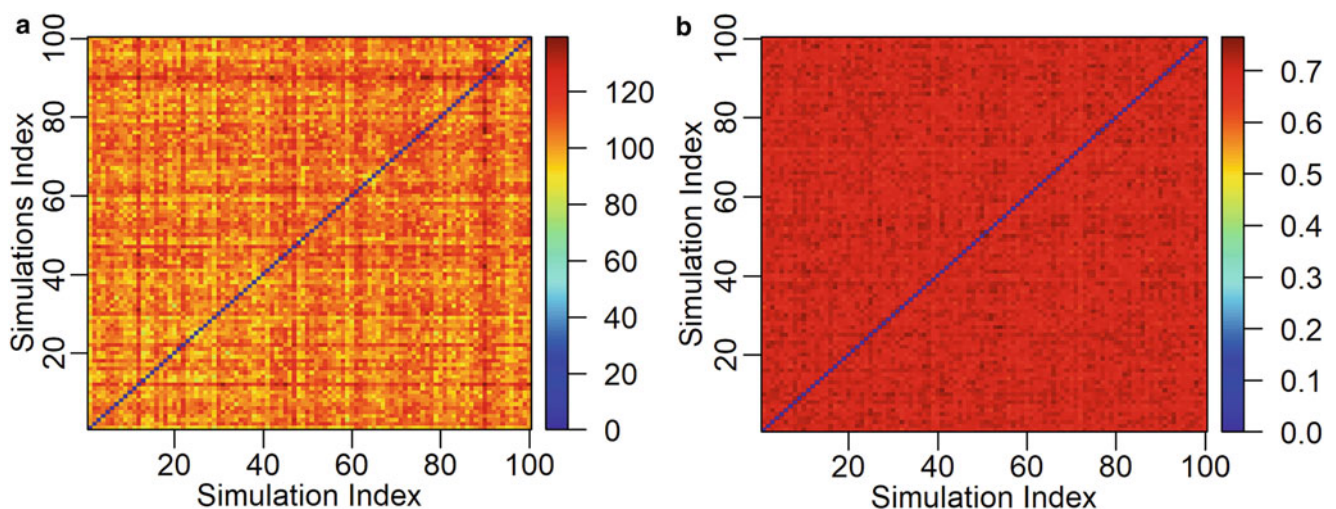
Application Examples

Multidimensional scaling is used in geostatistics for many purposes. MDS is used to model non-stationary spatial dependence structure through the so-called space deformation approach (Fouedjio 2016; Fouedjio et al. 2015). The stationary assumption of the spatial dependence structure may not be suitable because the Euclidean distance is not an appropriate measure of the spatial separation. The spatial dependence structure can, for instance, follow geological patterns that lead to non-stationarity. However, using a distance other than the Euclidean distance in an isotropic stationary spatial dependence structure model may result in an invalid second-

order structure model. In this case, one can use the metric MDS method to embed locations of the study domain in a Euclidean space where the distances between points approximate well the original distances; thus, the validity of the second-order structure model is guaranteed (Boisvert and Deutsch 2011; Rivest and Marcotte 2012; Ver Hoef et al. 2006; Løland and Høst 2003). MDS is also used to analyze an ensemble of geostatistical simulations (de Barros and Deutsch 2017; Tan et al. 2014). This is useful for identifying redundant simulations and selecting a subset of representative simulations for an accurate and realistic risk assessment and decision analysis. It is also helpful for falsification of simulations, where one tests if simulations are consistent (in Bayesian sense) with the data (Fouedjio et al. 2021).

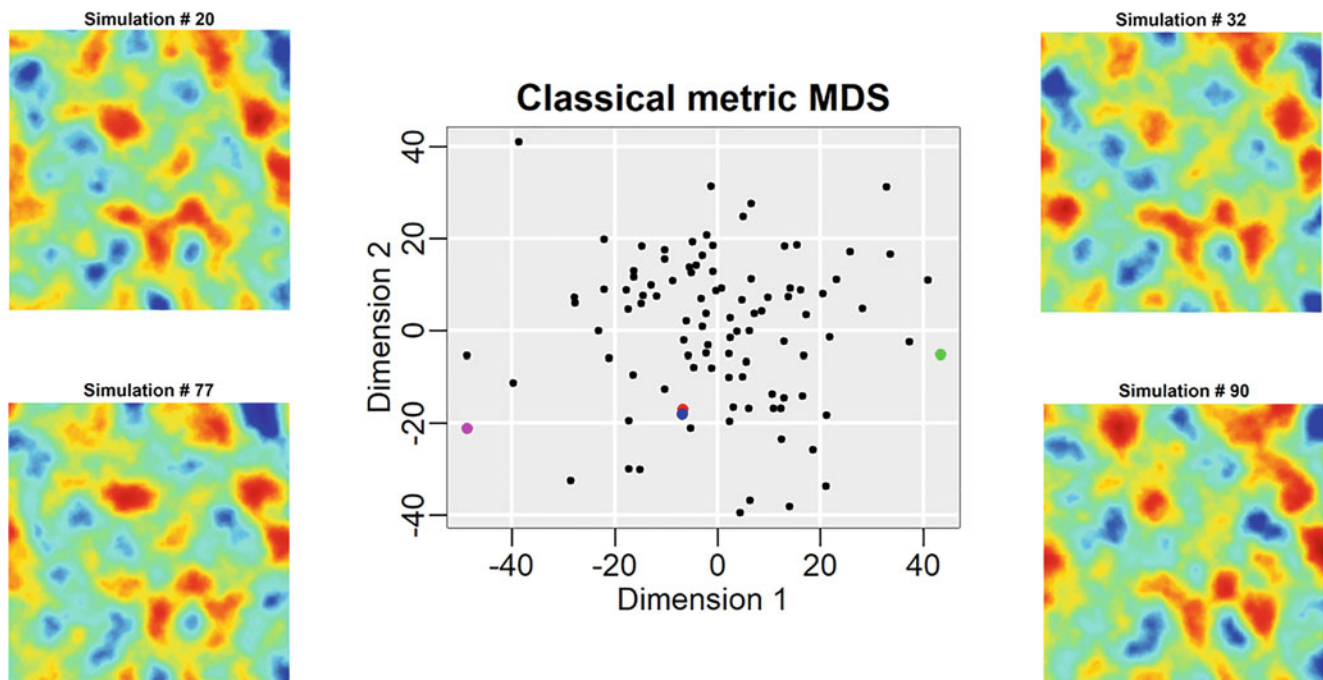
In the following, the application of MDS for visualizing an ensemble of geostatistical simulations is presented. Two examples considering continuous and categorical simulations are given. The first example considers 100 continuous geostatistical conditional realizations on 200×200 grid cells. The dissimilarity measure between two simulations (realizations) is defined as the square root sum of square differences between realizations, which is calculated cell-by-cell. So, the dissimilarity measure here is simply the Euclidean distance in the space spanned by the number of cells. The second example refers to 100 categorical geostatistical conditional realizations on 200×200 grid cells. There are four categories. The dissimilarity measure between two categorical simulations is defined using the Jaccard distance which is a distance measure dedicated to categorical data.

The classical metric MDS is applied to the resulting dissimilarity matrix in both examples. Figure 1 shows the



Multidimensional Scaling, Fig. 1 (a) (100×100) dissimilarity matrix (based on Euclidean distance) of an ensemble of 100 continuous geostatistical conditional realizations; (b) (100×100) dissimilarity

matrix (based on Jaccard distance) of an ensemble of 100 categorical geostatistical conditional realizations



Multidimensional Scaling, Fig. 2 Metric MDS representation of an ensemble of 100 continuous geostatistical conditional realizations; green dot represents the simulation #90; magenta dot represents the simulation

#77; red dot represents the simulation #20; blue dot represents the simulation #32; black dots represent the other simulations

dissimilarity matrix in the two examples. The 2D MDS representation of the ensemble of 100 continuous geostatistical conditional realizations is given in Fig. 2. Four simulations are also shown in the figure. Simulations #20 and #32 represent very close realizations; low and high values are well reproduced in both of the associated realizations. Simulations #77 and #90 represent realizations with large differences, which are reflected in their position in the MDS plot. Figure 3 depicts the 2D MDS representation of the ensemble of 100 categorical geostatistical conditional realizations as well as four realizations. Realizations #26 and #58 are very close in the MDS plot and show similar patterns. Realizations #74 and #98 have significant differences, which are reflected in their location in the MDS plot. In both examples, the classical metric MDS recreates the configuration of points in the $q = 99$ dimensional space that exactly matches the observed dissimilarities. Figure 4 presents the number of dimensions versus the goodness-of-fit criteria defined in Eq. (2).

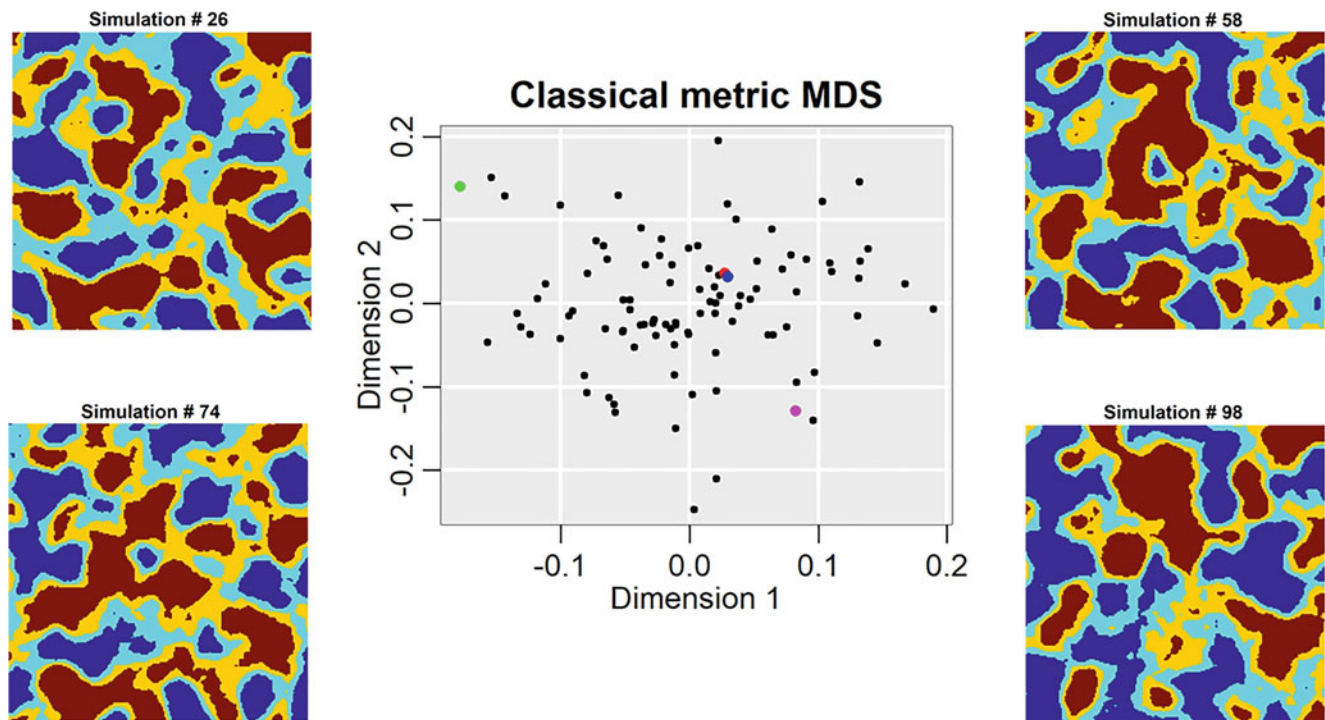
Conclusion

Multidimensional scaling (MDS) represents (dis)similarities among objects by mapping points (representing the objects) into a multidimensional space in such a way that the inter-

point distances best agree with the (dis)similarities between the objects. Under MDS, one can visually inspect the (dis)similarities among the objects and investigate the principle underlying the organization of the (dis)similarities among these objects.

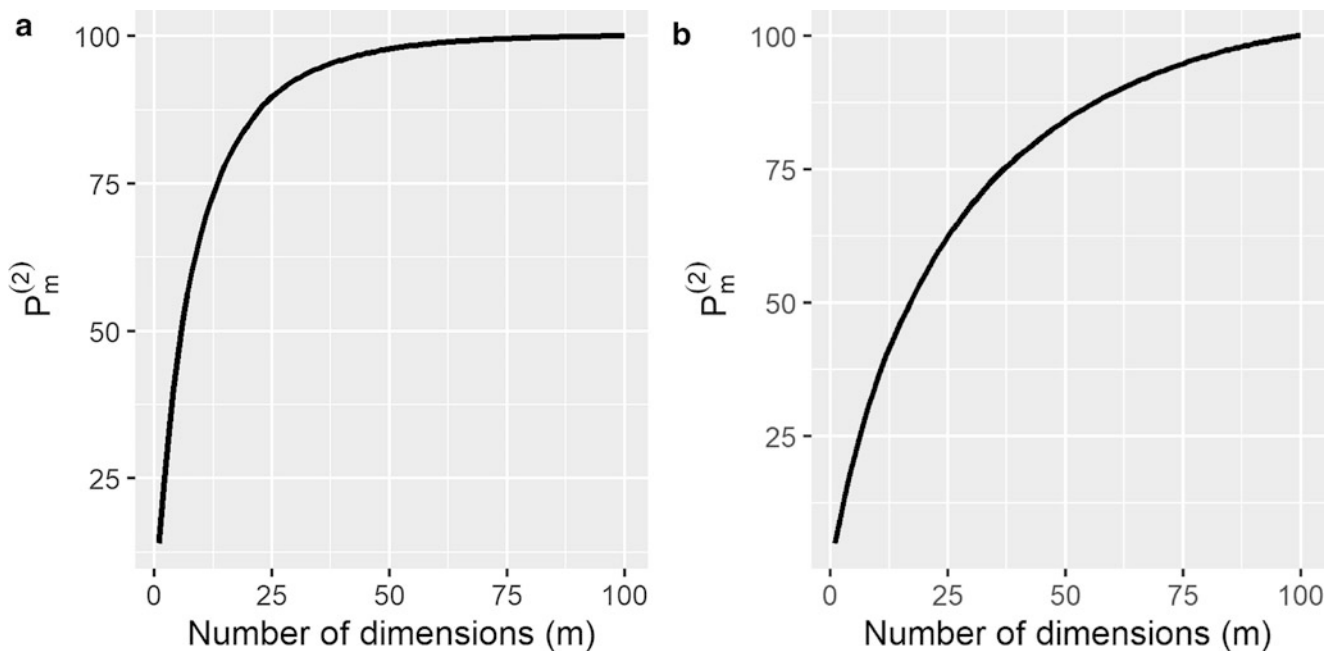
Although one can globally distinguish two types of MDS (metric multidimensional scaling and non-metric multidimensional scaling), there are various MDS models (or algorithms) depending on some factors such as the goodness-of-fit criterion used, the optimization method adopted, and the underlying transformation function chosen.

Like principal component analysis (PCA), MDS is also a dimension reduction technique because it aims to find a set of points in a low-dimensional space that approximate the possibly high-dimensional configuration represented by the original (dis)similarity matrix. However, MDS notably differs from PCA. Under the PCA, the data points were directly observable as n points in m dimensional space, but under MDS, one can only observe a function of the data points, namely, the dissimilarities. Like PCA, MDS can be used with supplementary or illustrative elements projected onto the dimensions after being computed. MDS is often combined with cluster analysis. Indeed, cluster analysis is performed on the low-dimensional representation of data resulting from MDS.



Multidimensional Scaling, Fig. 3 Metric MDS representation of an ensemble of 100 categorical geostatistical conditional realizations; green dot represents the simulation #74; magenta dot represents the simulation

#98; red dot represents the simulation #26; blue dot represents the simulation #58; black dots represent the other simulations



Multidimensional Scaling, Fig. 4 Number of dimensions versus goodness-of-fit of classical metric MDS representation: (a) ensemble of 100 continuous geostatistical conditional realizations and (b) ensemble of 100 categorical geostatistical conditional realizations

Cross-References

- [Cluster Analysis and Classification](#)
- [Principal Component Analysis](#)
- [t-Distributed Stochastic Neighbor Embedding](#)

Bibliography

- Boisvert JB, Deutsch CV (2011) Programs for kriging and sequential Gaussian simulation with locally varying anisotropy using non-Euclidean distances. *Comput Geosci* 37(4):495–510
- Borg I, Groenen P (2005) Modern multidimensional scaling: theory and applications, Springer series in statistics. Springer, New York
- Borg I, Groenen P, Mair P (2018) Applied multidimensional scaling and unfolding, Springer briefs in statistics. Springer International Publishing
- Cox T, Cox M (2000) Multidimensional scaling, Chapman & Hall/CRC monographs on statistics & applied probability, 2nd edn. CRC Press
- de Barros G, Deutsch CV (2017) Optimal ordering of realizations for visualization and presentation. *Comput Geosci* 103:51–58
- Fouedjio F (2016) Space deformation non-stationary geostatistical approach for prediction of geological objects: case study at El Teniente mine (Chile). *Nat Resour Res* 25(3):283–296
- Fouedjio F, Desassis N, Romary T (2015) Estimation of space deformation model for nonstationary random functions. *Spat Stat* 13:45–61
- Fouedjio F, Scheidt C, Yang L, Wang Y, Caers J (2021) Conditional simulation of categorical spatial variables using Gibbs sampling of a truncated multivariate normal distribution subject to linear inequality constraints. *Stoch Env Res Risk A* 35(2):457–480
- Løland A, Høst G (2003) Spatial covariance modelling in a complex coastal domain by multidimensional scaling. *Environmetrics* 14(3):307–321
- Rivest M, Marcotte D (2012) Kriging groundwater solute concentrations using flow coordinates and nonstationary covariance functions. *J Hydrol* 472:238–253
- Tan X, Tahmasebi P, Caers J (2014) Comparing training-image based algorithms using an analysis of distance. *Math Geosci* 46(2):149–169
- Ver Hoef JM, Peterson E, Theobald D (2006) Spatial statistical models that use flow and stream distance. *Environ Ecol Stat* 13(4):449–464

Multifractals

Renguang Zuo

State Key Laboratory of Geological Processes and Mineral Resources, China University of Geosciences, Wuhan, China

Definition

Multifractal measures, or multifractals, proposed by Benoit Mandelbrot in 1972, are used to describe the phenomenon that the distribution of physical or other quantities is self-similar across many different scales in a geometric support (Feder 1988; Evertsz and Mandelbrot 1992; Cheng and Agterberg 1996). Multifractals are common in nature

including, for example, fully developed turbulence and distribution of gold across the Earth's surface (Halsey et al. 1986). Multifractals can be characterized by a continuous spectrum of fractal dimensions (Cheng and Agterberg 1996). Fractal is a mathematical concept that describes a pattern that repeats itself at many different scales (Mandelbrot 1967, 1977). Fractals are related to the notion of self-similarity, according to which the geometrical pattern on a specific length scale is the same as that on another length scale (Mandelbrot 1967). The fractal dimension (D) can be estimated from the negative slope of best fitting straight line on a log–log plot of number of objects versus the scale. The fractal dimension D is a positive real number, and reflects the complexity of the object.

Introduction

Multifractals have been shown to provide a useful mathematical concept in studies of the spatial distributions of physical and chemical quantities with geometrical supports. Many researchers have advocated the use of multifractals in the geosciences (e.g., Schertzer and Lovejoy 1991; Korvin 1992; Turcotte 1992; Agterberg 1994; Cheng et al. 1994; Cheng and Agterberg 1995, 1996). Multifractals have been successfully employed for characterizing both the frequency and spatial distributions of geological point processes, which occur widely in the geosciences, e.g., for mineral deposits, oil-producing wells, earthquakes, and landslides (e.g., Carlson 1991; Cheng and Agterberg 1995). At the same time, various multifractal models and related indices have been developed for measuring the complexity of geospatial patterns, such as maps of fractures (Agterberg et al. 1996) and geochemical elements (Xie and Bao 2004), and for separating geochemical anomalies from background in the context of mineral exploration (Cheng et al. 1994, 1999; Cheng 2007). The following sections introduce several multifractal models in the field of mathematical geoscience.

Multifractal Spectrum

The multifractal spectrum, a continuous function of singularity exponent α , is a commonly used tool to characterize multifractals. A monofractal corresponds to any single point in a multifractal spectrum, because it only has one singularity. Multifractal spectra are typically generated using one of three methods: the histogram method, the method of moments, or the direct determination method. The first two methods were introduced by Evertsz and Mandelbrot (1992), while the third method was proposed by Chhabra and Jensen (1989). In practice, the method of moments involving Legendre transformation is more widely used because it can yield better results than the other two methods (Agterberg 2014).

Meanwhile, the wavelet-based method (Arneodo et al. 1995) also can be used to construct a multifractal spectrum. For generating a multifractal spectrum for a geospatial pattern such as a geochemical map, let $\mu_i(\varepsilon)$ be the total metal amount of an element in the i th cell on a linear scale ε . The partition function $\chi_q(\varepsilon)$ then is given by (Evertsz and Mandelbrot 1992)

$$\chi_q(\varepsilon) = \sum_{N(\varepsilon)} \mu_i^q(\varepsilon), \quad (1)$$

where $N(\varepsilon)$ is the total number of cells with size ε . From a multifractal viewpoint, the relationship between $\chi_q(\varepsilon)$ and cell size ε can be modeled using a power-law statistic with $-\infty \leq q \leq +\infty$, i.e.,

$$\chi_q(\varepsilon) \propto \varepsilon^{\tau(q)}, \quad (2)$$

where $\propto$ represents proportionality, and $\tau(q)$ is the mass exponent of order q . The singularity exponent $\alpha(q)$ and the multifractal spectrum $f(\alpha)$ then can be obtained using the Legendre transformation (Evertsz and Mandelbrot 1992)

$$\alpha(q) = \frac{d\tau(q)}{dq}, \quad (3)$$

and

$$f[\alpha(q)] = \alpha(q)q - \tau(q), \quad (4)$$

Two indices (the multifractality and the asymmetry index) can be applied to quantify a multifractal spectrum. Cheng (1999) proposed an index based on the second derivative of the mass exponent function $\tau(q)$ at $q = 1$ to measure multifractality, i.e.,

$$\tau''(1) = \tau(2) - 2\tau(1) + \tau(0). \quad (5)$$

The multifractality provides a quantitative measure of irregularities (or regularities) in spatial distribution patterns. The higher the multifractality, the higher the degree of irregularity. The asymmetry index R (Xie and Bao 2004; Cheng 2014) is a measure of the symmetry of a multifractal spectrum:

$$R = \frac{\alpha(0) - \alpha_{\min}}{\alpha_{\max} - \alpha(0)}, \quad (6)$$

where $\alpha(0)$ is the value of α when $q = 0$, and $\alpha_{\min}$ and $\alpha_{\max}$ are the minimum and maximum values of α , respectively. $R > 1$ corresponds to a left-deviation multifractal spectrum with local enrichment; $R < 1$ corresponds to a right-deviation multifractal spectrum with local depletion; $R = 1$ suggests a symmetric multifractal spectrum.

Multifractal Inverse Distance Weighted Model

The multifractal inverse distance weighted model (MIDW) was developed by Cheng (2008). In the geosciences, sample point data are routinely interpolated into raster maps. The inverse distance weighted method (IDW), which is a frequently used interpolation method, assumes that the values of closer sample points contribute more to the interpolated values than the values of more distant sample points. The IDW method has the following disadvantages: (1) it does not consider the spatial variance in the values, and (2) it suffers from smoothing effects. The MIDW approach accounts for both the distance between neighboring points and the local structures and singularity of the field. It results in (Cheng 2008)

$$\rho(x_0, \varepsilon) = \left(\frac{\varepsilon}{\varepsilon_0}\right)^{\alpha(x_0)-E} \sum_{x_i \in \Omega(x_0, \varepsilon_{\max})} \lambda(\|x_i - x_0\|) Z(x_i), \quad (7)$$

where $\rho(x_0, \varepsilon)$ is the density function, $\alpha(x_0)$ is the singularity exponent at the location of x_0 , E is the Euclidian dimension of the pixel $\Omega(x, \varepsilon)$ (e.g., for a two-dimensional case, $E = 2$), ε is the pixel size, and $Z(x)$ is the value at point location x . In addition, $\lambda(\|x_i - x_0\|)$ is the weighting contribution of location x_i as a function of distance from x_i to x_0 , and $\sum \lambda(\|x_i - x_0\|) = 1$.

Number-Size Model

The number-size multifractal model (N-S), proposed by Mandelbrot (1982), is a basic fractal model for determining the relationship between the object size (S) and the number of objects (N) with sizes greater than or equal to a given size. The N-S model is parameterized as

$$N(\geq S) = cS^{-D}, \quad (8)$$

where c is a constant.

Concentration-Area Model

The concentration-area multifractal model (C-A) was proposed by Cheng et al. (1994). This model quantifies the relationship between the concentration and the area (A) where the concentration is greater than or equal to a given elemental concentration (ρ). This model has been used for identifying the threshold for separating geochemical anomalies from background. The C-A model is parameterized as

$$A(\geq \rho) = c\rho^{-D}, \quad (9)$$

On a log-log plot of $A(\rho)$ versus ρ , more than two straight lines can be fitted using the least squares method. The optimal

threshold for separating geochemical anomalies from background is the cutoff value of the fitted straight lines.

Spectrum–Area Model

The spectrum–area multifractal model (S–A) was developed by Cheng et al. (1999), for identifying geochemical anomalies in the frequency domain. The S–A model quantifies the power spectrum density (P) and area (A) with the power spectrum density above a given value, and is parameterized as

$$A(\geq P) = cP^{-D}. \quad (10)$$

The S–A model was developed based on the C–A model. The C–A model deals with data in the spatial domain, while the S–A model processes data in the frequency domain. Geospatial patterns can be converted into the frequency domain using Fourier transformation. A double logarithmic plot of P versus A with the spectral density above P can be constructed. M straight lines ($M \geq 2$) can be typically fitted, based on the S–A plot, using the least squares method. M straight lines define M filters for decomposing the original map into several components. Anomalous and background components can be obtained by converting each of these

components from the frequency domain into the spatial domain, using inverse Fourier transformation.

Local Singularity Analysis

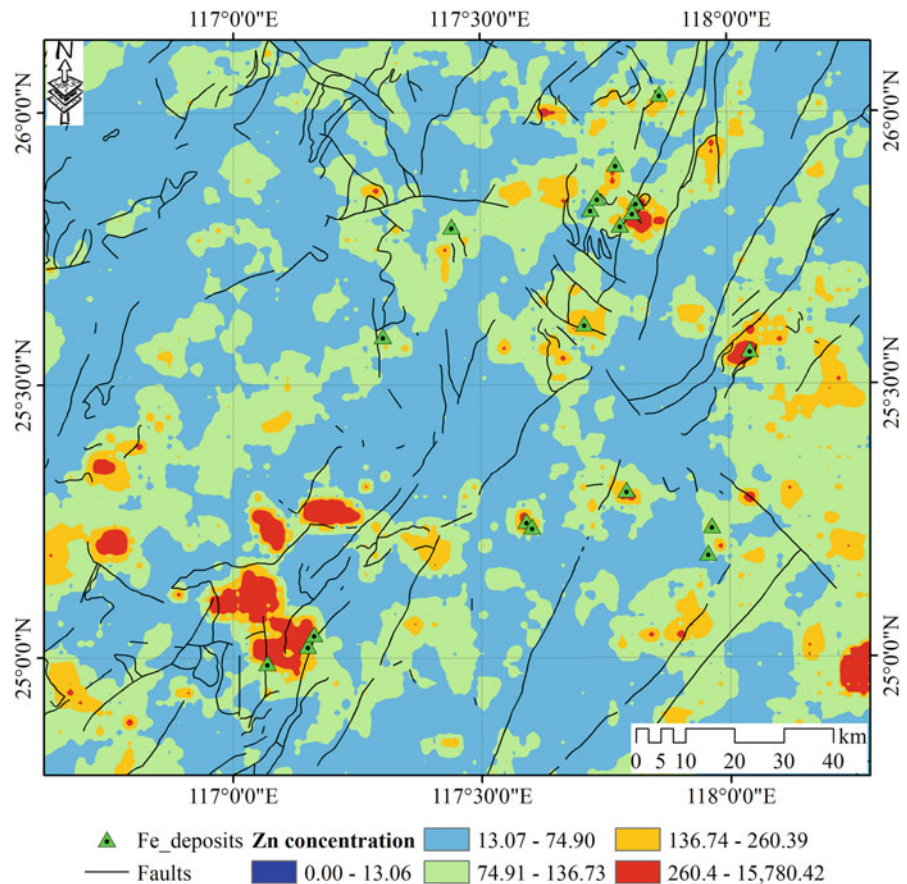
Geological singular events, such as earthquakes and mineralization, typically lead to anomalous energy release and/or mass accumulation within a relatively narrow spatial-temporal interval (Cheng 2007). The energy release and/or mass accumulation patterns exhibit multifractal characteristics. Local singularity analysis (LSA) has been developed to quantify the multifractal distribution of geospatial patterns. For instance, when measuring the enrichment or depletion for a geochemical pattern, LSA yields the following power-law relation:

$$\mu(A) = cA^{\frac{\alpha}{2}}. \quad (11)$$

where $\mu(A)$ represents the total amount of metal within a region with area A , c is a constant, and α is the singularity exponent. The singularity exponent can be calculated using the ratio of the logarithmic transformation of measure μ and area A as follows:

Multifractals,

Fig. 1 Geochemical map of Zn in the Southwest Fujian Province of China



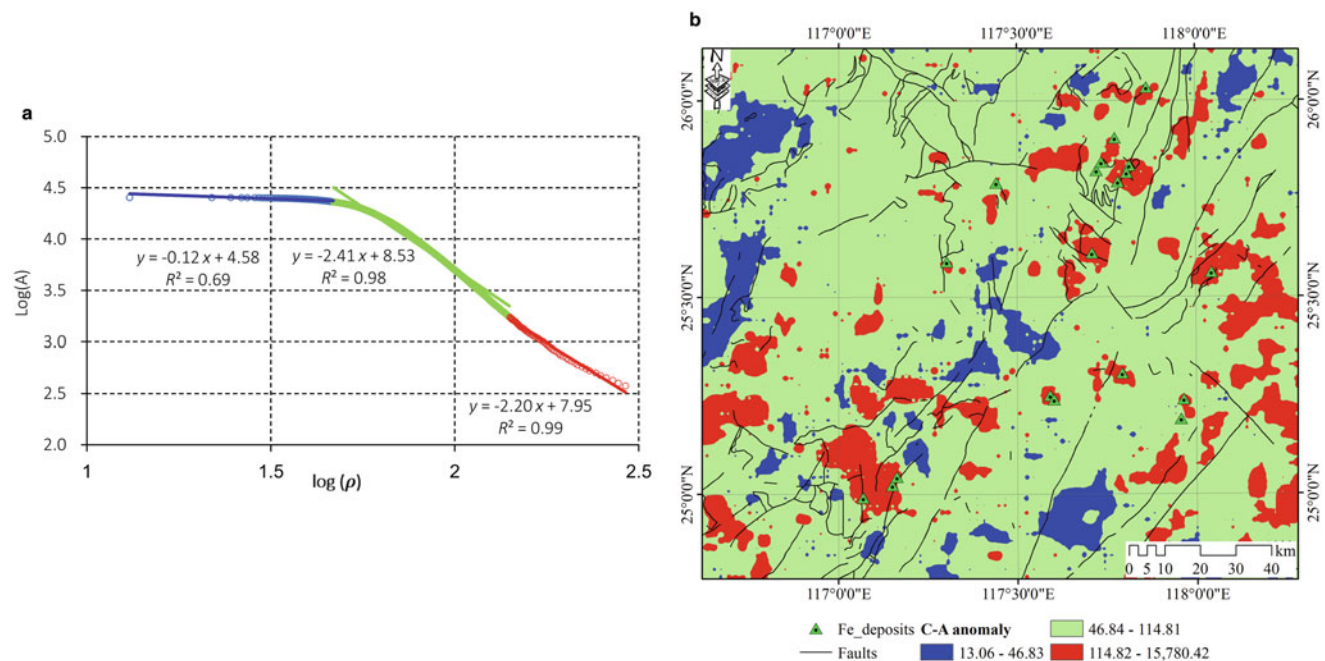
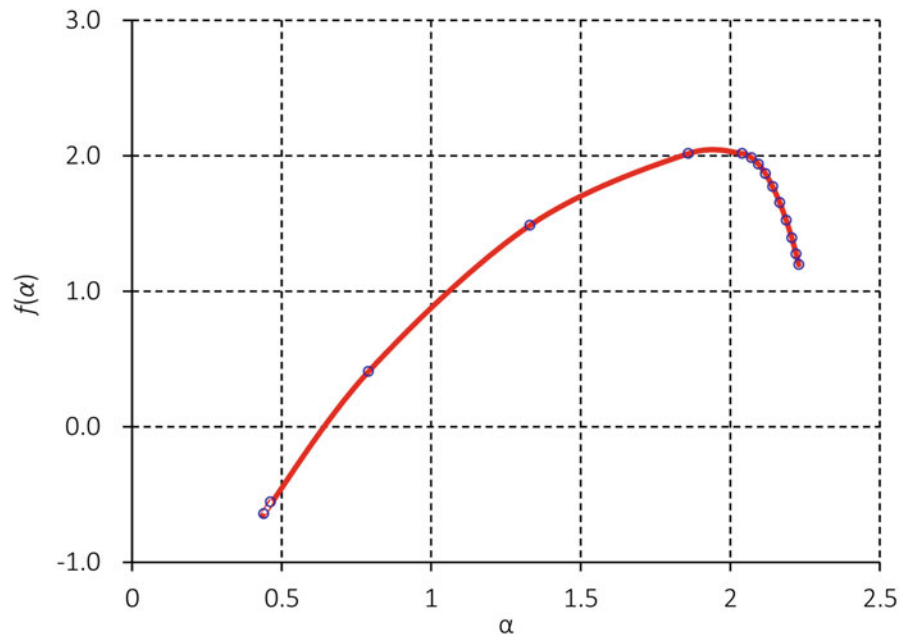
$$\alpha = \log\left(\frac{\mu_1}{\mu_2}\right) / \log\left(\frac{A_1}{A_2}\right)^{\frac{1}{2}}, \quad (12)$$

The sliding window technique can be used to estimate the local singularity exponent. For example, for a two-

dimensional (2D) geochemical map, let $A(r)$ be a set of square sliding windows centered at a certain location x , and let the size of each window be r , calculate the average concentration value $C[A(r_i)]$ for each window size r_i . On the log-log plot of $C[A(r_i)]$ versus r_i , a straight line can be fitted as

Multifractals,

Fig. 2 Multifractal spectrum of the geochemical pattern of Zn



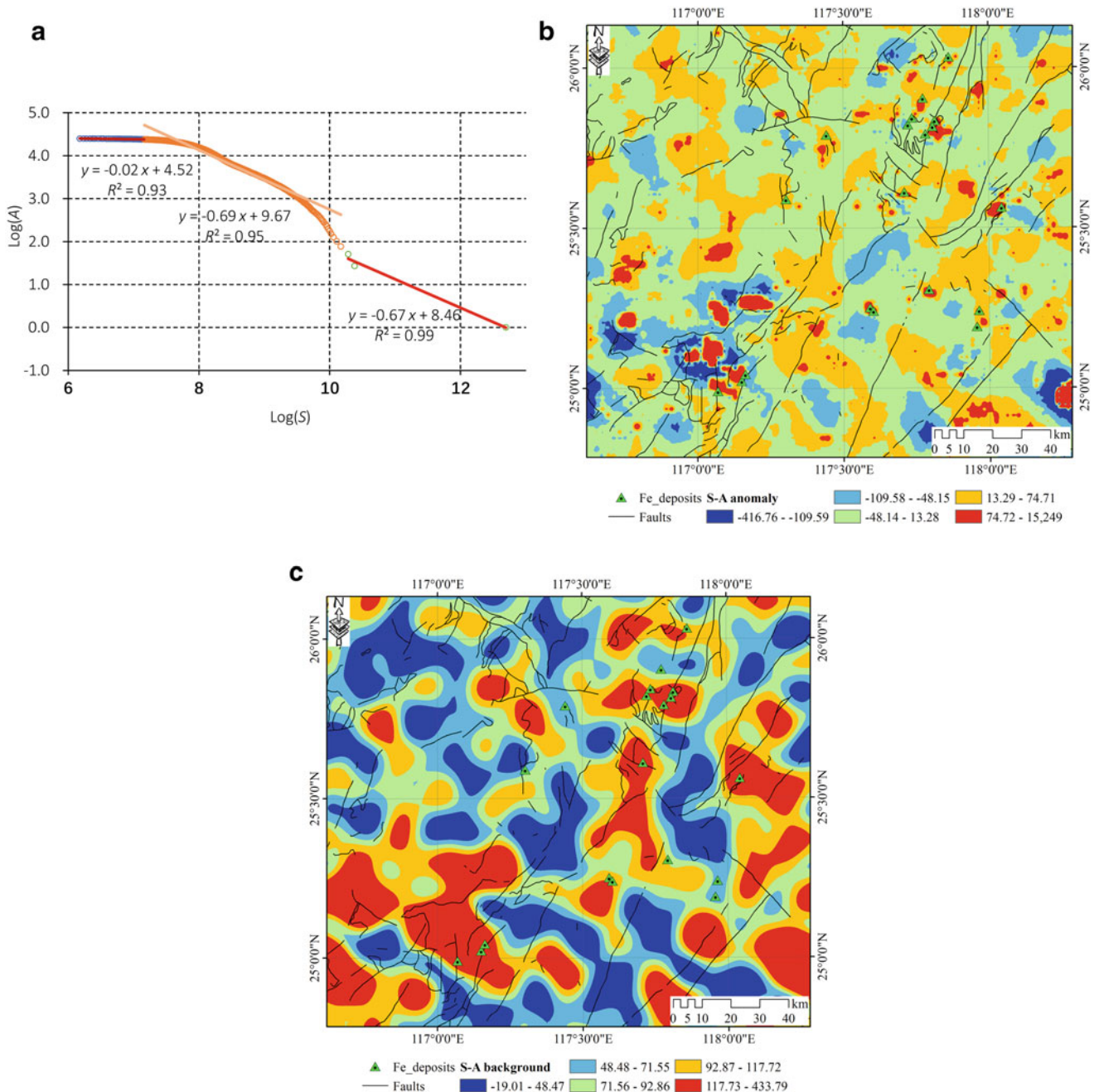
Multifractals, Fig. 3 (a) Log-log plot of the cumulative area versus the concentration of Zn, and (b) the map of Zn anomalies identified using the C-A model

$$\log\{C[A(r)]\} = c + (\alpha - 2) \log(r). \quad (13)$$

The value of $\alpha-2$ can be estimated as the slope of the best-fitting straight line. For every location on a 2D field, $\alpha > 2$ suggests a depletion pattern, $\alpha < 2$ suggests an enriched pattern, and $\alpha = 2$ suggests a normal pattern. LSA can effectively detect weak anomalies in a complex geological setting (Cheng 2007).

Case Study

A geochemical map of Zn (Fig. 1) from the southwest Fujian Province of China was studied from a multifractal viewpoint in support of ArcFractal, which is an Add-in ArcGIS for processing geospatial data using fractal and multifractal models (Zuo and Wang 2020). The Zn data were collected from the Chinese National Geochemical Mapping Project



Multifractals, Fig. 4 (a) Log–log plot of the cumulative area versus the concentration spectrum density, and (b) anomaly and (c) background maps of Zn identified using the S–A model

(Xie et al. 1997). For this dataset, the density of the geochemical data points was about one sample per km^2 , out of which four samples were combined into one sample, capturing 4 km^2 . The raster map of Zn with a cell size of $1 \text{ km} \times 1 \text{ km}$ was created using the IDW method. Previous studies had suggested Zn was one of indicator elements for the presence of Fe-polymetallic mineralization in the study area (Zuo 2018). The multifractal spectrum of Zn is a continuous curve, not a single point (Fig. 2), indicating a multifractal distribution. Meanwhile, R is 3.65 (greater than 1.0), suggesting a left-deviation of the multifractal spectrum and a local enrichment pattern of Zn.

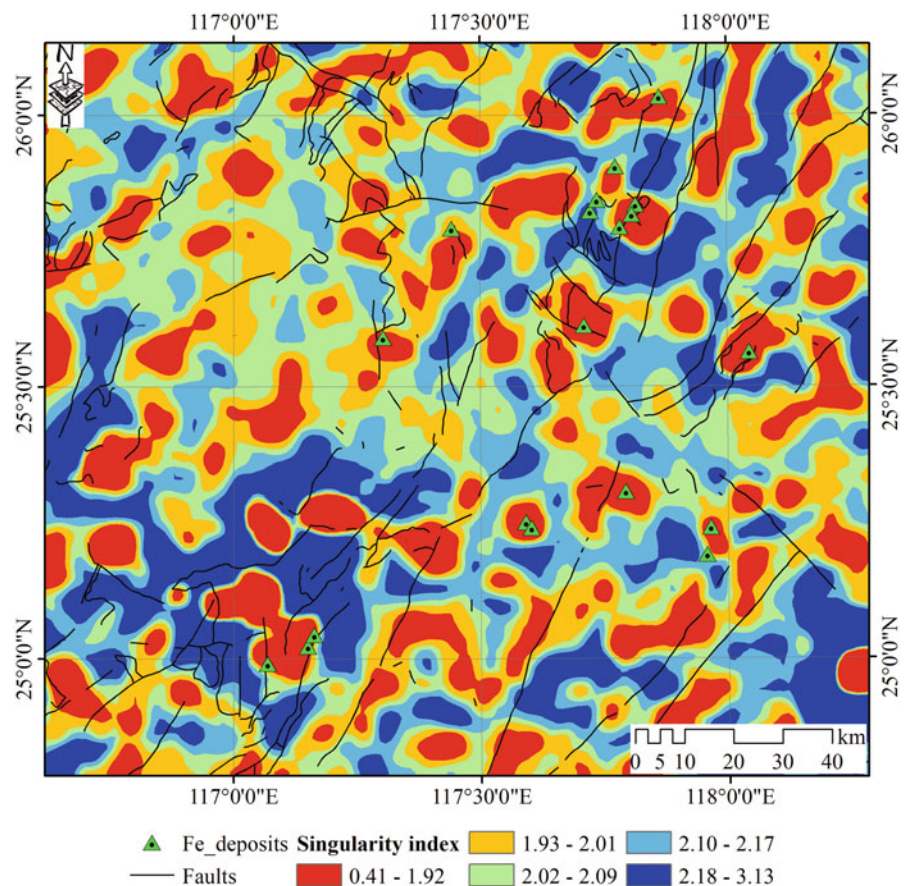
In the C–A plot, three straight lines could be fitted with different slopes (Fig. 3a). The right, middle, and left lines represent high anomalies, moderate anomalies, and background (Fig. 3b), respectively. Three straight lines also could be fitted in the S–A plot (Fig. 4a) representing noise, anomaly (Fig. 4b), and background (Fig. 4c) patterns of Zn, from left to right. The results of the S–A plot analysis show that the geochemical background of Zn (Fig. 4c) varies throughout the study area, meaning that different subareas have different background and anomaly values. These

observations suggest a complex geological setting of the study area. The anomalous patterns of Zn detected by the C–A, S–A, and LSA (Fig. 5) show close spatial correlations with the locations of known mineral deposits, indicating that these multifractal models can effectively identify geochemical anomalies on varied geochemical backgrounds linked to complex geological settings. It should also be noted that the S–A model suffers from edge effects (Ge et al. 2005), and several parameters (such as window shape and window size) affect the results of the LSA (Zuo et al. 2013).

Summary

Various geological processes or events exhibit complex and nonlinear characteristics, which can be modeled by using the concept of multifractals. Multifractals have gained significant attention in the field of mathematical geosciences, and have been used for processing geoscience data or exploring geological phenomena. This allows a full understanding of the origin, evolution, and future of Earth processes, with the aim of predicting and assessing its resources, environments, and

Multifractals, Fig. 5 Spatial distribution of the Zn singularity



natural hazards. Future studies should focus on obtaining insight into the mechanisms of geological processes and events from the multifractal viewpoint. Meanwhile, more case studies related to multifractal models are required for supporting mineral exploration, environmental assessment, and the effective prevention of natural disasters.

Cross-References

- [Concentration-Area Plot](#)
- [Local Singularity Analysis](#)
- [Inverse Distance Weight](#)
- [Spectrum-Area Method](#)

Bibliography

- Agterberg FP (1994) Fractals, multifractals, and change of support. In: Dimitrakopoulos R (ed) *Geostatistics for the next century*. Kluwer, Dordrecht, pp 223–234
- Agterberg FP (2014) *Geomathematics: theoretical foundations, applications and future developments*. Springer, Cham, p 569
- Agterberg FP, Cheng Q, Brown A, Good D (1996) Multifractal modeling of fractures in Lake du Bonnet Batholith. *Manitoba Comput Geosci* 22:497–507
- Arneodo A, Bacry E, Muzy J (1995) The thermodynamics of fractals revisited with wavelets. *Physica A* 213:232–275
- Carlson CA (1991) Spatial distribution of ore deposits. *Geology* 19: 111–114
- Cheng Q (1999) Multifractality and spatial statistics. *Comput Geosci* 25: 949–961
- Cheng Q (2007) Mapping singularities with stream sediment geochemical data for prediction of undiscovered mineral deposits in Gejiu, Yunnan Province. *China Ore Geol Rev* 32:314–324
- Cheng Q (2008) Modeling local scaling properties for multiscale mapping. *Vadose Zone J* 7:525–532
- Cheng Q (2014) Generalized binomial multiplicative cascade processes and asymmetrical multifractal distributions. *Nonlinear Process Geophys* 21:477–487
- Cheng Q, Agterberg FP (1995) Multifractal modeling and spatial point processes. *Math Geol* 27:831–845
- Cheng Q, Agterberg FP (1996) Multifractal modeling and spatial statistics. *Math Geology* 28:1–16
- Cheng Q, Agterberg FP, Ballantyne SB (1994) The separation of geochemical anomalies from background by fractal methods. *J Geochem Explor* 51:109–130
- Cheng Q, Xu Y, Grunsky E (1999) Integrated spatial and spectral analysis for geochemical anomaly separation. In: Lippard SJ, Naess A, Sinding-Larsen R (eds) *Proceedings of the fifth annual conference of the international association for mathematical geology*, Trondheim, Norway 6–11th August, vol 1, pp 87–92
- Chhabra A, Jensen RV (1989) Direct determination of the $f(a)$ singularity spectrum. *Phys Rev Lett* 62(12):1327–1330
- Evertsz CJG, Mandelbrot BB (1992) Multifractal measures (appendix B). In: Peitgen HO, Jurgens H, Saupe D (eds) *Chaos and fractals*. Springer, New York, pp 922–953
- Feder J (1988) *Fractals*, vol 283. Plenum, New York
- Ge Y, Cheng Q, Zhang S (2005) Reduction of edge effects in spatial information extraction from regional geochemical data: a case study based on multifractal filtering technique. *Comput Geosci* 31: 545–554
- Halsey TC, Jensen MH, Kadanoff LP, Procaccia I, Shraiman BI (1986) Fractal measures and their singularities: the characterization of strange sets. *Phys Rev A* 33(2):1141
- Korvin G (1992) *Fractal models in the earth sciences*. Elsevier, Amsterdam, 369 p
- Mandelbrot BB (1967) How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science* 156:636–638
- Mandelbrot BB (1977) *Fractals: form chance and dimension*. W.H. Freeman, San Francisco
- Mandelbrot BB (1982) *The fractal geometry of nature*. W.H. Freeman, San Francisco
- Schertzer D, Lovejoy S (eds) (1991) *Nonlinear variability in geophysics*. Kluwer Academic Publishers, Dordrecht, 318 pp
- Turcotte DL (1992) *Fractals and chaos in geology and geophysics*. Cambridge University Press, Cambridge, 221p
- Xie S, Bao Z (2004) Fractal and multifractal properties of geochemical fields. *Math Geol* 36:847–864
- Xie X, Mu X, Ren T (1997) Geochemical mapping in China. *J Geochem Explor* 60:99–113
- Zuo R (2018) Selection of an elemental association related to mineralization using spatial analysis. *J Geochem Explor* 184:150–157
- Zuo R, Wang J (2020) ArcFractal: an ArcGIS add-in for processing geoscience data using fractal/multifractal models. *Nat Resour Res* 29:3–12
- Zuo R, Xia Q, Zhang D (2013) A comparison study of the C-A and S-A models with singularity analysis to identify geochemical anomalies in covered areas. *Appl Geochem* 33:165–172

Multilayer Perceptrons

C. Özgen Karacan

U.S. Geological Survey, Geology, Energy and Minerals Science Center, Reston, VA, USA

Definition

Artificial neural networks (ANNs) are adaptable systems that can solve problems that are difficult to describe with a mathematical relationship. They seek relationships between different types of datasets with their abilities to learn either with supervision or without. ANNs recognize patterns between input and output space and generalize solutions, in a way simulating the human brain's learning experience with many relatively simple individual processing elements, called neurons. Neurons are networked (network topology) in a number of ways depending on the problem type and complexity. One of the most widely used ANN learning techniques is supervised learning coupled with a multilayer perceptron (MLP) topology due to its flexible applicability to a wide range of modeling problems involving both general classification and regression. ANNs, due to this flexibility, have been applied to many fields since the 1990s and their theory, types (such as radial basis functions, random forest, self-organizing maps,

and generalized feedforward network), and vast applications have been both documented in literature (e.g., Maier and Dandy 2000; Grima 2000; Karacan 2008, 2009; Rogerson 2019) and have been implemented in different software packages (e.g., Khan 2020).

Elements of an MLP Network

An MLP simulates a highly interconnected structure and parallel computing ability of a human brain. The input-output flow of the MLP is determined by the connections between neurons. The operation of a single neuron is dependent on the weighted sum of the incoming signals and a bias term, fed through an activation function, resulting in an output. For multiple neurons in a single-layer network topology, this process can be shown in mathematical terms as (Grima 2000):

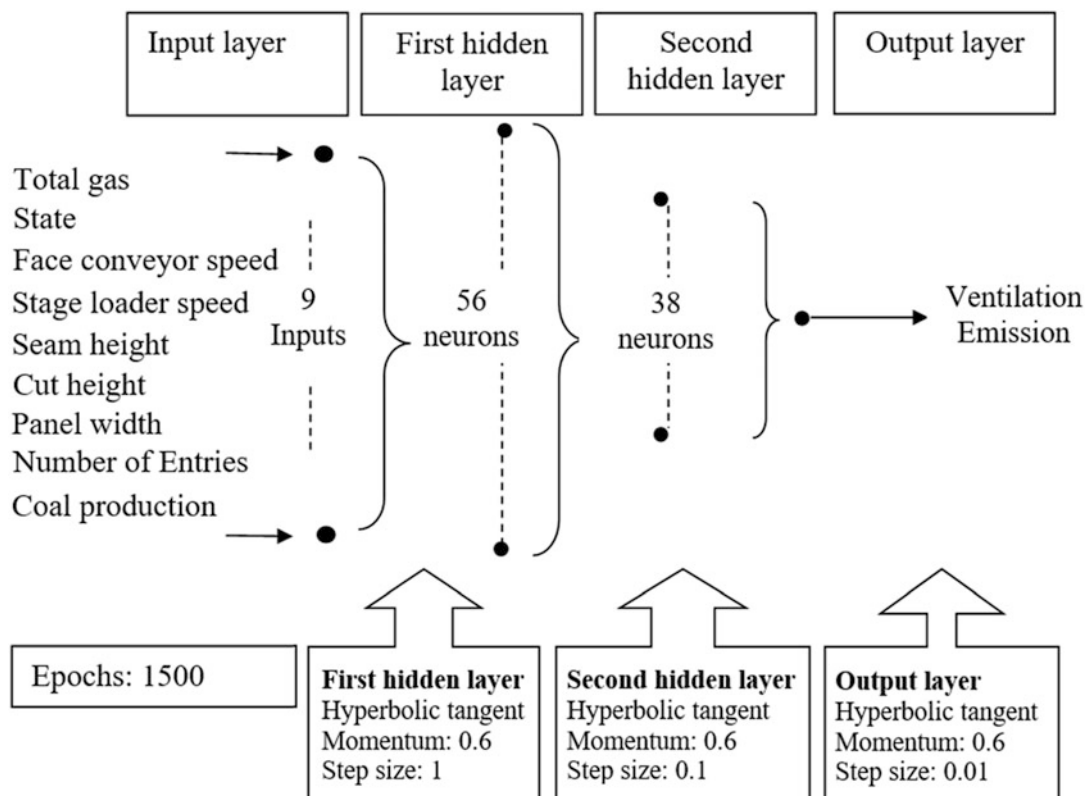
$$y_i = f \left(\sum_{r=1}^n w_{ir} f \left(\sum_{j=1}^m v_{rj} u_j + b_r \right) + c_i \right) \quad i = 1, \dots, l \quad (1)$$

where u is the $m \times 1$ input vector, y is the $l \times 1$ output vector, n the number of neurons in the hidden layer, v and w are the

weight factors, $f(\cdot)$ is the activation function, and b_r and c_i are the bias values of the neurons in the hidden and output layers, respectively. Depending on the problem type and complexity, neurons are networked in a number of ways. Figure 1 shows a typical MLP structure with its parameters.

The simplest MLP structure has three layers: input layer, one hidden layer, and an output layer. Usually the number of neurons in the input layer is equal to the number of inputs, and the number of output neurons are equal to the desired number of outputs. For classification problems, number of outputs is at least two. One of the critical issues in developing MLP-type ANNs is the choice of the number of neurons, and thus weights, in the hidden layers. Both the number of hidden layers and the number of neurons in each of those layers may vary depending on the complexity of the problem, which should be optimized, for effective functioning and predictive capability (please see next section).

Within the layers, the inputs are summed and processed by a nonlinear function called a transfer function. There are different types transfer functions, but the “hyperbolic tangent” is generally the preferred type for function approximation problems. The nonlinear nature of this function plays an important role in the performance of the network. In the case of classification, the objective of the output layer is to bin the



Multilayer Perceptrons, Fig. 1 Structure and parameters of an MLP network that was developed and used to predict methane emissions from ventilation system of longwall mines. (Modified from Karacan 2008)

information propagating from the hidden layers into classes with either higher probability or lower probability. The transfer function of choice in classification problems is the “softmax” function in the output layer. This function scales the outputs between 0 and 1 so that the higher probability class will have a higher value compared to others.

Learning and Parameter Optimization

The predictive capability of a network is determined by optimizing the weights between the connections in hidden layers, and input and output layers. The process of finding a suitable set of weights is called “training.” Since the weights and the network characteristics are used in making predictions, it is one of the most important steps in the development of the MLP. The most common way of training is using supervised training algorithms with gradient descent, which requires repeated feed (epochs) of both input vectors and the expected outputs of the training dataset to the network. During each epoch, an error of prediction is calculated for a given pattern (input-output) and summed for all output neurons to calculate an average mean squared error (MSE), which is propagated back to adjust the interconnection weights to minimize the error of prediction in next epochs. Due to the constant check of previous errors and update of learning parameters, the algorithm is called a “back propagation” algorithm. Depending on the complexity of the problem, 1500–2500 epochs may be needed. However, this is a critical process and the number of neurons and epochs should not be too high to avoid the system memorizing the patterns instead of learning them. Memorization may give successful training errors between predicted and actual outputs, but the system may fail in performance in the face of new inputs in test cases. Training may involve trials with different numbers of hidden layers and neuron quantity to find the best combination that can give an acceptable error with a preferably smaller, but robust, network.

Batch learning, where the error is stored up to the end of all training examples before the system is updated, is more stable and more common. During learning process, interconnection weights and learning rates are adjusted with change parameters, called step size and momentum (Fig. 1). Step size controls the change of weights, whereas momentum provides the properties from the previous weight update to make sure that the solution proceeds in the same direction as the previous iterations. During error minimization, a momentum factor between 0 and 1 is applied at each layer to increase the effective step size in shallow regions of the error surface and to move the solution toward a global optimum without being stuck at a local optimum. Also, the weight update for a given error, or “learning rate,” is controlled by a small number called the “step size.” These parameters are not static and

can be adjusted to obtain a lower error and more efficient error minimization.

Once the training phase is complete, the performance of the network needs to be validated on an independent data set using “cross validation.” Cross validation is a model evaluation method that gives an indication of how well the MLP will behave with new data. It is important that the validation data should not be part of training set. Cross validation is further followed by testing to evaluate the performance of the MLP. Due to the multiple stages of learning and testing, MLP usually requires a large dataset to be trained and optimized. Therefore, the available data should be randomized to avoid biases, normalized, and then partitioned for training, cross validation, and testing phases.

Which Inputs to Use? Reducing Dimensionality

MLP can tackle multivariate problems to find relations between different input and output patterns. An increased number of input parameters increases the need for using more hidden layers and neurons in each of the hidden layers. Also, some of the input parameters may not be as influential or may be highly correlated with others. Therefore, it is advantageous to reduce the number of input parameters in such a way as to reduce the dimensionality of the network but still to cope with the complexity, nonlinearity, and multivariable nature of the problem. Performing a principle component analysis (PCA) externally and prior to MLP development may help to reduce the complexity of the problem and allow for a more targeted search of the relations between inputs and outputs. Hybrid models can also address the complexity reduction issue internally, although presenting a large number of inputs to the MLP by relying on the network to determine the critical model inputs usually increases network size. Such models that select variables automatically, and thus eliminate the knowledge and expertise of the user, can be used when the modeled system is not well understood. If the contribution of each input parameter is understood, there are distinct advantages in using statistical methods, such as PCA, to determine the inputs prior to developing MLP models. One other advantage of reducing input variables is that more input-output patterns require more training data. Therefore, especially, if the training data are limited, it is advantageous to limit the number of inputs to only significant ones.

Summary and Conclusions

MLP-type ANN networks can perform function approximation and classification successfully, and they have been implemented in various software packages. Depending on

the nature of the problem and its complexity, the data requirements can be significant for a robust network. Therefore, employing a statistical method beforehand to determine significant parameters for the problem can reduce the network size and training data requirements. Also, the MLP performance can be improved by adjusting the network architecture, transfer functions, and learning parameters during the training phase. The objective of developing an MLP-type ANN should be minimization of training, cross validation, and testing errors; and developing it so that the network will not just memorize the patterns in the dataset at hand but will be flexible to model completely new data.

Cross-References

- [Backprojection Tomography](#)
- [Principal Component Analysis](#)
- [Topology in Geosciences](#)

Bibliography

- Maier HR, Dandy GC (2000) Neural networks for the prediction and forecasting of water resources variables: a review of modeling issues and applications. *Environmental Modeling and Software* 15: 101–124
- Grima MA (2000) Neuro-fuzzy modeling in engineering geology. A.A. Balkema, Rotterdam. 244 pp
- Rogerson J (2019) Recent progress in artificial neural networks. *Clanrye International*, New York. 191 pp
- Karacan CÖ (2008) Modeling and prediction of ventilation methane emissions of U.S. longwall mines using supervised artificial neural networks. *Int J Coal Geol* 73(3–4):371–387
- Karacan CÖ (2009) Degasification system selection for US longwall mines using an expert classification system. *Comput Geosci* 35(3): 515–526
- Khan S (2020) Backpropagation-based multi layer perceptron neural networks. <https://www.mathworks.com/matlabcentral/fileexchange/66477-backpropagation-based-multi-layer-perceptron-neural-networks>. MATLAB Central File Exchange. Retrieved November 12, 2020

Multiphase Flow

Juliana Y. Leung
University of Alberta, Edmonton, AB, Canada

Definition

A fluid phase refers to a homogeneous region of matter. It may consist of more than one chemical component. Phases are separated by distinct boundaries. For example, an aqueous

phase consists of water (H₂O) and other water-soluble solutes. In the context of mathematical geosciences, multiphase flow refers to the flow of more than one fluid phase simultaneously through a porous medium.

Introduction

Multiphase flow in geologic media is of great importance in a variety of subsurface engineering problems. In this entry, the fundamental mathematical formulation of the governing equations and relevant constitutive relationships are discussed. Solution approaches to these problems are highlighted. Finally, areas of current investigations and knowledge gaps are examined.

Mathematical Formulation of Multiphase Flow

The conservation of mass for component $i \in I, I = \{1, 2, \dots, N_c\}$ can be formulated as (Bird et al. 2007):

$$\frac{\partial \left[\phi \sum_{j=1}^{N_p} \rho_j S_j \omega_{ij} + (1 - \phi) \rho_s \omega_{is} \right]}{\partial t} = -\nabla \sum_{j=1}^{N_p} \left(\rho_j \omega_{ij} \vec{u}_j - \phi S_j \rho_j \vec{D}_{ij} \cdot \nabla \omega_{ij} \right) + R_i. \quad (1)$$

N_c and N_p refer to the total number of components and phases, respectively. R_i the source or sink term. ϕ and ρ_s denote porosity and solid density, respectively; ρ_j and S_j are the density and saturation of phase j . ω_{ij} and ω_{is} represent the mass fractions of component i in phase j or the solid (rock), respectively. $\vec{D}_{ij}$ is the dispersion tensor. For a special case with three components of water ($i = 1$), “oil” ($i = 2$), and “gas” ($i = 3$), if dispersion and fluid-rock interaction (e.g., adsorption) are omitted, the conservation of mass becomes:

$$\frac{\partial \left(\phi \sum_{j=1}^3 \rho_j S_j \omega_{ij} \right)}{\partial t} = -\nabla \sum_{j=1}^3 \rho_j \omega_{ij} \vec{u}_j + R_i. \quad (2)$$

It is further assumed that three immiscible phases exist: aqueous ($j = 1$), oleic ($j = 2$), and gaseous ($j = 3$). It is generally expected that the water component would be present only in the aqueous phase, while the oil and gas components may partition into both the oleic and gaseous phases. This conceptual model is described as a black-oil model, where “oil” and “gas” refer to the liquid and vapor

hydrocarbons at standard (surface) conditions, respectively. The conservation of mass for each phase is written as:

$$\text{Water : } \frac{\partial(\phi \rho_1 S_1)}{\partial t} = -\nabla \cdot (\rho_1 \vec{u}_1) + R_1; \quad (3)$$

$$\begin{aligned} \text{Oil : } & \frac{\partial[\phi(\rho_2 S_2 \omega_{22} + \rho_3 S_3 \omega_{23})]}{\partial t} \\ & = -\nabla \cdot (\rho_2 \omega_{22} \vec{u}_2 + \rho_3 \omega_{23} \vec{u}_3) + R_2; \end{aligned} \quad (4)$$

$$\begin{aligned} \text{Gas : } & \frac{\partial[\phi(\rho_2 S_2 \omega_{32} + \rho_3 S_3 \omega_{33})]}{\partial t} \\ & = -\nabla \cdot (\rho_2 \omega_{32} \vec{u}_2 + \rho_3 \omega_{33} \vec{u}_3) + R_3. \end{aligned} \quad (5)$$

Conservation of momentum in a porous medium is represented by Darcy's law (Bear 1972). The superficial velocity of phase j is denoted as $\vec{u}_j$:

$$\vec{u}_j = -\vec{k} \frac{k_{rj}(S_j)}{\mu_j(\omega_{ij}, \vec{u}_j)} \cdot (\nabla P_j + \rho_j \vec{g}); \quad (6)$$

$\vec{k}$ is the permeability tensor; $\vec{g}$ is the gravitational vector along the vertical direction; P_j , μ_j , and k_{rj} are the pressure, viscosity, and relative permeability, respectively, for phase j ; $\vec{\lambda}_j$ is the phase mobility tensor. It should be noted that Darcy's law is an approximation to the full conservation of momentum when the Reynolds number is approximately zero:

$$\frac{\partial(\rho_j \vec{v}_j)}{\partial t} = -(\nabla \cdot \rho_j \vec{v}_j) \vec{v}_j - \nabla \cdot \vec{\tau}_j - \nabla P_j - \rho_j \vec{g}. \quad (7)$$

$\vec{\tau}$ is the shear (deviatoric) stress tensor, and $\vec{v}$ is the interstitial velocity. Equation (7) can be simplified to Eq. (8) at steady-state conditions with low velocity:

$$\nabla \cdot \vec{\tau}_j = -\nabla P_j - \rho_j \vec{g}. \quad (8)$$

It can be shown that Darcy's law is obtained when setting $\nabla \cdot \vec{\tau}_j = \vec{\lambda}_j^{-1} \cdot \vec{u}_j$.

Constitutive Relationships

When multiple fluids are occupying a pore space, various types of interfacial forces would act between the fluids and the rock matrix. The distribution of these fluids within the

pore space (e.g., phase saturation, S_j) directly affects their transport behavior (e.g., phase relative permeability, k_{rj}).

Wettability is related to the interactions between the fluids and the solid. If two fluids are contacting a rock matrix (solid surface), the fluid whose molecules exhibit the strongest affinity for the atoms in rock will likely occupy much of the surface. Considering a system consisting of a solid (s) and two fluids (fluid-1 and fluid-3), the spreading coefficient (S) is defined as follows (Adler 1995):

$$S = \sigma_{s1} - \sigma_{s3} - \sigma_{13}. \quad (9)$$

where σ_{s1} , σ_{s3} , and σ_{13} are the interfacial tensions for the solid-fluid-1, solid-fluid-3, and fluid-1-fluid-3 interfaces, respectively. **Interfacial tension** is related to the imbalance of forces at the interface; it represents the differences in molecular attraction for molecules in the same phase, as opposed to the molecular attraction for molecules in another phase. If $S > 0$, fluid-3 is wetting the solid surface completely, with a contact angle of zero; otherwise, fluid-3 is only partially wetting the solid surface. **Contact angle** (θ) refers to the angle at which the interface between two fluids is intersecting the solid surface. It is defined according to Young's equation:

$$\cos \theta = \frac{\sigma_{s1} - \sigma_{s3}}{\sigma_{13}}. \quad (10)$$

If $\theta < 90^\circ$, fluid-3 is the wetting phase; if $\theta > 90^\circ$, fluid-1 is the wetting phase; if $\theta = 90^\circ$, the surface is neutral wet to both fluids.

If we consider the scenario where the non-wetting fluid (e.g., fluid-1) is injected into a rock sample, two cases would be observed. In the first case, if fluid-1 is present in a pore body and a connected (neighboring) pore throat is filled with a wetting fluid (e.g., fluid-3), fluid-1 would not spontaneously enter the pore throat; in fact, it only enters the pore throat if its pressure is greater than the pressure in fluid-3 by a certain threshold, which is the **capillary pressure** (P_c):

$$P_c = \sigma_{13} \left(\frac{1}{r_1} + \frac{1}{r_2} \right) = \frac{2\sigma_{13}}{r}. \quad (11)$$

The principal radii of curvature of the interface between the two fluids are denoted as r_1 and r_2 , while r is an "effective" radii of the interface. Capillary pressure is related to the difference in fluid pressure across an interface between two immiscible fluids in a confined space, such as a pore throat or a tube.

In the second case, the non-wetting fluid is present in a pore throat and a connected pore body is filled with a wetting fluid. There now exists an excess capillary pressure (ΔP_c), as given in Eq. (12). It is the difference in P_c for an interface in

the throat and the pore body. As a result, the non-wetting fluid would spontaneously enter the connected pore body.

$$\Delta P_c = \frac{2\sigma_{13}}{r_t} - \frac{2\sigma_{13}}{r_b} = 2\sigma_{13} \left(\frac{1}{r_t} - \frac{1}{r_b} \right). \quad (12)$$

r_t and r_b are the effective radii of the pore throat and pore body, respectively. The invasion of a non-wetting phase or draining of the wetting phase is called a **drainage** process, while the invasion of a wetting phase is called an **imbibition** process. As the capillary pressure increases, more non-wetting fluid would fill the porous medium, resulting in P_c being a function of fluid saturation. The P_c -saturation relationships are used as a constraint or auxiliary equation when solving the governing equations. In particular, it describes the dependency in pressure among the different fluids across an interface.

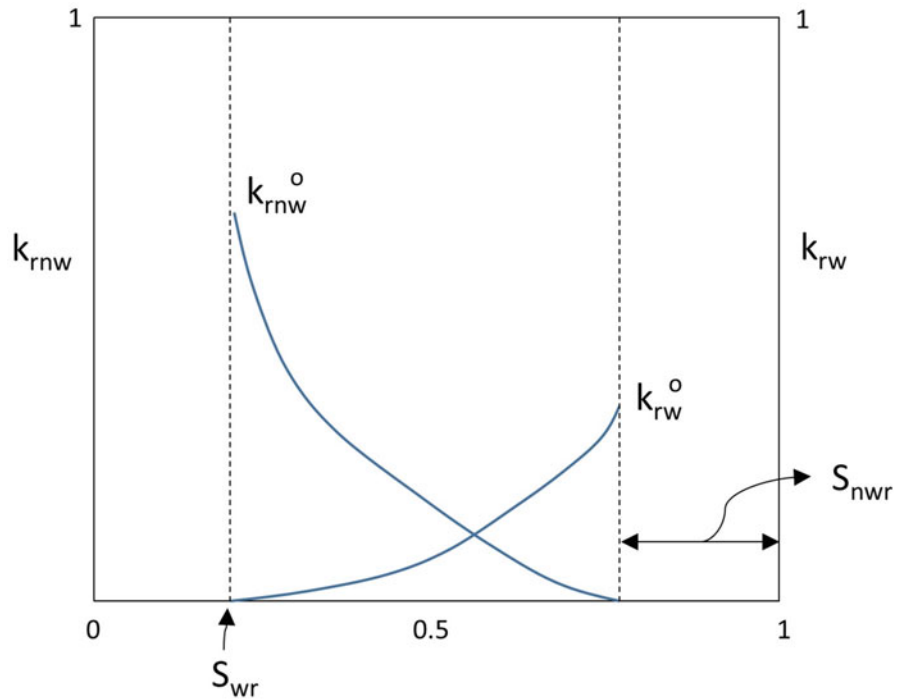
If there are three fluids present in a porous medium, one of them would be the wetting fluid, while the other two would be the non-wetting fluids (fluid-1 and fluid-2). However, fluid-1 and fluid-2 may exhibit different fluid-solid interfacial behaviors. If $\sigma_{13} > \sigma_{23}$, fluid-1 is the non-wetting phase and fluid-2 is the intermediate phase. The spreading phenomenon for a three-phase system is more much complex; further discussion of this subject, as well as contact line determination for a three-phase system, can be found in the cited references (Adler 1995; Blunt 2017).

Darcy's law offers a convenient way to describe flow behavior at a macroscopic level. The **relative permeability**

coefficients, in particular, are used to capture the relevant pore-scale physics; therefore, these relative permeability functions are strongly dependent on the pore structure and pore size distribution (Parker 1989). A set of connected pore bodies and throats can be represented quantitatively (or numerically) using a pore network model, which consists of a 3D network of nodes and channels with varying sizes (e.g., radii and lengths). This approach has gained significant popularity with advances in image analysis and digital rock physics. Different levels of capillary pressure are applied to the network to simulate the drainage and imbibition processes, and the macroscopic relative permeability relationships, as well as capillary pressure relationships, can be computed as a function of phase saturations and saturation history (or paths) (Blunt 2017). Although such theoretical computations using pore network models are feasible, empirical functions derived from experimental measurements involving actual rock samples are still most commonly adopted. A schematic of two-phase relative permeability functions for the wetting (k_{rw}) and non-wetting phases (k_{rnw}) are shown in Fig. 1.

Residual saturation is the saturation at which the relative permeability becomes zero, as denoted by S_{wr} or S_{nwr} for the wetting and non-wetting phases, respectively, in Fig. 1. As a result of capillary pressure, the non-wetting phase preferentially occupies the bigger pore bodies or centers of pore throats, while the wetting phase preferentially occupies the smaller pore throats, crevices between rock grains. The trapped non-wetting phase would likely pose a bigger

Multiphase Flow,
Fig. 1 Typical two-phase relative permeability functions



obstacle to the flowing wetting phase than vice versa. As a result, k_{rw}^o is generally higher than k_{rw}^o . In addition, the crossover between the two curves would typically occur where the wetting phase saturation is greater than 0.5 (Bear 1972).

Analytical Formulations Versus Numerical Simulations

The aforementioned governing equations (conservation of mass and Darcy's law) are solved by incorporating the relevant constitutive relationships (relative permeability and capillary pressure functions), as well as other auxiliary relationships such as a phase behavior or fluid model (e.g., an equation of state). Solutions to these problems may involve both analytical and numerical techniques. Boundary and initial conditions must be specified to complete the problem statement. Examples of boundary conditions may include no-flow boundary or constant pressure or flux. Analytical solutions are popular because of their simplicity and ease of application. They are widely adopted in the areas of pressure transient analysis or well testing (Lee et al. 2003). However, many assumptions regarding the flow regimes (e.g., transient flow, boundary-dominated flow, sequential depletion) or model set-up (e.g., geometric symmetry and homogeneous properties) are often invoked in analytical models.

Simulations entail numerical approximation of the solutions to the governing equations. Despite some recent developments in unstructured mesh generation and finite-element formulations for simulating multiphase flow, most existing commercial simulation packages predominantly focus on finite-difference/finite-volume schemes. A variety of structured and unstructured meshes, including Cartesian, corner point, perpendicular bisector (PEBI), and Delaunay triangular/tetrahedral, grids are available. Further details can be found in (Chen et al. 2006). Meshes should conform to the domain boundaries. The overall computational requirement would increase with more complex meshes and local refinement. Upscaling, a process that aims to replace a fine-scale detailed description of reservoir properties with a coarser scale description that has equivalent properties, is often used to generate a coarser mesh (Das and Hassanizadeh 2005).

Current Investigations and Knowledge Gaps

Much of the current research efforts have been focusing on the incorporation of additional physical phenomena. For example, coupling multiphase flow with geomechanical and geochemical considerations under non-isothermal conditions. This would require incorporating additional governing and auxiliary equations, such as energy balance, linear

momentum balance, and complex reaction kinetics, into the solution process. These coupled simulations are important for modeling more complex subsurface systems, including geothermal reservoirs, geologic storage of CO₂, and unconventional tight/shale formations with irregular multiscale fracture networks. More sophisticated numerical approximation or discretization methods (e.g., advanced finite element methods with mixed spaces) are needed to enhance computational efficiency and better capture the complex physical processes or geometries, which often introduce additional nonlinearities among the model parameters.

Another main challenge is the discrepancy between measurement and modeling scales, as well as the difficulty in representing heterogeneity (and its uncertainty) at these different scales. Detailed pore-scale models are computationally expensive and not suitable for field-level analysis. Macroscopic models are often employed. Darcy's law, in theory, is only applicable at scales greater than or equal to the representative elementary volume (REV) scale (Bear 1972). However, heterogeneity generally varies with scale. As a result, quantifying the sub-scale variability in multiphase flow and transport behavior, due to uncertain heterogeneity distribution below the modeling scale, is especially challenging (Vishal and Leung 2017).

Summary or Conclusions

Subsurface multiphase flow problems are encountered in a wide range of scientific and engineering applications. Complex heterogeneity observed in rock properties coupled with multitudes of physical phenomena render these problems to be both challenging and interesting. Models at both the pore- and macroscopic-levels are commonly adopted to analyze these problems. Future directions should focus on discovering, examining, and quantifying additional physical phenomena into existing or new models.

Cross-References

- [Geohydrology](#)
- [Porosity](#)
- [Rock Fracture Pattern and Modeling](#)

Bibliography

- Adler PM (ed) (1995) Multiphase flow in porous media. Kluwer Academic, Amsterdam
- Bear J (1972) Dynamics of fluids in porous media dynamics of fluids in porous media. Elsevier, New York
- Bird RB, Stewart WE, Lightfoot EN (2007) Transport phenomena, Revised 2nd edn. Wiley, New York

- Blunt MJ (2017) Multiphase flow in permeable media: a pore-scale perspective. Cambridge University Press, Cambridge
- Chen Z, Huan G, Ma Y (2006) Computational methods for multiphase flows in porous media, vol 2. SIAM, Philadelphia
- Das DB, Hassanizadeh SM (2005) Upscaling multiphase flow in porous media. Springer, Berlin
- Ertekin T, Abou-Kassem JH, King GR (2001) Basic applied reservoir simulation. Society of Petroleum Engineers, Houston
- Lee J, Rollings JB, Spivey JP (2003) Pressure transient testing, SPE textbook series, vol 9. SPE, Richardson
- Parker JC (1989) Multiphase flow and transport in porous media. Rev Geophys 27(3):311–328
- Vishal V, Leung JY (2017) A multi-scale particle-tracking framework for dispersive solute transport modeling. Comput Geosci 22(2):485–503

Multiple Correlation Coefficient

U. Mueller

School of Science, Edith Cowan University, Joondalup, WA, Australia

Definition

Given a set of random variables $\{X_1, X_2, \dots, X_K\}$, the multiple correlation coefficient $\rho_{X_i(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_K)}$ of the i^{th} random variable X_i on the remaining $K - 1$ variables is given by

$$\rho_{X_i(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_K)} = \sqrt{1 - \frac{\det(C(X))}{\det(C_i(X))}},$$

where $C_i(X)$ is the i^{th} principal minor of the correlation matrix

$$C(X) = \begin{bmatrix} 1 & \rho_{X_1 X_2} & \cdots & \rho_{X_1 X_K} \\ \rho_{X_2 X_1} & 1 & \cdots & \rho_{X_2 X_K} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{X_K X_1} & \rho_{X_K X_2} & \cdots & 1 \end{bmatrix},$$

i.e., the matrix obtained by removing the i^{th} row and the i^{th} column of $C(X)$ and $\rho_{X_i X_j}$ is the Pearson correlation coefficient between X_i and X_j .

Properties and Usage

The multiple correlation coefficient was first introduced by Pearson (1896) who also produced several further studies on it and related quantities such as the partial correlation coefficient (Pearson 1914). It is alternatively defined as the Pearson correlation coefficient between X_i and its best linear approximation by the remaining variables $\{X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_K\}$ (Abdi 2007). In contrast to the Pearson correlation coefficient between two random variables, which attains values between -1 and 1 , the multiple correlation coefficient is

nonnegative. In particular, the multiple correlation coefficient is equal to 0, when the random variables in $\{X_1, \dots, X_K\}$ are pairwise orthogonal, i.e., when $C(X) = I_K \times K$, i.e., when $\rho_{X_i X_j} = 0, i \neq j$. On the other hand, when the random variables in $\{X_1, \dots, X_K\}$ form a closed composition (see ► “Compositional Data”), which is a common situation when considering geochemical data, then $\rho_{X_i(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_K)} = 1$ for all $i, i = 1, \dots, K$. This property reflects nothing other than the constant sum constraint of compositional data, as the correlation matrix for the independent variables has 0 determinant.

The multiple correlation coefficient can be used to measure the degree of multicollinearity between variables as follows. The variance inflation factor VIF_i for the i^{th} variable (Hocking 2013) is defined as

$$VIF_i = \frac{1}{1 - \rho_{X_i(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_K)}^2}.$$

Its value is nonnegative, and it increases with decreasing magnitude of $\rho_{X_i(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_K)}$. There is a high degree of collinearity in the data if $VIF_i > 10$, for some $i, i = 1, \dots, 10$.

The most common application of the multiple correlation coefficient arises in the situation when there is a dependent variable Y and a set of predictor or dependent variables $\{X_1, X_2, \dots, X_K\}$. In this case, the multiple correlation coefficient may be computed from the matrix

$$C(Y, X) = \begin{bmatrix} \rho_{YY} & \rho_{YX_1} & \cdots & \rho_{YX_K} \\ \rho_{X_1 Y} & \rho_{X_1 X_1} & \cdots & \rho_{X_1 X_K} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{X_K Y} & \rho_{X_K X_1} & \cdots & \rho_{X_K X_K} \end{bmatrix}$$

as

$$\rho_{Y(X_1, X_2, \dots, X_K)} = \sqrt{1 - \frac{\det(C(Y, X))}{\det(C(X))}}$$

and the multiple correlation coefficient $\rho_{Y(X_1, X_2, \dots, X_K)}$ as a measure between the dependent variable and its predictors is well defined, and it does not matter whether the independent variables are random or fixed (Abdi 2007). The multiple correlation coefficient attains a particularly simple form, when the independent variables are pairwise orthogonal (Abdi 2007), in that case it is given as the sum of the square correlation coefficients:

$$\rho_{Y(X_1, X_2, \dots, X_K)} = \sqrt{\rho_{X_1 Y}^2 + \rho_{X_2 Y}^2 + \cdots + \rho_{X_K Y}^2}$$

This case arises, for example, when a regression is based on principal components (see ► “Principal Component Analysis”).

In regression settings, the square multiple correlation coefficient is interpreted as the fraction of the variation in the data explained by the model; this quantity is also known as the coefficient of determination.

The coefficient of determination $R^2_{Y(X_1, X_2, \dots, X_K)}$ is commonly used as one of the criteria for determining the best model from a set of competing regression models. As $R^2_{Y(X_1, X_2, \dots, X_K)}$ is biased upward, the adjusted R^2 is normally considered which is nothing other but $R^2_{Y(X_1, X_2, \dots, X_K)}$ multiplied by $\frac{n-1}{n-K-1}$ where n denotes the number of observations.

To test whether a given value of $R^2_{Y(X_1, X_2, \dots, X_K)}$ is statistically significant, the ratio

$$F = \frac{R^2_{Y(X_1, X_2, \dots, X_K)}}{1 - R^2_{Y(X_1, X_2, \dots, X_K)}} \times \frac{n - K - 1}{K}$$

is computed, and under the assumption of normally distributed errors and independence of the errors and estimates, this ratio follows an F -distribution with K and $n - K - 1$ degrees of freedom (Abdi 2007).

The usefulness of $R^2_{Y(X_1, X_2, \dots, X_K)}$ is context dependent: a low value which is statistically significant might be acceptable when the aim is to discover causal relationships, but even a relatively high R^2 might be unacceptable when accurate prediction is required (Crocker 1972). Already in 1914 Pearson cautioned about the temptation to add more variables with the aim to increase significance; this strategy will only succeed, if the variables to be added have weak correlation with the other predictors and among one another and high correlation with the dependent variable (Pearson 1914). It should be noted that multiple correlation or coefficient of determination is rarely used on their own when judging the quality of a statistical model but typically in combination with the Akaike or Bayes information criteria (James et al. 2013).

Applications

The multiple correlation coefficient is rarely used as a measure for the goodness of fit of a regression model. It is more common to use the coefficient of determination for this purpose. Exceptions are a study by Liu et al. (2019) in a context of fully mechanized coal mining where the height of the water conducting fractured zone in is modeled via nonlinear multiple linear regression with predictor variables thickness of coal seam, proportion of hard rock, length of panel, mined depth, and dip angle. The model is tested in detail for the Xiegou coal mine in the Shanxi Province, China. The multiple regression coefficient has also been used to evaluate competing models in mineral prospectivity mapping such as the case study on

iron ore presented by Mansouri et al. (2018). The general setting for these evaluations is often cross-validation.

In a geostatistical context, multiple correlation is also typically used in cross-validation as one criterion in the assessment of the quality of a geostatistical model (Webster and Oliver 2007).

Summary

The multiple correlation coefficient is a measure of association linked to multiple regression and is useful as a first-pass measure for assessing the quality of a regression model. In this case, it is calculated as the Pearson correlation between estimates and true values. However, it is also a valuable tool to appraise the input variables and their potential for variance inflation.

Cross-References

- [Compositional Data](#)
- [Principal Component Analysis](#)
- [Regression](#)

Bibliography

- Abdi H (2007) Multiple correlation coefficient. In: Salkind N (ed) Encyclopedia of measurement and statistics. Sage, Thousand Oaks
- Crocker DC (1972) Some interpretations of the multiple correlation coefficient. *Am Stat* 26(2):31–33
- Hocking RR (2013) Methods and applications of linear models: regression and the analysis of variance, 3rd edn. Wiley, Hoboken
- James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning with applications in R, 7th printing. Springer, New York
- Liu Y, Yuan S, Yang B, Liu J, Ye Z (2019) Predicting the height of the water-conducting fractured zone using multiple regression analysis and GIS. *Environ Earth Sci* 78:422. <https://doi.org/10.1007/s12665-019-8429-3>
- Mansouri E, Feizi F, Jafari Rad A, Arian M (2018) Remote-sensing data processing with the multivariate regression analysis method for iron mineral resource potential mapping: a case study in the Sarvian area, Central Iran. *Solid Earth* 9:373–384
- Pearson K (1896) Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia. *Philos Trans R Soc London Series A* 187:253–317. <https://doi.org/10.1098/rsta.1896.0007>
- Pearson K (1914) On certain errors with regard to multiple correlation occasionally made by those who have not adequately studied this subject. *Biometrika* 10(1):181–187
- Pearson K (1916) On some novel properties of partial and multiple correlation coefficients in a universe of manifold characteristics. *Biometrika* 11(3):233–238
- Webster R, Oliver MA (2007) Geostatistics for environmental scientists, 2nd edn. Wiley, Chichester

Multiple Point Statistics

Jef Caers¹, Gregoire Mariethoz² and Julian M. Ortiz³

¹Department of Geological Sciences, Stanford University, Stanford, CA, USA

²Institute of Earth Surface Dynamics, University of Lausanne, Lausanne, Switzerland

³The Robert M. Buchan Department of Mining, Queen's University, Kingston, ON, Canada

Definition

Multiple-point geostatistics is the field of study that focuses on the digital representation of the physical reality by reproducing high-order statistics inferred from training data, usually training images, that represent the spatial (and temporal) patterns expected in such context. Model-based and data-driven algorithms aim to reproduce such patterns in space-time by capturing and reproducing real world features through trends, hierarchies, and local spatial variation.

Multiple-point Geostatistics: A Historical Perspective

Geostatistics aims at modeling spatio-temporal phenomena, which encompasses a large class of types of data and modeling problems. From a niche statistical science in the 1970s, it has now evolved into a widely applicable and much used set of practical tools. The increased acquisition of massive spatio-temporal data and the vast increase in computational power is at the root of a renewed attraction to geostatistics. Fundamental societal problems, such as climate change, environmental destruction, and even pandemics, involve spatio-temporal data. Solution to these problems will require ingesting vast datasets while treating them in a repeatable, coherent, and consistent mathematical framework. Such coherent framework was first established by George Matheron (Matheron 1973), with the formulation of random function theory and extension of random variables to space-time. Space-time is more challenging than multivariate problems, since, as he discovered, mathematical models are required that address the special continuity that is space-time. While in multivariate analysis (Härdle and Simar 2015), covariance matrices are fundamental building block, geostatistics requires functions, such as covariance functions that need to be defined mathematically consistently (Matheron 1973), in as much as the covariance matrix needs to be positive definite to be useful in multivariate statistics.

For the first three decades, geostatistics followed classical principles outlined in probability theory. Estimating spatio-temporal behaviors was cast as a least-square problem of estimating stochastic functions. The unbiased linear solution to this least square formulation of the estimation problem was termed kriging (Matheron 1963). Kriging is one of the most general forms of linear regression, where all (linear) dependencies are accounted for. Today we see a resurgence of kriging in Computer Science and Machine learning, in the form of Gaussian process regression (Rasmussen and Williams 2006). Regression does not provide an uncertainty quantification on a function. The estimated functions are smoother than the true functions because covariance between the function and the data is not equal to the covariance of the data itself. A regression estimate is not necessarily a reflection of truth or true behavior. The quest for “truth” arrived in the form of stochastic simulation (e.g., Lantuéjoul 2013; Deutsch and Journel 1998). Instead of one best guess function, the aim is to build realistic models of the real world, where realism is, at least initially, based on modeling statistical properties of the data. In traditional geostatistics, a key statistical property is the semi-variogram, which essentially described the average square difference in value between any two locations in space. A marriage was found between variograms (covariances) and the multivariate Gaussian distribution. The Gaussian assumption was essentially needed because it leads to convenient analytical expressions that involved the variogram. As we now know, much of statistics in the nineteenth and twentieth century was driven by analytical closed form expressions, simply because of lack of computer power. Parsimonious parametric models, such as multi-Gaussian random fields, require estimating just a few parameters. Today, machine learning involves estimating possibly millions of parameters from even more data. A much due makeover in Geostatistics was arriving.

There was a cost to analytical convenience and that cost was physical plausibility and/or truthful behavior. Spatio-temporal models based on variance reflected the (linear) spatial correlation in the data but nothing more. The parsimonious part often led to a fairly narrow formulation of spatial uncertainty. As a result, geostatistical simulation models were often not “believed” by practitioners (Caers and Zhang 2004). This was certainly true in the natural sciences, for example, geology, where practitioners come with a background and prior knowledge about what they believe is truthful about spatial variation. In that context, spatial variation can be distinct and organized in a fashion that cannot be captured with linear models of correlations, such as variograms and Gaussian regressions. Nature is often organized in very specific geometries: forms of low-entropy type of variation and

not high-entropy such as a Gaussian distribution with given covariance.

In the late 1990s, it now became clear that just the measured data of the phenomenon at the study area of interest may not be sufficient and that additional prior information is required (Strebelle 2002; Guardiano and Srivastava 1993). And, alternatively, that so much data is available that parametric models can no longer describe their statistical complexity. The analogy to machine learning is now clearly visible: a model should be trained with data from “something else,” then applied, that is, generalized to the problem at hand. Since geostatistics deals with spatio-temporal data, the training data would need to be of that kind, somewhat exhaustive in space-time. The training image was born. The word “image” here reflects the spatial coherence of such data, just like the covariance function is mathematically consistent. Training “images” can be 4D (space-time). Training images are by definition “positive definite,” simply because they exist in the real world: every single statistic extracted from a training image, such as an exhaustive experimental variogram is, by definition, positive definite. With the advent of the training image, or some very large exhaustive dataset of space-time, came the decrease in the need for parsimonious parametric models. In fact, most algorithms in MPS, but not all, are decidedly nonparametric: the data is driving the model. Because now a large amount of data are available, statistical properties other than the variogram became accessible, simply because of the increase in frequency of observations. MPS algorithms described in the next section rely heavily on this fact. Calculations can now be made of the probability of occurrence at some location, given any values in the neighborhood (space-time), without a probability model, simply by counting. Alternatively, comparisons can be made between some pattern of data in the study area of interest and the patterns in the training image. This comparison, in terms of “distance,” leads to the birth of many MPS algorithm, some of which mimic ideas in computer graphics such as texture synthesis (Mariethoz and Lefebvre 2014).

Applications have always driven development in geostatistics. MPS was driven by such need as well as the explosion of data and computational power. While MPS focused initially on oil and gas applications (a funding reason), it has lately expanded to many fields of the Earth Sciences and beyond. The oil and gas application is worth mentioning because of the direct link between the realistic depiction of geological geometries and realism in the multiphase flow oil/water/gas in the subsurface (Caers 2005; Pyrcz and Deutsch 2014). Real-world applications drove the development of MPS. Current needs are also driven by ease of use. Fitting higher-order parametric statistical models (Mustapha and Dimitrakopoulos 2011) is extremely challenging even for a statistical expert, let alone the thousands of possible

nonexpert users. A certain form of expertise is needed. Current work in science is more going towards artificial intelligence, meaning procedures and processes that allow for relatively easy automation without the presence of the expert. The 1970s expert-based way of doing geostatistics is gradually diminishing simply due to the current technological revolution. MPS algorithms now are evolving to this form of self-tuning (e.g., Meerschman et al. 2013; Baninajar et al. 2019), because they rely on relatively small amounts of tuning parameters that can be learned from examples. In these types of algorithms, there is no need to assess convergence in a Markov chain, or worry about positive definite functions. This allows the domain of application to continue growing as more people discover the richness and convenience that MPS approaches bring to their real world.

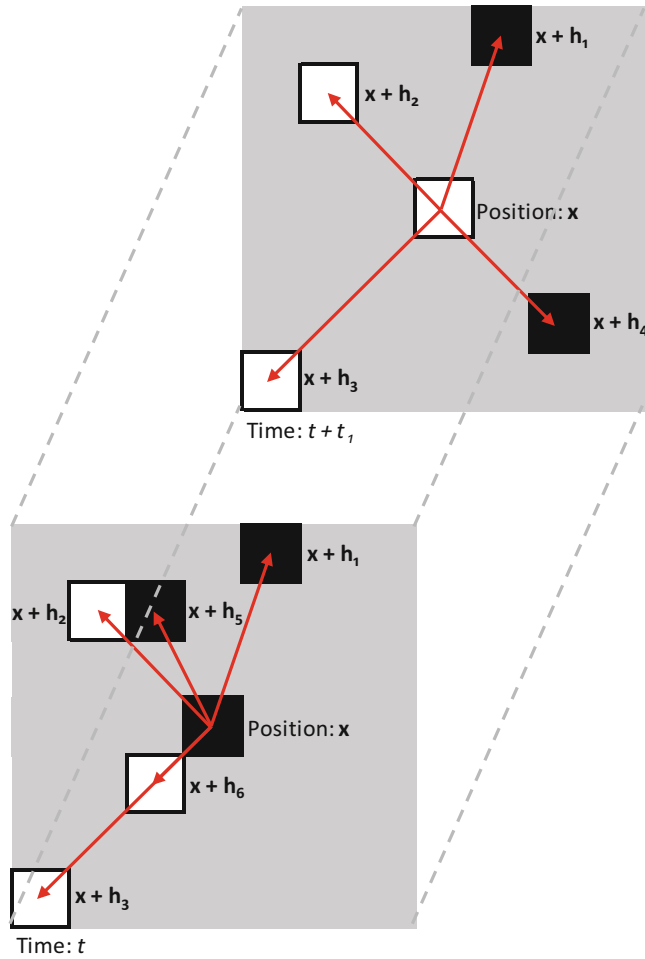
Methods

Simulation methods aim at generating multiple renditions of a random function, honoring conditioning data and reproducing a spatio-temporal structure, as depicted in the training information. In some cases, the goal is one of realism, where the reproduction of exact statistics of pattern configurations may not be a priority. In other cases, the goal is to match as closely as possible the multiple-point pattern frequencies, while respecting conditioning data. In order to solve this problem, various approaches have been developed over the last 30 years, from methods that attempt to solve this analytically, to methods that rely on the heuristics of a specific algorithm to define the higher-order patterns found in the resulting realizations. A review of all the methods is available in the book by Mariethoz and Caers (2015).

In this section, we provide some basic definitions and then dive into a high-level description of different methods and their differences.

Definitions

MPS simulation methods aim at reproducing patterns consistent with some reference information, often a training image, thereby characterizing the spatial (or spatio-temporal) continuity of one or more variables. Patterns are specific configurations of values and locations/time (Fig. 1). Furthermore, in practice, the pattern is composed of conditioning data, which may include previously simulated nodes, and the simulated node or patch, where values need to be assigned. This is referred to as a data event (Fig. 2). In order to inform the simulated node or patch, the training image is used to identify the frequency of values taken by the simulated nodes (red nodes in Fig. 2). An exact match or a similar pattern may be sought after, however, nodes within the simulated patch



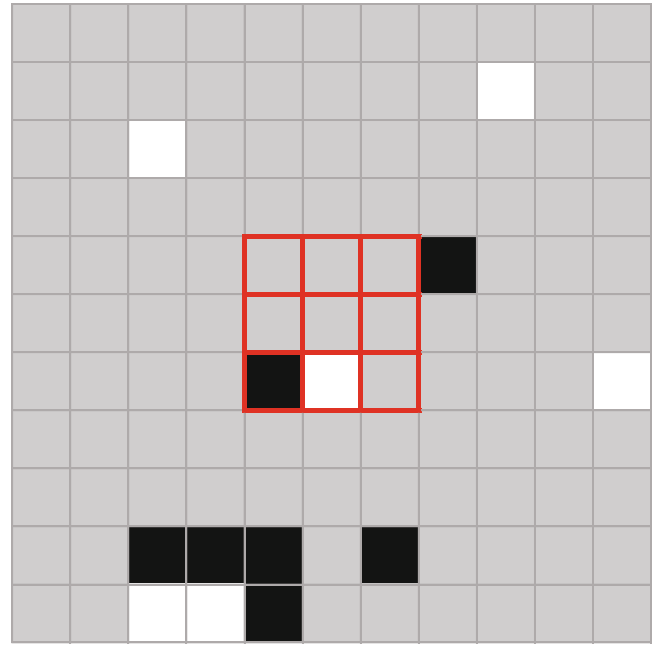
Multiple Point Statistics, Fig. 1 The concept of a pattern: the pattern is defined within its domain (the grey area) by the specific data values taken by one or more variables at multiple locations and times. In the example, values can be “black” or “white,” representing two categories. Locations are defined relative to a specific position in space and time (in this case, the center x , at time t)

should be matched exactly, or at least more closely, as these are key to avoiding artifacts in the simulation result.

The continuity of the spatio-temporal information is inferred from the training data, by recording the occurrence/nonoccurrence of a specific data event, by sampling or scanning the training image. This provides information about possible values the simulated node or patch may take.

Simulation Methods

A large variety of approaches exists to generate simulated models accounting for multiple-point statistics. As discussed in Sect. 1, most of these methods are driven by the algorithm; however, there are a few exceptions where the random function is analytically defined (model-based approaches) and a full description of the mathematical model that characterizes the statistical and spatial frequency of patterns is available. Algorithm-based methods, on the other hand, do not explicitly define the probability model. The conditioning data and



Multiple Point Statistics, Fig. 2 The concept of data event: the data event defines the relationship between the location (in space and time) of the conditioning values and the node or patch (as in this example) to be simulated. In this case, the patch to be simulated consists of 3 by 3 pixels, from which two are already informed. In addition to these two nodes, 11 additional conditioning points are available, some of them from a previously simulated patch (at the bottom of the domain)

frequency of different patterns, inferred from the training information, impose feature in the simulated models, but some features are a direct result of the algorithm.

Model-based Methods

Among model-based methods, Markov Random Fields (MRF) (Tjelmeland and Besag 1998) can be used to reproduce joint and conditional distributions of categorical variables by defining the interactions between pixels in the simulation grid. Reproduction of large-scale continuity can be achieved by imposing high-order interactions (beyond pairwise relationships), over a limited neighborhood. The conditional probability at a specific location $\mathbf{x}_i$ conditioned to the data event, $p(z(\mathbf{x}_i)|N(\mathbf{x}_i))$, is modeled as an analytical function, usually an exponential function (parametric model), from which the spatial law of the random function can be fully defined (up to a normalization constant). Conditional distributions can also be written as an exponential function:

$$p(z(\mathbf{x}_i)|N(\mathbf{x}_i)) \sim \exp\left(-\sum_{c:\mathbf{x}_i \in C} V_c(z_c)\right)$$

where z_c is a clique (a set of connected nodes within the pattern), V_c is a potential function that defines the parametric interactions between cliques. The challenge of MRF is to determine which cliques should be used and how to infer

the parameters θ of the potential functions. This is approached in theory by maximum likelihood; however, in practice it requires Markov chain Monte Carlo sampling, in order to approximate the likelihood function.

Markov mesh models (MMM) are a subclass of MRF that solve some of these implementation issues. They essentially work as a two- or three-dimensional time series. The conditioning path is kept constant, that is, a unilateral path is used, thus simplifying the problem and limiting the clique configurations. Furthermore, the potential functions can be parameterized by using generalized linear models, with parameters inferred by maximum likelihood estimation.

A second family of methods under the model-based category builds upon the concept of moments of the random function. Instead of limiting the inference and simulation to second-order moments as in Gaussian simulation, higher-order moments are considered. These are called spatial cumulants. This allows characterizing random functions with high-order interactions that depart from Gaussianity, thus imposing particular structure and connectivity of patterns. As with traditional variograms (or covariances), inference becomes an issue, thus requiring abundant training data.

Algorithm-based Methods

Many of the algorithm-based methods work under the sequential simulation framework. This means that every new simulated node or patch is conditioned by the hard data and previously simulated nodes within a search neighborhood. The typical sequential framework is summarized in Algorithm 1.

Algorithm 1

The sequential simulation framework

Inputs:

1. Variable(s) to simulate
 2. Simulation entity: pixel or patch
 3. Simulation grid and search neighborhood size
 4. Training data
 5. Conditioning points
 6. Auxiliary variables
 7. Algorithm parameters
-

Algorithm:

1. Assign conditioning data to closest nodes in simulation grid
 2. Visit nodes according to a path
 - 2.1 Search for conditioning data in search neighborhood
 - 2.2 Compute the conditional probability of the variable(s) in the simulation entity conditioned by the data and previously simulated values in the neighborhood (and constrained also with auxiliary variables)
 - 2.3 Draw a value (pixel or patch) from the conditional distribution and assign to location
 3. Postprocess realization
-

Outputs:

1. Simulated grid fully informed, honoring conditioning data and replicating patterns inferred from training data
-

Algorithms vary in the way they handle conditioning to data, inferring of the conditional distribution of node or patch values, and storing frequencies of nodes or patterns conditioned to different data events. The simplest approaches involve simulating a single node, noted pixel-based. Nodes are visited in a random path and at every uninformed location, a search is performed for nearby informed nodes (either conditioning data or previously simulated). A data event is then searched and retrieved from the training data.

Pixel Based Sequential Methods

The simplest method called Extended Normal Equation Simulation (ENESIM) directly scans the training image to determine the frequency with which the uninformed node to be simulated takes different values when conditioned by the data event (Guardiano and Srivastava 1993). This rather inefficient approach of scanning the training image every time can be improved by prescanning the training image and storing the different data events in a data structure for efficient retrieval. The Single Normal Equation Simulation (SNESIM) does this by using a search tree (Strebelle 2002), while the Improved Parallel List Approach method (IMPALA) uses a list (Straubhaar et al. 2011). Direct sampling (DS) replaces the inference of the full conditional distribution by finding a single case and using it as a direct sample of it (Mariethoz and Caers 2015).

Patch Based Sequential Methods

Other approaches seek faster results and a better reproduction of textures, by simulating a complete patch instead of a single node. The process is similar to the one for pixel-based methods. The specific data events as related to the patch to be simulated are inferred from the training image and stored either individually, grouped with a clustering technique or in some kind of data structure.

Stochastic Simulation with Patterns (SIMPAT) was originally conceived to perform this task creating a database of patterns, where the most similar to the data event is searched (using a distance metric). This makes it computationally expensive, which can be solved by clustering patterns that are similar into prototypes. This is done in several methods, where patterns are clustered after applying a dimensionality reduction technique. FILTERSIM uses filters to reduce the dimension of the feature space and then identify the most similar prototype to the data event. Based on this, a patch is drawn from the patterns within a prototype family. DISPAT performs a multidimensionality scaling (MDS) and clustering is done over this reduced space. WAVESIM uses wavelets coefficients for the pattern classification task.

Another variation of SIMPAT is provided by GROWTHSIM that uses a random path that starts at conditioning data locations and grows from there.

Pixel and Patch Based Unilateral Methods

Following ideas from texture synthesis (and similar to MMM), some methods use a unilateral path, instead of proceeding in the traditional sequential fashion with a random path. PATCHWORK simulation computes transition probabilities between patches in the unilateral path from the training image and draws from them depending on the data event, discarding patches that do not honor conditioning data in the nodes of the patch to be simulated. This approach is extended in cross-correlation based simulation (CCSIM), where the difference between patterns is computed rapidly by using a convolution with a cross-correlation distance function (Tahmasebi et al. 2014). Patches are drawn from the most similar set that matches the conditioning data. The patch can be split if no match is found, reducing the size of the simulated patches, thus simplifying the problem. Image Quilting (IQ) extends the CCSIM framework by optimizing the overlapping area between patches, using a minimum boundary cut.

Optimization Based Methods

Simulated annealing (SA) and Parallel Conditional Texture Optimization (PCTO-SIM) are optimization-based approaches (Mariethoz and Caers 2015). Both methods work by starting with the simulation grid filled randomly, except at conditioning locations, where the values are preserved. In SA, an objective function is defined to compute the mismatch between the current state of the simulation and the target statistics. These statistics are related to the frequency of patterns, which are inferred from a training source. A change in the current model is proposed, either by perturbing a node or swapping pairs of nodes, and the objective function is updated. Changes are conditionally accepted following the Boltzmann distribution, that is, favorable changes, those that make the simulation closer to the target, are always accepted, while unfavorable changes are accepted with probability. The probability of this acceptance is reduced as the simulation progresses. In the case of PCTO-SIM, iterations over the initial grid are applied in order to update patches that lead to the minimization of the mismatch function. This is resolved with an expectation-maximization algorithm, which is used to update patches of the simulation. SA can be extremely slow and depending on the definition of the objective function, fail at reproducing realistic representations of the phenomenon at the end of the simulation. More recent methods based on deep neural networks integrate the image reconstruction capabilities of deep learning methods in the spatial context.

Constraining Models to Data

Up to this point, we have discussed the simple case where conditioning data is a hard data, that is, a sample value without uncertainty at the same support of the simulation pixel, and that measures the same variable being simulated.

However, many of the methods described in the previous section can be extended to deal with multivariate data, where both the simulation grid and the training image need to contain multiple correlated variables. In this case, the direct and multivariate distributions are relevant (both statistically and over space/time). The extension to the multivariate case brings inference problems due to the combinatorial nature of the problem. It becomes harder to find specific patterns and every comparison becomes more expensive.

The case of dynamic data related to flow properties and geophysical information bring a different challenge. These variables are related to the simulated variables, but in an indirect manner, often through nonlinear relationships. Adhering to the relationship between the primary variable and the nonlinear information requires a different approach, often done with a Bayesian approach, requiring the definition of a likelihood function. Implementing such an approach requires specifying the prior distribution and then sampling from the posterior distribution according to Bayes' formulation. The prior can be based on training images; however, training-image based priors need to be consistent with data. If there is inconsistency, the sampling process may not converge to the posterior distribution.

Training Images

Requirements for a Training Image

Obtaining adequate training images is one of the most fundamental requirements to be able to use MPS, and at the same time it can be the largest hurdle. MPS simulation methods extract information from the training image and store it in different forms (e.g., a search tree for SNESIM, a list for IMPALA or even keeping the training image in its original image form for DS). Regardless of storage format, the stored training information is thereafter used equivalently as a model in the simulation process. Therefore, whichever simulation algorithm is chosen, it can only perform as well as the quality of the training image it is provided. This begs the question of defining what a good training image is. We define below three essential criteria, namely: stationarity, diversity, data compatibility.

A first requirement is that the training image should be stationary. This means that all areas of the training image represent similar statistical characteristics, otherwise it would be virtually impossible to calculate meaningful higher order frequencies or conditional probabilities. Stationarity in the traditional geostatistical sense, such as second order intrinsic stationarity, is now extended to higher order statistics. While there are ways of using nonstationary training images, those work by segmenting the nonstationary training image into several subdomains that are themselves stationary. This segmentation can take the form of well-defined zones

(de Vries et al. 2009) or continuously varying features that are described with an auxiliary variable, which defines correspondences between areas in the training image and in the simulation (Chugunova and Hu 2008). However, in all cases each segmented zone should be itself stationary, meaning that it must present a degree of recurrence of similar statistical characteristics.

The second requirement is that the training image should be diverse, that is, large enough to capture the properties of interest. But here the term “large” should not be taken in terms of size (e.g., number of pixels or voxels), but in terms of information content. Specifically, one can use a relatively small training image if the goal is to simulate simple patterns. But for complex and diverse structures, the training image needs to represent the entire range of features and their variability. Essentially, the features of the training image have to be replicated sufficiently in order to represent their diversity. If those features are large, the training image has to be large enough to contain a sufficient amount of them. One way of increasing the diversity of patterns in a training image is to use transform-invariant distances (Mariethoz and Kelly 2011), which allow considering not only the patterns present in a training image, but also the ensemble of their transformations (e.g., rotations, affinity transforms).

Finally, the third requirement is that a training image should be compatible with data are available for the simulated domain. Most practical cases involve conditional simulation, and therefore, such data exist. It can be, for example, scattered point measurements in the case of simulating rainfall intensity based on in situ rain gauge measurements, line data if simulating a 3D geological reservoir with information from wells, or it can be a known secondary variable which has a given relationship with the variable to be simulated, such as in geophysics. In such cases, the statistical properties of the simulated variable should correspond to the statistics of the data, and here again such requirement goes beyond mean and

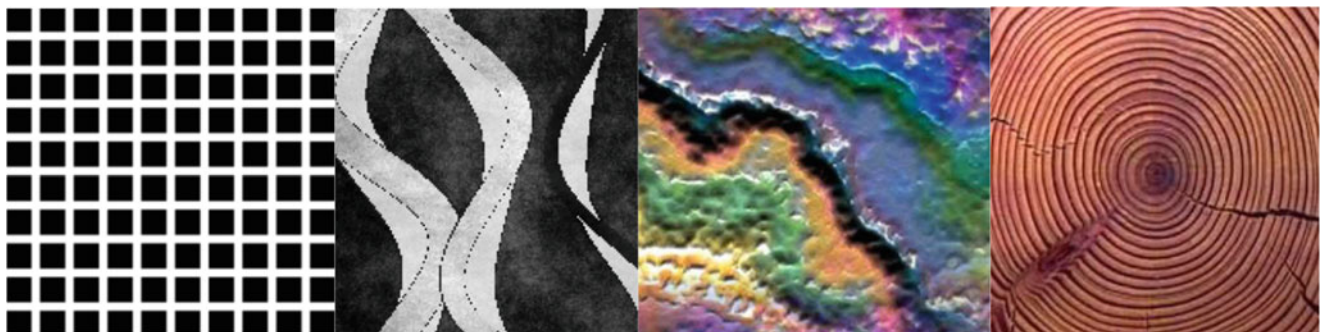
variance, but can include higher-order properties such as connectivity or the frequency of patterns. Validating a training image for data compatibility can be achieved in several ways. Often, one would compute the statistical properties of interest on the dataset and on a candidate training image. This can take the form of variograms (for point data), connectivity functions (if the data are spatially continuous, a requirement for computing connectivity metrics), or patterns frequencies (using specific patterns comparison algorithms that assess the frequency of spatial patterns in both training image and data). Examples of pattern comparison algorithms are given in Boisvert et al. (2010), Pérez et al. (2014), and Abdollahifard et al. (2019).

To illustrate our stated requirements, Fig. 3 provides some examples of inadequate training images, which are individually commented in the figure caption.

Data-Based Training Images

There are two main ways of obtaining adequate training images for a given simulation problem. The first one is in situations where large amounts of spatially registered data are available, such that the data themselves can be used as training image. Typical data-rich applications where data-based training images are common include geophysics, remote sensing, or time series analysis. Such cases can involve processes that repeat themselves in space (e.g., geophysical cross-sections across a valley, where measured transects are assumed to be statistically similar to unmeasured transects) or repetitions in time (e.g., repeated acquisitions of remote sensing imagery, where previous acquisitions are similar to acquisitions of the same area that have not yet taken place), or even a combination of both (e.g., a spatial network of weather stations that produce several meteorological time-series that show recurrent patterns in both space and time).

One issue with data-driven training image is that direct observation of natural processes often does not correspond to



Multiple Point Statistics, Fig. 3 Example of inadequate training images. From left to right: (a) A very repetitive training image that is overly redundant and could be reduced to a fraction of its size without loss of information. (b) A nonstationary training image where features on the right and left are fundamentally different. Such an image could however be used by defining stationary subdomains. (c) A nonstationary

image with a continuous gradation of colors and orientations implying almost repetitivity. (d) Circular structures resulting in nonstationarity. May be usable with an auxiliary variable that defines where each feature should be mapped in the simulation, or by using rotation-invariant distances

the requirements outlined above. While pattern diversity can be attainable if large amounts of data are at hand (e.g., with Earth observation satellites that generate daily terabytes of spatialized imagery), stationarity and data compatibility are often more problematic.

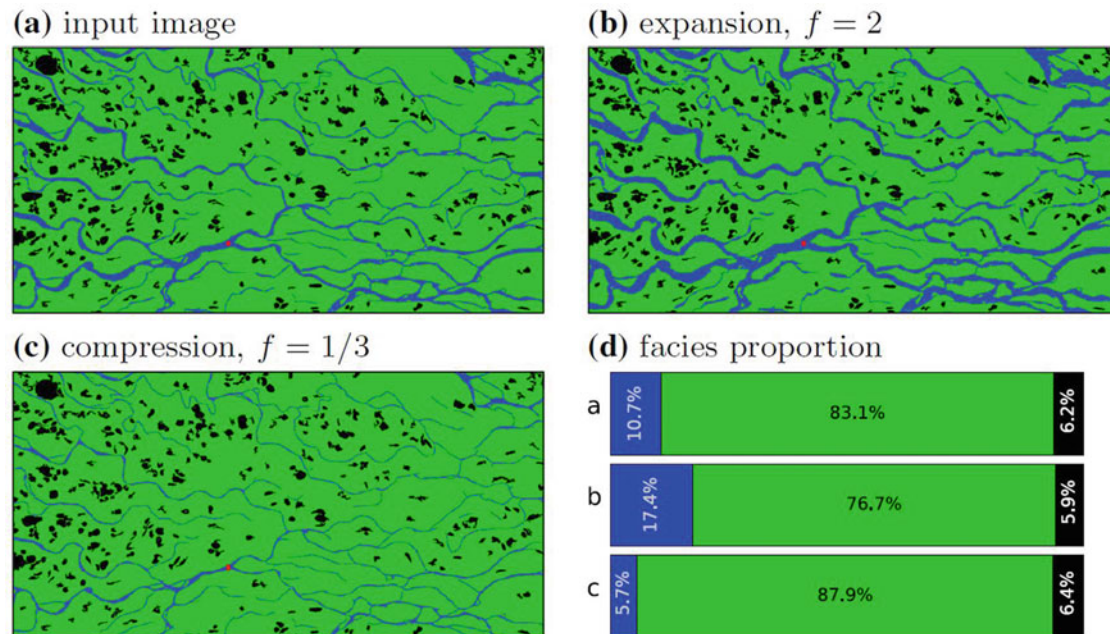
Natural processes are rarely stationary, often requiring detrending techniques to make them amenable for use as training images. The detrending can take many forms, depending on the type of nonstationarity to be addressed. One solution for large-scale trends is to separate the training image in two components: a deterministic trend and a residual, much like the traditional way of doing geostatistics. Under the assumption that such a decomposition is sufficient to describe the trend in the data, the residual component is then considered as stationary and used as training image (Rasera et al. 2020). Such a decomposition is common in time series simulation, where the trend can, for example, describe seasonal fluctuations in the mean and in the standard deviation of the data of a variable.

Data compatibility can also be an issue when using data-based training images. The typical recommendation is to find data that are representative of the modeled process; however, this can be difficult in practice. The training image generally does not have exactly the same statistical properties as the data. In many cases, this can be resolved by applying a transformation to the data, such as a histogram transformation, which results in the training image having the same distribution as the available data. Sometimes however the data transformation needs to be more complex. For example,

one may want to simulate the spatial features of a deltaic channelized river, but the only available training image has been acquired in high-water conditions and therefore the channels are broader than desired. In this case, a morphological transformation of the training image may be more appropriate, for example, by applying erosion or dilation operators. Interactive image transformation techniques have been proposed to either alter the proportion of categories in an image or “warping” an image to change its morphological characteristics while preserving the overall topology (Straubhaar et al. 2019). This is illustrated in Fig. 4.

Construction of Training Images

In many applications, the data are too scarce to provide data-driven training images. A typical case is the modeling of subsurface reservoirs, where the only data are often borehole cores which provide lithological information along 1D lines that represent only a very small proportion of the entire reservoir 3D volume. In these cases, using MPS requires to build a training image based on whatever information is available. The common approaches to achieve this often involve using other geostatistical methods (i.e., other than MPS) that can produce a realistic representation of the phenomenon to model. Several geostatistical methods are less data-hungry than MPS and still able to depict a high level of complexity. For example, object-based approaches have been commonly used to create 3D training images in reservoir modeling.



Multiple Point Statistics, Fig. 4 Illustration of image warping. (a) input image, (b) and (c) image after expansion and compression, (d) proportion of each category. (Figure from (Straubhaar et al. 2019))

Another way of constructing training images is by using process-based models. These models use physical laws to model the process of interest, for example, sediment transport (Sun et al. 2001) or terrain erosion (Braun and Willett 2013) or fluid flow in fractures (Boisvert et al. 2008). This is generally done by solving computationally demanding systems of partial differential equations. Process-based models are often deterministic, and in any case so expensive that it is often not possible to obtain more than one or a handful set of realizations, let alone to condition them to available data. MPS is then the ideal tool to use these few models as training images and expand them into a larger set of realizations that are statistically representative and moreover conditioned (Comunian et al. 2014; Comunian et al. 2014; Jha et al. 2015).

Applications

As in Part I, we will provide here more of a historical timeline of applications, partly to illustrate how MPS grew and how more applications entered the scene, how diversity of algorithms became linked to diversity of applications. Without doubt, the primary application of MPS was in reservoir modeling (Oil & Gas; Caers et al. 2003; Caers and Zhang 2004). Reservoir modeling (Pyrz and Deutsch 2014) requires building multiple realizations of several important properties such as lithology, porosity, and permeability. The purpose of these models is either appraisal or production planning. The latter requires simulation of multiphase flow. Flow simulation is sensitive not just to the magnitude in permeability but to the contrast between high and low permeability. This contrast, in the subsurface, offers itself in specific geological shapes, such as lithological or architectural elements, or due to diagenesis. The connectivity of high (or low) permeability is poorly characterized by Gaussian distributions with given covariance (the high entropy model). Often, multiple-point geostatistics offered a compromise between Boolean and pixel-based methods. Boolean methods are more difficult to constrain to data; hence, they can serve as training images once the objects are pixelized on a grid. In terms of approach, oil and gas application very much followed the “constructed” training image approach (see sect. 3). Well data (considered hard data) are scarce, yet geological interpretation based on them (as well as seismic) are very specific. Uncertainty in such interpretation was handled by proposing multiple training image. Validation of training images was often needed when integrating production data. Specific methods for training-image based inversion of flow data were developed (Caers 2003). To date, a range of algorithms are used in actual applications: SNESIM, DS, IQ. Like discussed in the previous section, there was initially confusion around the need for a stationary training. Model based on dense outcrop data or generated using

process methods are by very definition nonstationary. Either stationary training images are used in combination with seismic data, which offers a nonstationary trend in petrophysical properties (Caers et al. 2001), nonstationary images are accompanied by auxiliary variables (Chugunova and Hu 2008) (Fig. 5).

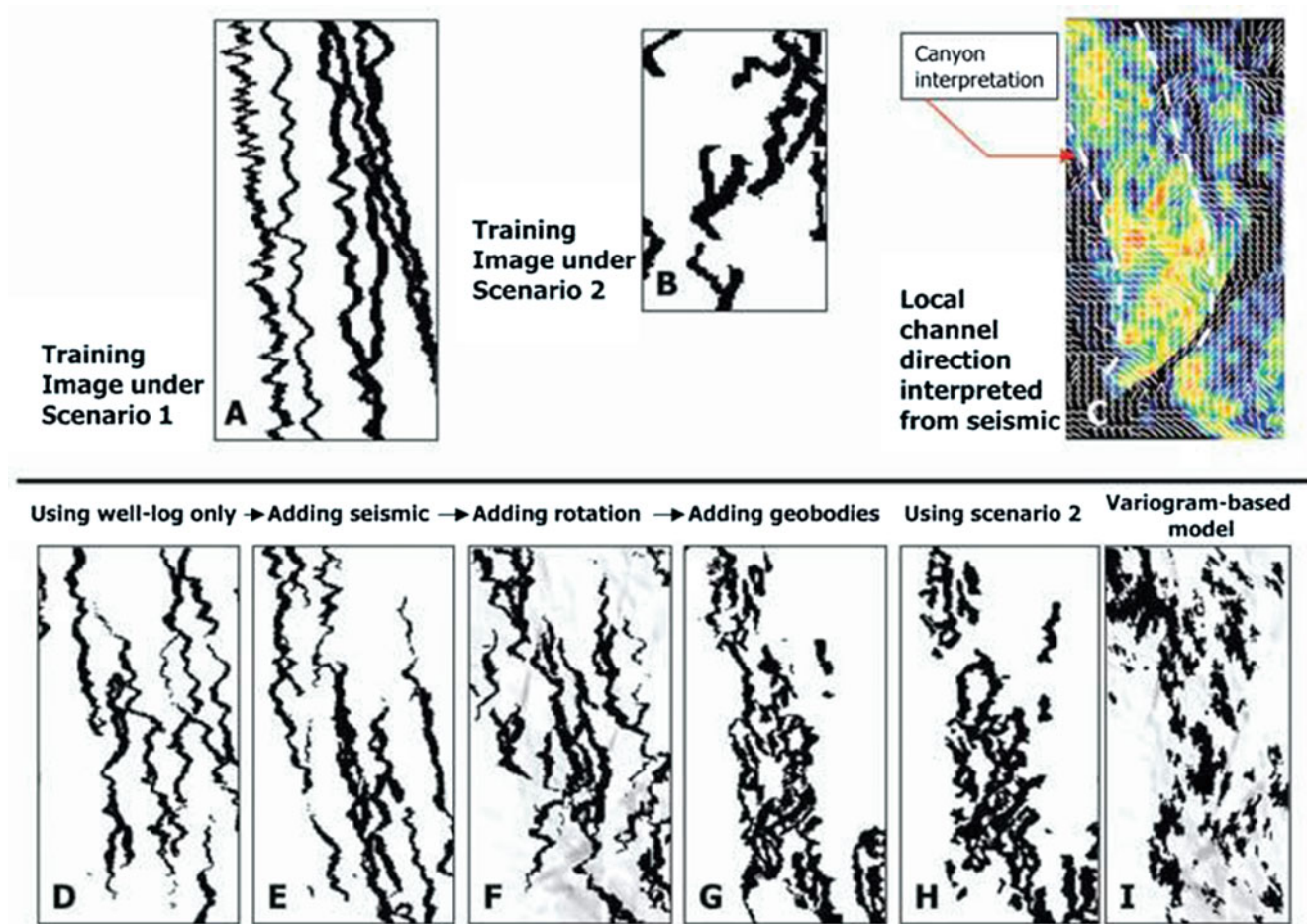
Groundwater modeling offers similar challenges compared to oil and gas, namely, in the realistic representation of subsurface heterogeneity. A more general similarity emerges, namely, these are problems with very few specific data (hard data, wells) and very specific information on variability of spatial geometries. A second type of application was then emerging which presents the exact opposite problem: very dense data.

With very dense data, we enter the realm of geophysical and remote sensing data. Such data sources are easier to get exhaustively than drilling wells and require either passive monitoring (remote sensing, gravity) or active source monitoring (electrical resistivity, seismic). Two types of problems occur in which MPS has started to play a role: data gap filling and downscaling.

Satellite data may present gaps because of cloud cover and orbital characteristics. Traditional approaches by filling gaps using kriging (Cressie et al. 2010) poses many problems. First, kriging provides only a smooth representation of actual data variability. Second, a computational problem occurs because of the need to invert very large covariance matrices. MPS provides a natural solution in that it uses the data itself to derive statistics to fill the gaps. Here, the neighboring data is the “training image,” although there is no explicit construction of such image. Gap filling is also needed with (subsurface) geophysical data. Often such data are collected along acquisition lines with certain line spacing. These linespaced data then need to be gap-filled to get a 2D or 3D realistic representation of the subsurface. Most geophysical inversion methods smooth the line data.

Downscaling constitutes a set of problems whereby coarse scale data or models are transposed to fine-scale resolution. This problem is not uniquely defined: given an upscaling process, one can devise many downscaled models from a single coarse model. Downscaling can be addressed in two ways. First, one may have areas (or space-time blocks) with both high-resolution and low-resolution data, which can be used as training data. Alternatively, one may have constructed, using physical models, high-resolution models that are then upscaled (another physical process) into low-resolution versions. An example presented here is the downscaling of topographic data in Antarctica.

Other applications worth mentioning are in mining. Mining is characterized by the availability of dense drill-hole data; hence, training images based on dense data can be constructed, then used in sparser drilled areas. MPS has also



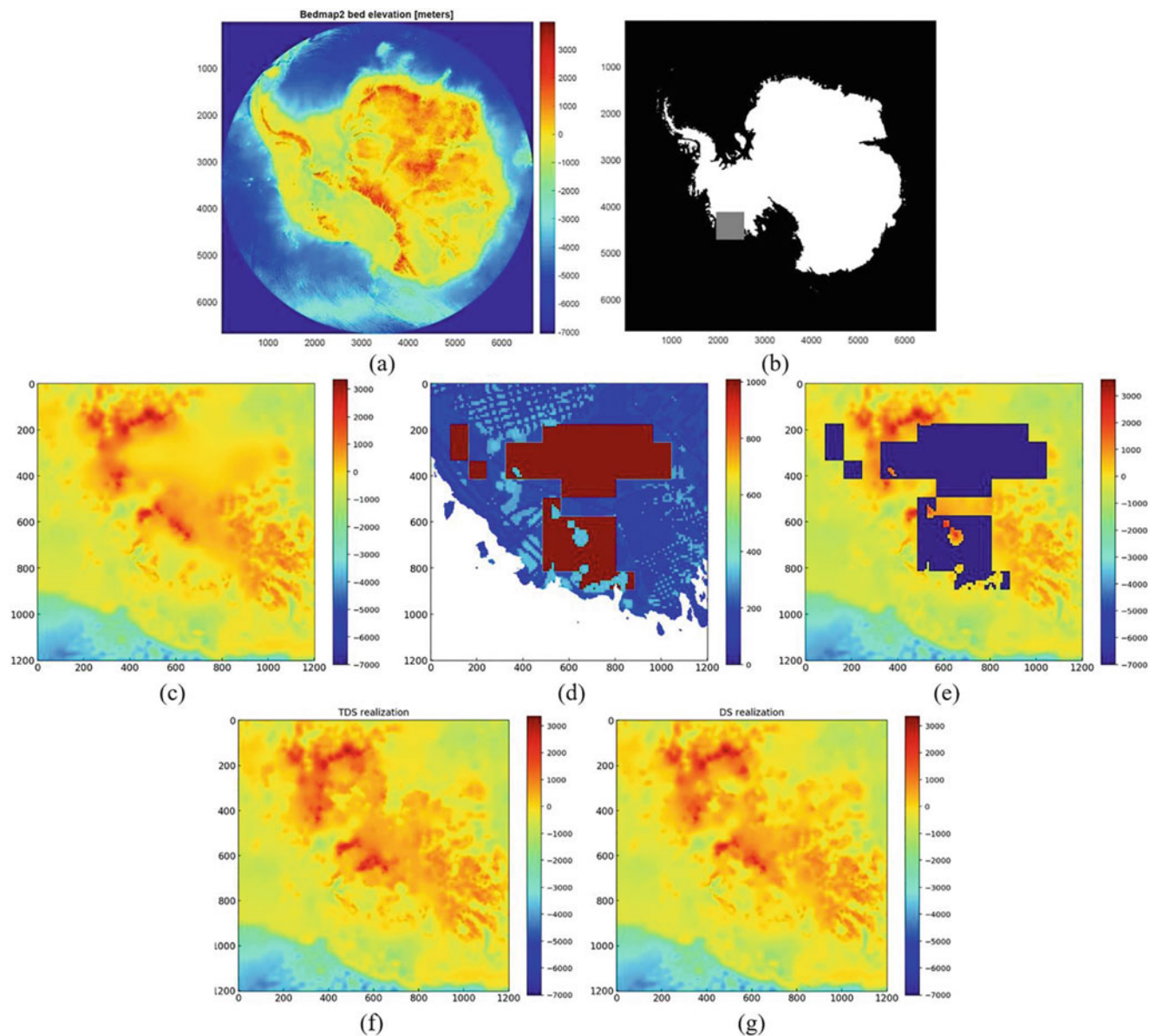
Multiple Point Statistics, Fig. 5 One of the first real-world application of MPS, generation of a reservoir model using a training image, constrained to well and probability maps deduced from seismic. (From Caers et al. 2003)

applications in temporal modeling such as the modeling of rainfall time-series (Fig. 6).

Summary and Conclusion

We presented a historical overview of multiple-point geostatistics (MPS) as well as some ongoing trends in this area. Fundamental to multiple-point geostatistics is recognizing that spatial covariance or variogram-based methods fail to capture complex spatio-temporal variation. Training images, either as constructed formats or as direct real-world observations, are explicit expressions of such variations. The aim of MPS algorithms is to capture and reproduce essential higher-order statistics from training images, at the same time, conditioning to a variety of data. Algorithms can be divided into two groups: model-based or data-driven. Model-based algorithms rely on the traditional probability theory framework while data-driven method share ideas with machine learning

and computer graphics. Training images are required to have certain characteristics to be useable with the defined MPS framework: stationarity, diversity, and data compatibility. Stationarity is required to be able to borrow meaningful statistical properties. Diversity is needed to have enough quantity of such meaningful statistical information and compatibility means that the statistical properties cannot contradict with actual conditioning data. The application of MPS methods and ideas continues expanding. From the initial subsurface focus, applications have grown to address the need of today's societal challenges: the Earth's landscape, climate, and the environment. The fundamental idea of MPS and the shift it brought to geostatistics will be permanent. The algorithms, data, and scope of applications will continue to grow, for example, more recent publications have started to focus on generative models that replace training images with neural networks. An example is the Generative Adversarial Network (GAN) that can be trained once and possibly ported to a variety of application with minimal modification.



Multiple Point Statistics, Fig. 6 Gap-filling the digital elevation in Antarctica. (a) Bed elevation model of Antarctica; (b) study area; (c) the kriging realization; and (d) kriging (e) training data (f–g) Direct sampling realizations. (From Zuo et al. 2020)

Cross-References

- [Geostatistics](#)
- [High-Order Spatial Stochastic Models](#)
- [Markov Random Fields](#)
- [Pattern](#)
- [Pattern Analysis](#)
- [Pattern classification](#)
- [Pattern recognition](#)
- [Simulation](#)
- [Spatial statistics](#)
- [Spatiotemporal](#)

Bibliography

- Abdollahifard MJ, Baharvand M, Mariéthoz G (2019) Efficient training image selection for multiple-point geostatistics via analysis of contours. *Comput Geosci* 128:41–50
- Baninajar E, Sharghi Y, Mariéthoz G (2019) MPS-APO: a rapid and automatic parameter optimizer for multiple-point geostatistics. *Stoch Env Res Risk A* 33(11):1969–1989. <https://doi.org/10.1007/s00477-019-01742-7>
- Boisvert JB, Leuangthong O, Ortiz JM, Deutsch CV (2008) A methodology to construct training images for vein-type deposits. *Comput Geosci* 34(5):491–502
- Boisvert JB, Pyrcz MJ, Deutsch CV (2010) Multiple point metrics to assess categorical variable models. *Nat Resour Res* 19(3):165–175

- Braun J, Willett SD (2013) A very efficient O(n), implicit and parallel method to solve the stream power equation governing fluvial incision and landscape evolution. *Geomorphology* 180–181:170–179
- Caers J (2003) History matching under a training image-based geological model constraint. *SPE J* 8(3):218–226, SPE # 74716
- Caers J (2005) *Petroleum Geostatistics*. Society of Petroleum Engineers, Richardson, 88 pp
- Caers J, Zhang T (2004) Multiple-point geostatistics: a quantitative vehicle for integrating geologic analogs into multiple reservoir models. In: Grammer GM, Harris PM, Eberli GP (eds) *Integration of outcrop and modern analog data in reservoir models*, AAPG memoir 80. American Association of Petroleum Geologists, Tulsa, pp 383–394
- Caers J, Avseth P, Mukerji T (2001) Geostatistical integration of rock physics, seismic amplitudes, and geologic models in North Sea turbidite systems. *Lead Edge* 20(3):308–312
- Caers J, Strebel S, Payrazyan K (2003) Stochastic integration of seismic data and geologic scenarios: a West Africa submarine channel saga. *Lead Edge* 22(3):192–196
- Chugunova T, Hu L (2008) Multiple-point simulations constrained by continuous auxiliary data. *Math Geosci* 40(2):133–146
- Comunian A, Jha S, Giambastiani B, Mariethoz G, Kelly B (2014) Training images from process-imitating methods. *Math Geosci* 46(2):241–260
- Cressie N, Shi T, Kang EL (2010) Fixed rank filtering for spatio-temporal data. *J Comput Graph Stat* 19(3):724–745
- de Vries L, Carrera J, Falivene O, Gratacos O, Slooten L (2009) Application of multiple point Geostatistics to non-stationary images. *Math Geosci* 41(1):29–42
- Deutsch CV, Journel AG (1998) *GSLIB: geostatistical software library and user's guide*. Oxford university press, New York
- Guardiano F, Srivastava M (1993) Multivariate geostatistics: beyond bivariate moments. In: Soares A (ed) *Geostatistics-Troia*. Kluwer Academic, Dordrecht, pp 133–144
- Härdle WK, Simar L (2015) *Applied multivariate statistical analysis*. Springer, Berlin/Heidelberg
- Jha S, Mariethoz G, Mathews G, Vial J, Kelly B (2015) A sensitivity analysis on the influence of spatial morphology on effective vertical hydraulic conductivity in channel-type formations. *Groundwater*
- Lantuéjoul C (2013) *Geostatistical simulation: models and algorithms*. Springer Science & Business Media. Springer, Berlin, Heidelberg
- Mariethoz G, Caers J (2015) *Multiple-point Geostatistics: stochastic modeling with training images*. Wiley, Hoboken, 374 pp
- Mariethoz G, Kelly BF (2011) Modeling complex geological structures with elementary training images and transform-invariant distances. *Water Resour Res* 47(7):1–14
- Mariethoz G, Lefebvre S (2014) Bridges between multiple-point geostatistics and texture synthesis: review and guidelines for future research. *Comput Geosci* 66(0):66–80
- Matheron G (1963) Principles of geostatistics. *Econ Geol* 58:1246–1266
- Matheron G (1973) The intrinsic random functions and their applications. *Adv Appl Probab* 5(3):439–468
- Meerschman E, Pirot G, Mariethoz G, Straubhaar J, Van Meirvenne M, Renard P (2013) A practical guide to performing multiple-point statistical simulations with the direct sampling algorithm. *Comput Geosci* 52:307–324
- Mustapha H, Dimitrakopoulos R (2011) HOSIM: a high-order stochastic simulation algorithm for generating three-dimensional complex geological patterns. *Comput Geosci* 37(9):1242–1253
- Pérez C, Mariethoz G, Ortiz JM (2014) Verifying the high-order consistency of training images with data for multiple-point geostatistics. *Comput Geosci* 70:190–205
- Pyrcz MJ, Deutsch CV (2014) *Geostatistical reservoir modeling*. Oxford University Press, New York
- Rasera LG, Gravey M, Lane SN, Mariethoz G (2020) Downscaling images with trends using multiple-point statistics simulation: an application to digital elevation models. *Math Geosci* 52(2):145–187
- Rasmussen CE, Williams CKI (2006) *Gaussian processes for machine learning*. The MIT Press, Cambridge, MA
- Straubhaar J, Renard P, Mariethoz G, Froidevaux R, Besson O (2011) An improved parallel multiple-point algorithm using a list approach. *Math Geosci* 43:305–328
- Straubhaar J, Renard P, Mariethoz G, Chugunova T, Biver P (2019) Fast and interactive editing tools for spatial models. *Math Geosci* 51(1):109–125
- Strebel S (2002) Conditional simulation of complex geological structures using multiple-point statistics. *Geology* 34(1):1–22
- Sun T, Meakin P, Jøssang T (2001) A computer model for meandering rivers with multiple bed load sediment sizes 2. *Computer simulations*. *Water Resour Res* 37(8):2243–2258
- Tahmasebi P, Sahimi M, Caers J (2014) MS-CCSIM: accelerating pattern-based geostatistical simulation of categorical variables using a multi-scale search in Fourier space. *Comput Geosci* 67(0):75–88
- Tjelmeland H, Besag J (1998) Markov random fields with higher-order interactions. *Scand J Stat* 25:415–433
- Zuo C, Yin Z, Pan Z, MacKie EJ, Caers J (2020) A tree-based direct sampling method for stochastic surface and subsurface hydrological modeling. *Water Res Res* 56(2). p.e2019WR026130

Multiscaling

Jaya Sreevalsan-Nair

Graphics-Visualization-Computing Lab, International
Institute of Information Technology Bangalore, Electronic
City, Bangalore, Karnataka, India

Definition

In several natural resource and environmental applications, e.g., mining, there exists a high degree of uncertainty in the information in the scenario (Sagar et al. 2018). This uncertainty can be addressed using *scale-dependent* phenomena. *Scale* in simple terms refers to the scope of observations in the respective physical dimensions, such as spatial length (Euclidean space) and time, or related spatial dimensions, such as image space. Multiple geological features occur at different scales in time and space, namely tectonic objects (faults, joints, etc.), sedimentary objects (bedding structures, etc.), intrusive and effusive objects (lava flows, dykes, etc.), and epigenetic objects (metamorphic units, etc.). They range in spatial scales of micrometers to kilometers, and the geophysical processes range in time scales of microseconds to millions of years.

Multiscaling is used to capture information at different scales where specific type of *locality* can be defined, e.g., spatial locality or temporal locality. This is especially necessary to faithfully model multiresolution data, which are

generated or acquired at multiple resolutions, by design. An illustrative example of multiscaling is in determining instantaneous friction and predicting time to failure in tectonic fault physics of slow earthquakes at the laboratory scale, and at the field scale (Bergen et al. 2019). The fault generated in the laboratory is effectively a single frictional patch in a homogeneous system, and a field fault in the earth is an ensemble of several frictional patches, occurring in heterogeneous earth materials.

Overview

The use of scale for studying physical models and simulations, and for analyzing data has been well known. However, this practice also contrasts with the ideal formulation of *scale-independent* entities in mathematics, e.g., “point” (Lindeberg 1998). In modern data scientific applications, scale-space theory has been extended to nonconventional spatial regions, e.g., images. Such applications are important in the domain, as multiscaling is applicable for modeling and simulation of geoscientific processes, and for data analysis for observed as well as simulated data of all types, including images. Multiscaling in image processing and computer vision is used to address scale-dependent variations in perspective, and noise generated during the image formation process.

Multiple scales naturally organize data in hierarchical structures, from coarsest to finest scales, i.e., from lowest to highest resolution or scope of information, presenting data in the form of a pyramid. Given the presentation of variability in the data, multiscaling also becomes a process by which uncertainty in the data and/or the data model can be alleviated. Uncertainty in data is owing to variability in the process that is being observed, which causes insufficiency in the sampling and/or quality of data. This type of uncertainty is called *aleatoric* or irreducible uncertainty. The other type of uncertainty stemming from the data model used owing to incomplete knowledge of the phenomenon is called *epistemic* or reducible uncertainty. Epistemic uncertainty may be reduced by adding more data. Generally, the aleatoric uncertainty tends to be scale dependent (Sagar et al. 2018), and multiscaling may be seen as a powerful tool to resolve this type of uncertainty.

The widely used multiscaling methods define scale as a contextually appropriate parameter, and focus on the best method of aggregating the information acquired at different scales. For LiDAR (light detection and ranging) point cloud analysis, local neighborhood can be a spherical one (Demantké et al. 2011) or k -nearest neighbors (Hackel et al. 2016), in which case the radius of the sphere or the integral neighbor count k is taken as the scale, respectively. For feature extraction applications, the multiscale aggregation methods include using information at an optimal scale (Demantké et al.

2011), concatenating information from all scales as a long vector (Hackel et al. 2016), or performing an appropriate *map-reduce* operation on the information from the different scales, e.g., averaging (Sreevalsan-Nair and Kumari 2017). Thus, the multiscaling approaches perform filtering, concatenation, or transformations, respectively.

Applications

There exist several different multiscaling applications in geosciences. Instead of an exhaustive survey, illustrative examples of multiscaling in spatial length, time, image space, and point clouds are discussed here.

Transfer functions are mathematical functions used to navigate the hierarchical structures of scales, i.e., for upscaling or for downscaling. Consider the example of transfer functions used in hydrogeology, which is a combined discipline of studying soil science (pedology) and hydrology to understand their interactions in the critical zone of the earth (Lin 2003). This involves studies at microscopic scale of length ($1\ \mu\text{m} - 1\ \text{cm}$), mesoscopic scale ($1\ \text{m}$), and macroscopic scale ($1\ \text{km}$). The microscopic structures include pores and aggregates, mesoscopic ones include pedons and catenas, and macroscopic ones include watersheds and regional and global regions. These structures also reflect varied temporal structures, going from minutes-days, days-months, to months-years, respectively. Given that the natural physical processes involve multiple scales, the process modeling, simulation, and data analysis also tend to involve multiscale bridging in the discipline. The different scales are incorporated in hierarchical structures of soil mapping and soil modeling, and these are used in generating different pedotransfer functions (PTF). PTFs are mathematical functions or models which map the soil survey database (the domain) to the soil hydraulic properties (the co-domain), and are formulated in a nested fashion, using the scales. The goal of the hydrogeological discipline in itself is to use these PTFs and other similar methods to connect the pedon-scale measurements (micro/mesoscopic) to the landscape-scale phenomena (macroscopic) in a three-way interconnection between soil structure, preferential flow, and water quality. Similarly, transfer functions have been used in a range of temporal scales of recharge and groundwater level fluctuations as input and output, respectively, to fractured crystalline rock aquifers, to quantify the aquifer recharge. Hyperspectral remote sensing data of different spatial and temporal scales has been used to monitor vegetation at laboratory (plot), field (local), and landscape (regional) levels. Different geometric resolutions in these images provide heterogeneity of vegetation indices, where the information between scales is harmonized by using transfer functions.

Coupling across different scales to communicate information is another one of the multiscaling methods. Consider the instance of interdisciplinary study of earth sciences involving interconnections between thermal, hydrous, mechanical, and chemical (THMC) behavior of the earth (Regenauer-Lieb et al. 2013). There exists a multiphysics model of THMC coupling across different scales, in both spatial length and time, which enables communication needed for the transition between scales. As an example, for earth resource processes, the spatial scales range from nanometers to 100 kilometers and the time scales range from femtoseconds to $1e+8$ years. These are the scales involved for energy and mineral resource discovery, involving a geodynamic framework. One of the unifying concepts across scales is where the macroscale of one process is used as the microscale of a process at a higher level, thus facilitating multiscaling. Statistical mechanics and thermodynamics offer tools to solve such problems, e.g., macroscopic thermodynamic properties of a continuous medium are derived using stochastic modeling of multiple microscale interactions. The poor availability of order parameters, microscale intrinsic length scales, and energy-based interactions is a roadblock for using the concept as is. To overcome this, the larger scale entities use inputs computed using percolation renormalization at the microscopic scale through coarse graining, and an asymptotic homogenization at the macroscopic scale. Percolation theory has been conventionally used for determining global behavior, e.g., phase transitions, in a random heterogeneous system, e.g., randomly structured media.

It is widely known that earth observations and analysis are sensitive to the scale of data acquisition (Zhao 2001). Thus, an important application of object detection in very-high-resolution imagery entails estimation of scale parameters at different resolutions to segment the images appropriately. Segmentation provides the geometric representation of the spatial objects present in the region captured in the images. The scale parameter depends on local variance of the object heterogeneity in the scene as well as the rate of change of the local variance across consecutive scales. This is because the structural information of the region is derived from the pixel information, which encodes the object heterogeneity. This structural information is effectively obtained from the features extracted from the remotely sensed images. In addition to the hierarchical organization of scales and navigation using transfer functions, scale-space filtering is a commonly used method in feature extraction using multiscaling (Lindeberg 1998). The scale-space theory states that the smaller-scale structures disappear sooner than the larger-scale ones, when the scale-space representation is defined using smoothing. The smoothing is done using Gaussian or wavelet filters. The smaller-scale structures are also perceived as noise, thus, using the scale-space representation for denoising images or filtering noise from the images. This also works

on the principle of *persistence* of *significant features* across multiple scales. An example application for feature extraction using multiscaling is for land cover-based image classification of Landsat images.

Case Study: Feature Extraction in LiDAR Point Clouds

Consider a deeper study on how multiscaling is used for feature extraction in point clouds acquired using airborne LiDAR or terrestrial LiDAR technology. Use of multiple scales in point cloud analysis is a well-established practice to address the uncertainty in environmental datasets. These features are further used for semantic classification, i.e., object-based classification of the point clouds. The classification is an important data analytic step in point cloud analysis. In point clouds, spatial locality is exploited using the neighborhood defined by the scale parameter. The local neighborhood itself can be determined using different approaches. The local neighborhood of each point in the point cloud has been defined in the state of the art using either the limiting shape of the neighborhood and/or the size of the neighborhood set. A neighbor of a point p is defined as a point which satisfies the distance criterion defining the local neighborhood of p . The different kinds of local neighborhoods generally used for point cloud analysis are spherical, cylindrical, cubical, and k -nearest neighborhoods. The distance used is usually the Euclidean distance or l_2 norm, and in certain applications, its limiting case of Chebyshev distance or max-norm is used. The size of the local neighborhood is, then, used as the scale parameter. For instance, the radius of the spherical or cylindrical neighborhood is the scale parameter, so is the length of side in the cubical neighborhood, and the value of k in the k -nearest neighborhoods.

Conventional workflow involves feature extraction at each scale for a set of consecutive scales, with lower and upper bounds (s_{min} and s_{max}), and the step size (Δs) of the scale parameter defined for each application. Determining the scales for effective feature extraction is in itself a difficult problem. Applications which use an optimal scale s_{opt} for the final feature vector are not as critically dependent on the choice of $[s_{min}, s_{max}, \Delta s]$ as the applications where feature vectors at all scales are used for further data processing.

Before discussing how features across different scales are aggregated meaningfully for any application, the feature extraction at each scale must be discussed. For a given point and its selected neighborhood, the local geometric descriptor is computed. A widely used local geometric descriptor is the covariance matrix or tensor, C . It is a positive semidefinite second-order tensor (Sreevalsan-Nair and Kumari 2017), computed as the sum of outer product of distance vector between the point and each of its local neighbor. Tensor

voting is an alternative for computing local geometric descriptors, as it generates positive semidefinite second-order tensor fields. The nonnegative eigenvalues of C are used as features, upon its eigenvalue decomposition. There is also a variant of C , where the distances of the neighbors are computed with respect to the centroid of the local neighborhood, instead of the concerned point itself. This is effectively equivalent to performing a principal component analysis of the local neighborhood of the point. Thus, the eigenvalue-based features essentially capture the shape of the local neighborhood. The shape of the local neighborhood in itself gives the geometric classification of the concerned point to a linear, areal/surface, or volumetric feature (Sreevalsan-Nair and Kumari 2017). The eigenvalue-based features provide a probabilistic geometric classification, as these features are scale dependent. Overall, the eigenvalue-based features of a point are essential in interpreting the point in its spatial context. For semantic classification, height-based features of the point are also significant as the eigenvalue-based ones, thus making the best optimal count of significant features to be 21. Now, consider three different ways the feature vectors across multiple scales are used.

The first method, which is a widely used method, is to determine an optimal scale where the eigenvalue-based features provide the minimum entropy (Demantké et al. 2011). The scale with lowest entropy is the scale at which the local neighborhood has the optimal shape, i.e., the local neighborhood has the least uncertainty in realizing the geometric class of the point. There is a variant of the method, where the optimal scale uses local minimum value instead of global minimum value in entropy. Overall, the method of aggregating multiple scales involves determining the optimal scale s_{opt} , and using the feature vector computed at s_{opt} . This is a conservative method, as the multiscale aggregation does not involve feature aggregation. The second method involves concatenation of multiscale features into a long vector (Hackel et al. 2016). This method is apt for both supervised learning using random forest classifiers as well as neural networks, for classification applications. The long feature vector is the most effective where locally significant features are determined and used, and is equivalent to the optimal scale method. The third method involves performing a map-reduce operation on the multiscale features. One way is to average the features, however the semantics of the features must be considered before performing the averaging operation (Sreevalsan-Nair and Kumari 2017). For instance, viewing different scales as mutually exclusive events justifies the averaging operation as the computation of the likelihood of the value of the concerned feature. The feature vectors computed using each of these three methods have been found to be effective for classification of different airborne and terrestrial laser scanned point clouds.

Future Scope

Scale selection continues to be an important problem to solve for multiscaling. One such solution, for images, uses a general methodology for spatial, temporal, and combined spatiotemporal dimensions (Lindeberg 2018). Most often, the scale selection in image processing and computer vision is applied sparsely in regions of interest (ROIs). The solution of the general methodology, on the contrary, performs dense scale selection where local extrema, such as minimum entropy, are computed at all points in the image and at all time instances. However, using mechanisms involving local extrema poses a problem of the strong dependency on the local order used for finding the locally dominant differential structure. This problem of phase dependency of local scales can be reduced by performing either post-smoothing or local phase compensation.

Machine learning (ML) methods have been applied to several problems in solid Earth geosciences (sEg), but with limited impact, owing to the limitations in data quality. Various physical systems seen in sEg are complex and nonlinear, implying that the data they generate are complex, with multiresolution structures in both space and time, e.g., fluid injection and earthquakes are strongly nonstationary (Bergen et al. 2019). Hence, the need of the hour is to generate more annotated data using multiscaling, which can model multiresolution data effectively. This can remove biases in frequent or well-characterized phenomena, thus improving the effectiveness of ML algorithms, especially those that rely on training datasets.

Altogether, using these illustrative examples, multiscaling can be seen as a powerful tool that has been successfully used in several geoscientific applications, with an equally large scope for influencing future applications.

References

- Bergen KJ, Johnson PA, Maarten V, Beroza GC (2019) Machine learning for data-driven discovery in solid Earth geoscience. *Science* 363(6433):eaau0323
- Demantké J, Mallet C, David N, Vallet B (2011) Dimensionality based scale selection in 3D LiDAR point clouds. *Int Arch Photogramm Remote Sens Spat Inf Sci* 38(Part 5):W12
- Hackel T, Wegner JD, Schindler K (2016) Fast semantic segmentation of 3D point clouds with strongly varying density. *ISPRS Ann Photogram Remote Sens Spatial Inf Sci* 3(3):177–184
- Lin H (2003) Hydrogeology: bridging disciplines, scales, and data. *Vadose Zone J* 2(1):1–11
- Lindeberg T (1998) Feature detection with automatic scale selection. *Int J Comput Vis* 30(2):79–116
- Lindeberg T (2018) Dense scale selection over space, time, and space-time. *SIAM J Imag Sci* 11(1):407–441
- Regenauer-Lieb K, Veveakis M, Poulet T, Wellmann F, Karrech A, Liu J, Hauser J, Schrank C, Gaede O, Trefry M (2013) Multiscale coupling and multiphysics approaches in earth sciences: theory. *J Coupled Syst Multiscale Dyn* 1(1):49–73

- Sagar BSD, Cheng Q, Agterberg F (2018) Handbook of mathematical geosciences: fifty years of IAMG. Springer International Publishing AG, Springer Nature, Gewerbestrasse 11, 6330 Cham, Switzerland.
- Sreevalsan-Nair J, Kumari B (2017) Local geometric descriptors for multi-scale probabilistic point classification of airborne lidar point clouds. In: Modeling, analysis, and visualization of anisotropy. Springer, Cham, pp 175–200
- Zhao W (2001) Multiscale analysis for characterization of remotely sensed images. PhD thesis, Louisiana State University, Baton Rouge

Multivariate Analysis

Monica Palma¹ and Sabrina Maggio²

¹Università del Salento-Dip. Scienze dell'Economia, Complesso Ecotekne, Lecce, Italy

²Dipartimento di Scienze dell'Economia, Università del Salento, Lecce, Italy

Synonyms

Multiple analysis; Multivariable statistics; Multivariate statistics

Definition

Multivariate analysis refers to a collection of different methods and techniques which allow studying the relationships characterizing data sets concerning two or more variables of the phenomenon of interest. These techniques are usually applied in the stage of exploratory data analysis, with the object to identify a reduced number of variables/factors which adequately describes the main characteristics shown by the data.

Multivariate Analysis in Geosciences

In several studies of environmental sciences, the phenomenon of interest is often the result of the simultaneous interaction of different variables observed at some locations of the survey area. In this case, the available data present a *multivariate spatial structure*, and the study is usually conducted on both (a) the spatial correlation which characterizes each observed variable and (b) the relationships existing among the variables. The first aspect is directly connected to the spatial structure of the data (spatially closer locations have often similar values compared to observations recorded at more distant locations), while the second aspect is related not only to the spatial features, but also to the nature of the phenomenon and the physical-chemical characteristics that link the variables to each other.

Often in the study of natural phenomena, such as climatic and atmospheric conditions in meteorology, or the concentrations of pollutants in environmental sciences, the variables of interest are observed at different points of the survey area for several instants of time. In this case, the data set presents a *multivariate space-time structure*, and the analyses must consider jointly the space-time correlation which characterizes each variable, as well as the relationships existing among the variables.

In geosciences, large data sets of multiple variables measured at several spatial or spatiotemporal points of the study area can be analyzed by using (a) the classical multivariate statistical techniques or (b) the multivariate geostatistical tools. Evidently, the research objectives are different even if complementary. Indeed, techniques of type (a) are usually applied in the stage of exploratory data analysis and most frequently as data reduction techniques; on the other hand, geostatistical tools are used with the aim of properly defining a model for the spatial/spatiotemporal multivariate correlation shown by the data and predicting the variables at unobserved points of the domain.

It is worth pointing out that often in geosciences compositional data arises, which is multivariate data and consists of vectors of positive components subject to a unit-sum constraint (Aitchison 1982). In such a context, the use of traditional multivariate analysis, which have been developed for unconstrained data, is not suitable. Because of the properties of compositional data (Aitchison 1994), it is necessary to work in terms of the corresponding log-ratio vectors and consequently to apply compositional data analysis in a log-ratio space.

Multivariate Statistical Techniques

The traditional multivariate statistical techniques, widely used for applications in social sciences, are usually applied in the stage of exploratory data analysis. A non-exhaustive list of multivariate techniques includes:

- Principal component analysis (PCA), which linearly transforms the observed variables into few uncorrelated variables which maximizing the data variability
- Canonical correlation analysis (CCA) that studies the relationships between two groups of variables representing different attributes of the same phenomenon
- Correspondence analysis (CorA), which is usually applied for studying the association between categorical variables, representing the variable's categories in a low-dimensional space
- Cluster analysis (CA) that tries to identify homogeneous groups of observations such that the dissimilarity among groups is remarkable

- Discriminant analysis (DA), which allows the identification of a linear combination of the observed variables that best separates two or more classes of observed data

Among these techniques, PCA is one of the most used because of its fast time of processing (from a computational point of view), as well as for the simplicity of results interpretation. Moreover, interesting insights into the variables' exploration come from the results obtained through the CCA. The fundamentals of the above mentioned techniques of classical multivariate analysis are reported in the next sections.

Main Aspects of PCA

PCA is typically a technique for data dimensionality reduction: PCA transforms (linearly) several correlated variables into few unrelated variables which maximize the variance characterizing the data (Hotelling 1933; Lebart et al. 1984; Jobson 1992; Jolliffe 2002).

PCA can be used also to understand possible relationships between variables, as well as to pre-process the data and eliminating redundancy caused by variables' correlation, before variographical analysis and modeling.

It is worth highlighting that in the PCA's applications the data are considered apart from the spatial or spatiotemporal points in which they are collected. In this context, the sample points represent the *cases*.

Let $\mathbf{X}$ be the $(N \times r)$ data matrix with N observations recorded for r variables (obviously, r is much greater than 2). In other words, the columns of $\mathbf{X}$ refer to the variables under study, while the rows of $\mathbf{X}$ refer to the measurements of the variables at the spatial or spatiotemporal points.

Through the PCA, the data matrix $\mathbf{X}$ is linearly transformed into an $(N \times r)$ matrix $\mathbf{Y}$, as follows:

$$\mathbf{Y} = \mathbf{X}\mathbf{Q},$$

where $\mathbf{Y}$, usually called *score matrix*, contains r orthogonal and zero mean components, while $\mathbf{Q}$, called *loading matrix*, is the $(r \times r)$ matrix of the eigenvectors of the following variance-covariance matrix

$$\mathbf{\Sigma} = \frac{1}{N}(\mathbf{X} - \mathbf{M})(\mathbf{X} - \mathbf{M})^T, \quad (1)$$

where the $(N \times r)$ matrix $\mathbf{M}$ is the matrix of the variables' mean values. Matrix $\mathbf{Q}$ is determined such that its first column represents the direction with the maximum variance in the data, while the second column of $\mathbf{Q}$ represents the direction of the next largest variance and so on.

Hence, PCA corresponds simply to the following eigenvalue analysis of the variance-covariance matrix $\mathbf{\Sigma}$

$$\mathbf{\Sigma} = \mathbf{Q}\mathbf{\Psi}\mathbf{Q}^T, \text{ with } \mathbf{Q}^T\mathbf{Q} = \mathbf{I},$$

where $\mathbf{\Psi}$ is the diagonal eigenvalues matrix of $\mathbf{\Sigma}$.

Moreover, as $\mathbf{\Sigma}$ is a positive semi-definite matrix, its eigenvalues are all non-negative; thus by arranging the eigenvalues in descending order, it is possible to select those components which correspond to the greatest eigenvalues; these components are named *principals* since they explain most of the total variance in the observed data. Therefore, if $r' < r$ is the number of the chosen principal components which correspond to the highest eigenvalues, then the data matrix $\mathbf{X}$ is adequately approximated as follows

$$\mathbf{X} \cong \tilde{\mathbf{Y}}\tilde{\mathbf{Q}}^T,$$

where $\tilde{\mathbf{Y}}$ is the $(N \times r')$ matrix of r' selected components and $\tilde{\mathbf{Q}}$ is the $(r \times r')$ matrix of the corresponding eigenvectors.

Note that the variance-covariance matrix $\mathbf{\Sigma}$ in (1) can be used if the variables under study are expressed with the same measurement units; otherwise, when the measurement units of the variables differ in size and type, the observed variables need first to be standardized, and successively the standardized variables are used for the computation of the correlation matrix which is scale-independent.

An interesting application of PCA on a multivariate environmental data set is in De Iaco et al. (2002).

Main Aspects of CCA

CCA is a classic multivariate statistical technique that studies the relationships between two groups of variables, which represent different attributes of the same phenomenon, for example, air pollutants and atmospheric variables, or concentrations of minerals and soil geologic features.

By performing CCA it is possible to identify pairs of factors relating to the two groups of variables under study, among which the correlation is maximum (Lebart et al. 1984; Jobson 1992).

Given the $(N \times r)$ data matrix $\mathbf{X}$ with N measurements (matrix rows) for the r variables (matrix columns) under study, the following matrix

$$\bar{\mathbf{X}} = \mathbf{X} - \mathbf{M}$$

is considered for the CCA, where $\mathbf{M}$ is the matrix of the variables' mean values.

According to the main features of the analyzed phenomenon, it is possible to define two different groups of, respectively, r_1 and r_2 variables, with $r_1 + r_2 = r$. Consequently, $\bar{\mathbf{X}}$ can be obtained through the horizontal concatenation of two matrices: the $(N \times r_1)$ matrix $\bar{\mathbf{X}}_1$ and $(N \times r_2)$ matrix $\bar{\mathbf{X}}_2$. In this way, the r columns of $\bar{\mathbf{X}}$ correspond to r_1 columns of $\bar{\mathbf{X}}_1$ plus r_2 columns of $\bar{\mathbf{X}}_2$, i.e.

$$\bar{\mathbf{X}} = [\bar{\mathbf{X}}_1 | \bar{\mathbf{X}}_2]$$

Given the following variance-covariance matrix for $\bar{\mathbf{X}}$

$$\Sigma = \frac{1}{N} (\bar{\mathbf{X}}^T \bar{\mathbf{X}}) = \frac{1}{N} [\bar{\mathbf{X}}_1 | \bar{\mathbf{X}}_2]^T [\bar{\mathbf{X}}_1 | \bar{\mathbf{X}}_2] = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

where

- Σ_{11} is the $(r_1 \times r_1)$ covariance matrix for the variables of the first group.
- Σ_{22} is the $(r_2 \times r_2)$ covariance matrix for the variables of the second group.
- Σ_{12} and Σ_{21} are, respectively, the $(r_1 \times r_2)$ and $(r_2 \times r_1)$ matrices of the covariances between the variables of the two groups, under the condition that $\Sigma_{12}^T = \Sigma_{21}$.

The CCA is applied in order to identify the matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ which maximize the correlation between

$$\mathbf{Y}_1 = \bar{\mathbf{X}}_1 \mathbf{A}_1 \quad \text{and} \quad \mathbf{Y}_2 = \bar{\mathbf{X}}_2 \mathbf{A}_2.$$

Note that the linear combination $\mathbf{Y}_1$ and $\mathbf{Y}_2$ represent the *canonical variates*, while $\mathbf{A}_1$ and $\mathbf{A}_2$ the canonical weights.

The canonical correlation coefficient given by

$$\rho_{(\mathbf{Y}_1, \mathbf{Y}_2)} = \frac{\mathbf{A}_1^T \Sigma_{12} \mathbf{A}_2}{\sqrt{(\mathbf{A}_1^T \Sigma_{11} \mathbf{A}_1)(\mathbf{A}_2^T \Sigma_{22} \mathbf{A}_2)}}, \quad (2)$$

measures the correlation between $\mathbf{Y}_1$ and $\mathbf{Y}_2$.

Hence, finding $\mathbf{A}_1$ and $\mathbf{A}_2$ such that the coefficient (2) is maximum corresponds to solving the following problem:

$$\max_{(\mathbf{A}_1, \mathbf{A}_2)} \mathbf{A}_1^T \Sigma_{12} \mathbf{A}_2,$$

under the condition that

$$\mathbf{A}_1^T \Sigma_{11} \mathbf{A}_1 = \mathbf{A}_2^T \Sigma_{22} \mathbf{A}_2 = \mathbf{1}.$$

Finally, the inspection of the magnitudes and signs of $\mathbf{A}_1$ and $\mathbf{A}_2$ allows the analyst to detect those variables which maximize the correlation, as well as to have insights into the contrasts among the variables under study.

A very interesting application of CCA on a multivariate environmental data set is in De Iaco (2011).

Main Aspects of CorA

The theory of CorA is discussed in several books (Benzécri 1983; Lebart et al. 1984; Greenacre 1989); therefore, the main characteristics of this multivariate technique will be reviewed.

Let $\mathbf{X}$ be the observed $(l \times p)$ two-way contingency table, where each entry x_{ji} represents the joint absolute frequency associated to the j -th and i -th attributes of two categorical variables, with $j = 1, \dots, l$, and $i = 1, \dots, p$.

By performing CorA, the best simultaneous representation of the rows and columns of $\mathbf{X}$ is determined in a low-dimensional space.

Let

$$\mathbf{D}_l = \text{diag}[f_{j\cdot}] \quad (3)$$

and

$$\mathbf{D}_p = \text{diag}[f_{\cdot i}], \quad (4)$$

be two diagonal matrices of the marginal relative frequencies computed as follows

$$f_{\cdot j} = \sum_{i=1}^p f_{ji}, \quad j = 1, \dots, l, \quad (5)$$

and

$$f_{j\cdot} = \sum_{i=1}^p f_{ji}, \quad i = 1, \dots, p, \quad (6)$$

where

$$f_{ji} = \frac{x_{ji}}{\sum_{j=1}^l \sum_{i=1}^p x_{ji}}, \quad j = 1, \dots, l, \quad i = 1, \dots, p. \quad (7)$$

Through CorA, a vector $\mathbf{u}$, which maximizes

$$\mathbf{u}^T [\mathbf{D}_l^{-1} \mathbf{F} \mathbf{D}_p^{-1}]^T \mathbf{D}_l [\mathbf{D}_l^{-1} \mathbf{F} \mathbf{D}_p^{-1}] \mathbf{u}, \quad (8)$$

under the following constraint

$$\mathbf{u}^T \mathbf{D}_p^{-1} \mathbf{u} = \mathbf{1}, \quad (9)$$

is finding in a p -dimensional space.

In the same way, through CorA a vector $\mathbf{v}$ which maximizes

$$\mathbf{v}^T [\mathbf{D}_p^{-1} \mathbf{F}^T \mathbf{D}_l^{-1}]^T \mathbf{D}_p [\mathbf{D}_p^{-1} \mathbf{F}^T \mathbf{D}_l^{-1}] \mathbf{v}, \quad (10)$$

under the constraint that

$$\mathbf{v}^T \mathbf{D}_l^{-1} \mathbf{v} = \mathbf{1}, \quad (11)$$

is determined in an l -dimensional space. The following factors

$$\Phi = \mathbf{D}_p^{-1} \mathbf{u} \text{ and } \Psi = \mathbf{D}_l^{-1} \mathbf{v}$$

define the plane where the rows and columns of $\mathbf{X}$ are projected. CorA diagram (*biplot*) allows the user to find latent relationships among the categories of the observed variables (Lebart et al. 1984).

An interesting application of CorA on a spatiotemporal environmental data set can be found in Palma (2015).

Main Aspects of CA

CA is a technique of multivariate analysis used for grouping a set of variables in such a way that variables which are “similar” to each other belong to the same group (called *cluster*), while variables which are “not similar” belong to different groups. The concept of similarity is in the sense that variables which present characteristics homogeneous within each group and different from those of the variables belonging to other groups can be considered similar (Everitt 1974).

This goal can be achieved by various algorithms which deeply differ in the way the final clusters are identified. In the literature, several clustering algorithms have been developed which can be classified in *hierarchical* and *not hierarchical* or *partitionial* clustering algorithms.

Hierarchical clustering techniques generate the well-known *dendrogram* (from the Greek words “dendro” which means tree and “gramma” drawing), i.e., a particular tree diagram which shows the clustering sequences. These methods of clustering can be:

- (a) *Divisive*, if it starts from a single cluster and step-by-step splits the cluster(s) until each cluster consists of a single variable
- (b) *Agglomerative*, if it starts with so many clusters as the observed variables, then step-by-step the most similar clusters are merged together until all the variables are assigned to a single cluster.

On the other hand, not hierarchical clustering approaches divide the data set into a user selected number of clusters, minimizing certain optimization criteria (e.g., a square error function). Therefore these techniques are based on iterative algorithms which find the optimum number of clusters when a specific stopping criteria is satisfied (i.e., a maximum number of iterations fixed by the user).

In Zhou et al. (2018), a good application of CA on geochemical data is discussed.

Main Aspects of DA

DA, introduced by Fisher (Fisher 1936) to classify some fossil remains as monkeys or humanoids, is a very useful technique of multivariate analysis. In particular, this technique allows the definition of a linear combination of the observed variables which best separates two or more classes of cases. By applying DA on a multivariate data set, a model (called *discriminant function*) for the classification of the cases under study is defined.

Unlike CA, this method starts from a known classification of the cases under study; successively, by defining a latent discriminant function which synthesizes the explanatory (quantitative) variables, new cases can be assigned to one of the groups of the primary (qualitative) variable. Hence, DA can be applied for descriptive aims, namely, to classify cases, as well as for prediction purposes, namely, to assign a new case to a specific group.

Let's consider a $(N \times r)$ matrix $\mathbf{X}$ of $r \geq 2$ variables observed on N sample cases. Moreover, let's assume that the cases under study belong to K mutually exclusive groups defined by K different attributes of a categorical variable. Each group is composed by N_k cases, such that $\sum_{k=1}^K N_k = N$.

By applying DA on matrix $\mathbf{X}$, a linear combination

$$\mathbf{Y} = \mathbf{A}^T \mathbf{X}$$

of the r variables is obtained, such that the K groups are best separated. In other words, among all the linear combinations of the observed variables, DA allows the identification of those linear combinations characterized by the maximum between groups variance (in this way the difference between groups are maximized) and the minimum within-group variance (in this way the spread within the groups are minimized).

Hence, the ratio of the between-group variance on the within-group variance has to be maximized with respect to the vector $\mathbf{A}$.

The function which maximizes this ratio is called *first discriminant function*. Then, the obtained function can be used for predictive purposes in order to assign new cases to each group.

Lonoce et al. (2018) proposed an interesting application of DA on a multivariate data set concerning archaeological measurements.

Multivariate Geostatistical Tools

In geostatistics, one of the main objectives is to estimate the variable of interest at unsampled points of the domain. For this aim, it is required a model which is able to adequately describe the spatial or spatiotemporal correlation of the variable under study. In the univariate context, techniques for

modeling and estimating spatial/spatiotemporal data are widely developed (Chilés and Delfiner 1999; Cressie 1993). Moreover, the existence in the literature of large classes of admissible correlation functions (covariances or variograms), from which selecting the most suitable function for the variable under study, greatly simplifies the modeling stage. On the other hand, in multivariate geostatistics, only few models have been proposed that can explain both the correlation which characterizes each variable (often called *direct correlation*) and the one existing between the observed variables (known as *cross-correlation*). Furthermore, there are several issues related to (a) the spatial or spatiotemporal sampling, (b) the choice of valid multivariate models, (c) the lack of automatic fitting procedures, which make modeling and estimation tasks quite complex.

In multivariate geostatistics, the concept of *regionalization*, first introduced by Matheron (1965) to highlight the close link between the observed values of one variable and the locations at which the values are measured, has been extended at the *coregionalization* concept, referring to regionalized variables which are spatially or spatiotemporally dependent on each other.

The probabilistic approach to coregionalization is based on the concept of multivariate random function (MRF) in space or space-time.

Let $\{Z_i(\mathbf{s}), \mathbf{s} \in D \subseteq \mathbb{R}^d\}$, $i = 1, \dots, r$, be r random functions (RFs) on the spatial domain D , where $\mathbf{s} = (s_1, s_2, \dots, s_d)$ are the spatial coordinates (usually $d \leq 3$).

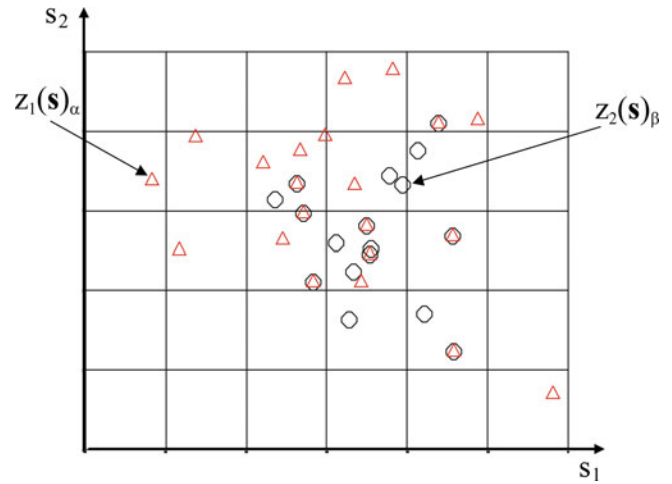
The vector

$$\mathbf{Z}(\mathbf{s}) = [Z_1(\mathbf{s}), Z_2(\mathbf{s}), \dots, Z_r(\mathbf{s})]^T \quad (12)$$

represents a spatial MRF. Hence, the RFs $Z_i(\mathbf{s})$, $i = 1, \dots, r$, which compose the vector (12) are associated with the observed variables. In particular, the measurements $z_i(\mathbf{s})_\alpha$, $\alpha = 1, \dots, N_i$, $i = 1, \dots, r$, are considered as realizations of a spatial MRF $\mathbf{Z}$ (Fig. 1).

As already pointed out, natural phenomena are usually characterized by a space-time evolution and are the result of the joint interaction of different correlated variables. Evidently, the analysis of these processes requires the measurements of two or more variables, at some locations of the survey area and for several time points. Therefore, the data set presents a multivariate space-time structure, and the analysis of its spatiotemporal distribution can be adequately performed through space-time multivariate geostatistical techniques. In this case, the MRF definition can be extended to the spatiotemporal case. In particular, let

$$\{Z_i(\mathbf{s}, t), (\mathbf{s}, t) \in D \times T \subseteq \mathbb{R}^{d+1}\}, \quad i = 1, \dots, r,$$



Multivariate Analysis, Fig. 1 Example of a sampling scheme for two variables over a 2D spatial domain

be r spatiotemporal RFs over the domain $D \times T$ where $\mathbf{s} \in D \subseteq \mathbb{R}^d$ (usually $d \leq 3$) and $t \in T \subseteq \mathbb{R}$ is the time-point.

Hence, the vector

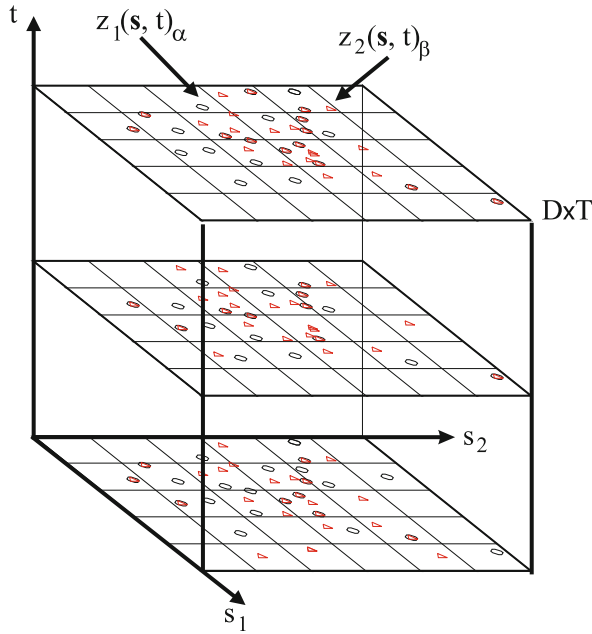
$$\mathbf{Z}(\mathbf{s}, t) = [Z_1(\mathbf{s}, t), \dots, Z_r(\mathbf{s}, t)]^T \quad (13)$$

represents a spatiotemporal MRF.

Given the RFs Z_i , $i = 1, \dots, r$, related to the $r \geq 2$ spatiotemporal variables under study, the measurements $z_i(\mathbf{s}, t)_\alpha$, $\alpha = 1, \dots, N_i$, $i = 1, \dots, r$, are assumed as realizations of the spatiotemporal MRF $\mathbf{Z}$ (Fig. 2).

In multivariate geostatistics, the modeling procedure is essentially based on the identification of a spatial/spatiotemporal correlation model between the components of the MRF, which is able to describe the overall behavior of the phenomenon under study. For multivariate spatial data, several models have been explored (Goovaerts 1997; Wackernagel 2003; Gelfand et al. 2004; Gneiting et al. 2010), while in the space-time domain, modeling approaches and methodologies have been discussed in Christakos (1992); Christakos et al. (2002); Fassó and Finazzi (2011); and De Iaco et al. (2019). However, both in spatial and spatiotemporal context, the *linear coregionalization model* (LCM) is still the most used model to explain the correlation among the RFs of a multivariate process. The computational flexibility which characterizes the LCM represents its main strength, despite some limitations of the model which will be pointed out later on.

Let $\mathbf{Z}(\mathbf{u}) = [Z_1(\mathbf{u}), Z_2(\mathbf{u}), \dots, Z_r(\mathbf{u})]^T$ be a vector of second-order stationary RFs in a spatial domain when $\mathbf{u} = \mathbf{s} \in D \subseteq \mathbb{R}^d$ or in a space-time domain when $\mathbf{u} = (\mathbf{s}, t) \in D \times T \subseteq \mathbb{R}^{d+1}$, with $d \leq 3$. Then, first- and second-order moments of $\mathbf{Z}$ are, respectively



Multivariate Analysis, Fig. 2 Example of spatiotemporal sampling scheme for two variables (2D spatial domain)

$$E[\mathbf{Z}(\mathbf{u})] = \mathbf{M} = [\mu_1, \dots, \mu_r]^T, \quad \forall(\mathbf{u}) \in D \times T, \quad (14)$$

where $\mu_i = E[Z_i(\mathbf{u})]$, $i = 1, \dots, r$, for all points $\mathbf{u}$ of the domain, and

$$\mathbf{C}(\mathbf{h}) = [C_{ik}(\mathbf{h})] \quad (15)$$

with, respectively, $C_{ii}(\mathbf{h}) = \text{Cov}\{Z_i(\mathbf{u}), Z_i(\mathbf{u}')\}$, the direct covariances and $C_{ik}(\mathbf{h}) = \text{Cov}\{Z_i(\mathbf{u}), Z_k(\mathbf{u}')\}$, the cross-covariances ($i \neq k$) of the r RFs of $\mathbf{Z}$; the vector $\mathbf{h}$ represents the separation vector between the points $\mathbf{u}$ and $\mathbf{u}'$; in particular $\mathbf{h} = \mathbf{h}_s$ in the spatial case where $\mathbf{u} = \mathbf{s}$ and $\mathbf{u}' = \mathbf{s}'$, while $\mathbf{h} = (\mathbf{h}_s, h_t)$ with $\mathbf{h}_s = (\mathbf{s}' - \mathbf{s})$ and $h_t = (t' - t)$ in the spatiotemporal case where $\mathbf{u} = (\mathbf{s}, t)$ and $\mathbf{u}' = (\mathbf{s}', t')$.

Given the orthogonal vectors $\mathbf{Y}_p(\mathbf{u}) = [Y_p^1(\mathbf{u}), \dots, Y_p^r(\mathbf{u})]^T$, $p = 1, \dots, P$, whose r components are second-order stationary RFs, with

$$E[Y_p^v(\mathbf{u})] = 0, \quad v = 1, \dots, r, \quad p = 1, \dots, P,$$

and

$$\text{Cov}[Y_p^v(\mathbf{u}), Y_{p'}^{v'}(\mathbf{u}')] = \begin{cases} c_p(\mathbf{h}) & \text{if } v = v', \quad p = p', \\ 0 & \text{otherwise,} \end{cases}$$

by assuming the LCM, the MRF $\mathbf{Z}$ is defined as follows

$$\mathbf{Z}(\mathbf{u}) = \sum_{p=1}^P \mathbf{A}_p \mathbf{Y}_p(\mathbf{u}) + \mathbf{M} \quad (16)$$

where $\mathbf{A}_p$, $p = 1, \dots, P$, are the $(r \times r)$ matrices of coefficients and $\mathbf{M}$ is the vector defined as in (14).

In other words, the LCM assumes that each Z_i is described through a linear combination of r orthogonal RFs Y_p^v , $v = 1, \dots, r$, $p = 1, \dots, P$, at P different variability scales. Hence, in the LCM, the matrix $\mathbf{C}(\mathbf{h})$ in (15) can be expressed as a linear combination of the basic covariance functions c_p , i.e.

$$\mathbf{C}(\mathbf{h}) = \sum_{p=1}^P \mathbf{B}_p c_p(\mathbf{h}), \quad (17)$$

where the $(r \times r)$ matrices

$$\mathbf{B}_p = \mathbf{A}_p \mathbf{A}_p^T, \quad p = 1, \dots, P,$$

must be positive-definite for the admissibility of the above model.

In spite of the computational simplicity and flexibility of the coregionalization model, both in spatial and spatiotemporal context (Wackernagel 2003; De Iaco et al. 2010, 2012), some limitations of the model need to be highlighted.

In particular, in the coregionalization model, the cross-covariance functions satisfy the following properties of symmetry:

$$C_{ik}(\mathbf{h}) = C_{ki}(\mathbf{h}), \quad (18)$$

$$C_{ik}(\mathbf{h}) = C_{ik}(-\mathbf{h}), \quad (19)$$

with $i, k = 1, \dots, r$, $i \neq k$.

Evidently, in the spatiotemporal context, $\mathbf{h} = (\mathbf{h}_s, h_t)$ and $-\mathbf{h} = (-\mathbf{h}_s, -h_t)$.

As known, in general the cross-covariance functions are not symmetric and do not satisfy conditions (18) and (19). In some applications the correlation between two variables is asymmetric with respect to the separation vector $\mathbf{h}$, therefore assuming symmetry for the cross-covariances can produce unreliable results. In this context, it could be useful the definition of covariance functions which describe the direct and cross-correlation among the variables and do not satisfy the above properties.

It is worth noting that in 1993 Grzebyk (1993) extended the LCM to the complex domain allowing the use of not-even cross-covariance functions, and more recently Li and Zhang (2011) discussed asymmetric cross-covariance functions.

In the coregionalization model, the cross-covariance functions satisfy not only the symmetry assumptions (viz., invariance with respect to the spatial or spatiotemporal separation

vector, as well as to the exchange of the variables) but also the non-separability condition. In other words, it is assumed that the cross-covariances C_{ik} , $i, k = 1, \dots, r$ are not proportional, namely, they are not described by the following intrinsic model

$$C_{ik}(\mathbf{h}) = a_{ik}\rho(\mathbf{h}),$$

where $\rho(\cdot)$ represents a correlation function, while a_{ik} , $i, k = 1, \dots, r$, represent the elements of an $(r \times r)$ matrix which is positive-definite.

Several methodologies have been developed in the literature, in order to assess symmetry and separability/non-separability conditions of the cross-covariances. Therefore, before adopting the coregionalization model to synthesize the multivariate spatial or spatiotemporal correlation among the analyzed variables, it is convenient to check if symmetry and non-separability conditions are satisfied, as thoroughly discussed in De Iaco et al. (2019) and Cappello et al. (2021).

Summary

The study of a multivariate data set concerning two or more variables of the phenomenon of interest can be done by using either classical multivariate statistical techniques mainly for exploratory purposes (detection of specific features or identification of a reduced number of hidden factors which are common to the observed variables) or multivariate geostatistical techniques to model the spatial/spatiotemporal correlation which characterizes the analyzed data.

Cross-References

- [Compositional Data](#)
- [Exploratory Data Analysis](#)
- [Principal Component Analysis](#)
- [Spatiotemporal](#)
- [Spatial Data](#)

Bibliography

- Aitchison J (1982) The statistical analysis of compositional data. *J R Stat Soc Ser B Methodol* 44(2):139–177
- Aitchison J (1994) Principles of compositional data analysis. Institute of Mathematical Statistics Lecture Notes, Monograph Series, Editor(s) Anderson TW, Fang KT, Olkin I, 24: 73–81
- Benzécri JP (1983) Histoire et préhistoire de l'analyse des données. Dunod, Paris
- Cappello C, De Iaco S, Palma M, Pellegrino D (2021) Spatio-temporal modeling of an environmental trivariate vector combining air and soil measurements from Ireland. *Spat Stat* 42:1–18
- Chilés J, Delfiner P (1999) Geostatistics - modeling spatial uncertainty. Wiley, New York
- Cressie N (1993) Statistics for spatial data. Wiley, New York
- Christakos G (1992) Random field models in earth sciences, 1st edn. Academic Press
- Christakos G, Bogaert P, Serre M (2002) Temporal GIS. Advanced functions for field-based applications. Springer
- De Iaco S (2011) A new space-time multivariate approach for environmental data analysis. *J Appl Stat* 38:2471–2483
- De Iaco S, Maggio M, Palma M, Posa D (2012) Chapter 14: Advances in spatio-temporal modeling and prediction for environmental risk assessment. In: Haryanto B (ed) *InTech, Air pollution: a comprehensive perspective*, IntechOpen, Croatia, 365–390
- De Iaco S, Myers DE, Posa D (2002) Space-time variograms and a functional form for total air pollution measurements. *Comput Stat Data Anal* 41(2):311–328
- De Iaco S, Myers DE, Palma M, Posa D (2010) FORTRAN programs for spacetime multivariate modeling and prediction. *Comput Geosci* 36(5):636–646
- De Iaco S, Palma M, Posa D (2019) Choosing suitable linear coregionalization models for spatiotemporal data. *Stochastic Environ Res Risk Assess* 33:1419–1434
- Everitt B (1974) Cluster analysis. Social Science Research Council, Heinemann, London
- Fassó A, Finazzi F (2011) Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. *Environmetrics* 22:735–748
- Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Ann Eugenics* 7(2):179–188
- Gelfand AE, Schmidt AM, Banerjee S, Sirmans CF (2004) Nonstationary multivariate process modeling through spatially varying coregionalization. *Sociedad de Estadística e Investigación Operativa. Test* 13:263–312
- Gneiting T, Kleiber W, Schlather M (2010) Matérn cross-covariance functions for multivariate random fields. *J Am Stat Assoc* 105(491):1167–1177
- Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York
- Greenacre MJ (1989) Theory and applications of correspondence analysis. Academic Press, London
- Grzebyk M (1993) Ajustement d'une corégionalisation stationnaire, Doctoral Thesis, Ecoles des Mines de Paris, France
- Hotelling H (1933) Analysis of a complex of statistical variables into principal components. *J Educ Psychol* 24(6):417–441
- Jobson JD (1992) Applied multivariate data analysis, 2nd edn. Springer, New York
- Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, Berlin
- Lebart L, Morineau A, Warwick KM (1984) Multivariate descriptive statistical analysis. Wiley, New York
- Li B, Zhang H (2011) An approach to modeling asymmetric multivariate spatial covariance structures. *J Multivar Anal* 102:1445–1453
- Lonoce N, Palma M, Viva S, Valentino M, Vassallo S, Fabbri PF (2018) The Western (Buonfornello) necropolis (7th to 5th BC) of the Greek colony of Himera (Sicily, Italy): site-specific discriminant functions for sex determination in the common burials resulting from the battle of Himera (ca. 480 BC). *Int J Osteoarchaeol* 28:766–774
- Matheron G (1965) La Théorie des Variables Régionalisées et ses Applications. Masson, Paris
- Palma M (2015) Correspondence analysis on a space-time data set for multiple environmental variables. *Int J Geosci* 6:1154–1165
- Wackernagel H (2003) Multivariate geostatistics: an introduction with applications. Springer, Berlin
- Zhou S, Zhou K, Wang J, Yang G (2018) Application of cluster analysis to geochemical compositional data for identifying ore-related geochemical anomalies. *Front Earth Sci* 12:491–505

Multivariate Data Analysis in Geosciences, Tools

John H. Schuenemeyer

Southwest Statistical Consulting, LLC, Cortez, CO, USA

Definition

Multivariate data occurs frequently in the geosciences. Graphical and statistical tools are applied to extract information from large, complex data sets. Methods are suggested for different types of problems.

Introduction

Multivariate analysis in the geosciences begins with an exploratory procedure that allows the analyst to gain a better understanding of a complex data set. Data is obtained in many ways. The preferred method would be from designed experiments, often categorized by a formal experimental design or sampling design since they yield more generalizable results. However, because of the cost of sampling (the need for ships, satellites, deep drilling), time sensitivity (floods and earthquakes), and accessibility of sites (access to land, harsh conditions), this may not be possible.

This entry discusses procedures by categories keyed to data analytic tasks. The tasks within categories include reducing the number of variables, clustering observations, gaining insight into correlation among variables, and classifying observations. Most of the variables we present are metric variables, which are variables that can be measured quantitatively. These include interval, ratio, and continuous variables.

Examples are from the geosciences, broadly defined to include energy, geology, archaeology, climate, and ecology. Procedures used to understand data obtained in these disciplines overlap. Geochemical data is common in geosciences. Often it is partly compositional data, which will be presented after a discussion of categories. Selected references are provided. A general reference with geoscience examples is by Schuenemeyer and Drew (2011).

Categories

The categories below list specific tasks that need to be addressed while performing multivariate exploratory data analysis. Under each category, one or more approaches are presented. In the section after this, alternative approaches are presented:

1. **Graphical procedures.** Graphical procedures are usually the first stage in multivariate exploratory data analysis and frequently are an extension of those used in univariate or bivariate analysis. They are used in the methods described in the following categories. Boxplots are one of the most effective ways to compare distributions of several variables expressed in the same unit. Density plots are alternatives. Using three dimensions plus color, size, and shape can extend a two-dimensional scatter plot to six dimensions. In addition, this system can be rotated. Faceting or grouping allows plotting of multiple variables on a single graph such as a histogram or boxplot.
2. **Dimensionality reduction.** Data sets in the earth sciences sometimes have dozens to hundreds of variables, partly due to the high initial cost of sampling and the relatively low marginal cost of sampling additional variables. In such cases, the variables can be highly correlated, meaning that the dimensionality of the system is less than the number of variables. This can result in interpretation, storage, and computational problems. Some approaches to transform a data set in high dimensional space to one in a lower dimensional space are:
 - (a) Principal component analysis (PCA) is an orthogonal transformation of a system of correlated variables into one of uncorrelated variables. The result is an ordered system by percent of variability accounted for by principal components $PC_1, \dots, PC_p$, where p is the number of variables. The hope is that a significant percent of the total variability will be accounted for by m variables where $m \ll p$. For example, we show (Schuenemeyer and Drew 2011) that 78% of the total variability of 21 variables from the Freeport coal bed was accounted for by seven variables.
 - (b) A second, more modern but more difficult, procedure to implement is reduction of the number of variables by machine learning. In a machine learning approach, criteria are established that variables must meet to be retained. These could include variability, distance, and passage of statistical tests.
3. **Clustering of observations by group where the variables have similar characteristics.** Cluster analysis seeks to group observations by like variables. As a result of clustering, observations in each group should be more similar than those in different groups. This reduces the number observations, as opposed to dimensional reduction of variables. Examples of cluster analysis in earth and environmental science include:
 - Identifying waste sites based on variables describing measures of toxicity.
 - Defining climate zones based on measures of pressure, precipitation, and temperature.

- Identifying archaeological sites based on the chemical composition of pottery sherds.

The number of approaches to finding clusters is too numerous to describe in this report, but here are possibilities:

- (a) Hierarchical clustering is a tree-based approach divided into two types of algorithms – agglomerative and divisive. An *agglomerative approach* begins at the lowest level, that is, each observation is a cluster. The first step upward is to combine the two observations that are closest according to some criterion, which could be as simple as medians. The path upward continues until a stopping rule or one cluster consisting of all observations is reached. The inverse of this procedure is a *divisive hierarchical approach*. It begins with all observations in a single cluster and then splits groups to achieve the most diversity.
- (b) Clustering by partitioning algorithms is another set of procedures to use. Unlike with hierarchical algorithms, the number of clusters is specified in advance, and observations can be moved from cluster to cluster until some measure of closure is achieved. The k-mean algorithm is frequently used. Other methods include fuzzy set and method clustering.
- (c) Multidimensional scaling (MDS) is a dimension reducing visualization method that may be effective when the number of dimensions can be reduced to three or fewer. The purpose is to reduce dimensionality while approximating the original distances or similarities. Dzwinel et al. (2005) illustrate its use in investigating earthquake patterns. MDS can also be used as a nonparametric alternative to factor analysis (FA) because it does not need the assumption of multivariate normality or a correlation matrix.
- (d) Correspondence analysis (CA) is similar to MDS.

There is no single best approach to clustering as it depends on the pattern in your data set such as elliptical or a snake-like curvature pattern. Achieving a satisfactory partitioning of data into clusters is largely a trial and effort procedure.

4. **Gaining insight into correlation among variables.** Factor analysis is a popular method used to perform this task. There are similarities to PCA, but there are important differences. PCA is a mathematical transformation. FA is a statistical model. In addition, the goal of FA is to find correlation structures. The goal of PCA is to maximize the variance of the new variables. FA is used as a statistical model when the number of factors is known and specified in advance. Since this is rarely the case, FA is more often an EDA method where several estimates of the number of factors are made. Typical output includes a factor loading

matrix where the columns are factors and each observation down a column represents the factor and variable. For example, in a two-factor solution, the variables Ca and SO₄ may have a high correlation with Factor 1 and low correlation with Factor 2. Conversely, variables Na and Cl may have a high correlation with Factor 2 and a low correlation with Factor 2. R-mode and Q-mode are alternative forms of factor analysis.

5. **Determining separation between known groups and/or classifying an object of unknown origin into one of two or more groups.** For example, an archaeologist has identified two nearby sites from buried samples from different time periods. Two questions arise: (1) How do we determine the spatial overlap, if any, among sites? (2) How do we determine with probability, to which site an observation of unknown origin belongs? We examine two approaches:

- (a) Discriminant analysis (DA) is a procedure used to determine separation between known groups and/or to classify an observation of unknown origin into one of two or more predefined groups. Two models are linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). The LDA is a model in which the population variance-covariance matrices are assumed to be the same for all sites. For QDA the homoscedastic assumption does not hold. In the archaeology example, the data set consists of two sites labeled K and N. The sites were identified by samples having variables Cu and Pb.
- (b) Tree-based modeling is a nonparametric alternative to LDA. Variables can be continuous, counts, or categorical. The procedure begins at the root node and splits on the variable that meets a classification rule that would make the splits at termination as pure as possible.

Automating Geoscience Analysis

Automated systems to understand large, complex data in many geological science disciplines are relatively new. They are discussed below.

1. **Structural equation modeling (SEM).** This methodology (Grace 2020) represents an approach to statistical modeling that focuses on the study of complex cause-effect hypotheses about the mechanisms operating in systems. SEM is used increasingly in ecological and environmental studies.
2. **Data mining versus machine learning.** Data mining (Bao et al. 2012) is a tool used to gain insight into data. The

- algorithms listed in the Categories section fall under the category of data mining. Even an algorithm to implement a procedure as simple as liner regression belongs to the data mining category because it yields information from a set of linear data. A slightly more complex algorithm would be one that distinguishes between linear and quadratic data. However, we typically think of data mining as yielding information from a larger and more complex data set. A distinguishing feature between data mining and machine learning is that machine learning (Korup and Stolle 2014; Srivastava et al. 2017) can teach itself from previous analyses of data. Data mining may be a procedure used by machine learning.
3. **Supervised versus unsupervised learning.** Most statistical learning processes fall into two categories, supervised and unsupervised learning (James et al. 2017). A statistical learning process is building a model where the data contains an input and an output. An example is regression where a possible set of explanatory variables is the input and the explanatory variable is the output. This is the training set. The supervised learning algorithm learns a function which can estimate output for future data. In unsupervised learning, there is only one set of data and the algorithm learns from the data. An example is clustering.

Compositional Data Analysis (CoDa)

This refers to the analysis of compositional data, also known as constant sum data (van den Boogaart and Tolosana-Delgado 2013; Pawlowsky-Glahn et al. 2015). Geochemical data occurs frequently in the geosciences. Compositional data are portions of a total. They include variables known as components that are expressed as fractions, percentages, or parts per million. This data has an induced correlation because, when there are p variables, only $p - 1$ are independent. Because of this closure, a variance matrix will always be singular, and the variables cannot be normally distributed. Therefore, standard statistical methods do not apply. Alternative approaches include performing a log-ratio transformation prior to analysis.

A Numerical Example

Data in the geosciences is often compositional; however, it is not always analyzed in that manner. We present a small, constructed example of a subset of a compositional data set consisting of three variables (Cu, Pb, and Zn) and 92 observations and four sites. A sample is shown in Table 1. Typically, there would be more compositional variables that may be of little interest to the investigator. The units here are percent but parts per million is another way to express compositional data.

The problem is to see if linear discriminant analysis (LDA) is a reasonable model to identify separation of existing sites using the three variables identified above. The two procedures we consider are (1) to use the data shown above and (2) to divide each of the three elements by their sum so that the variables in each observation sum to 1. A third procedure (not shown) would be to add a residual variable to the elements that would be the sum of the omitted variables.

Table 2 compares the prior (known) data with the LDA posterior results using the original data as the new data set in the LDA prediction algorithm.

The prior column shows the number of observations by site. The posterior row shows the allocation made by the prediction. The diagonal shows the number of correct allocations based on the maximum probability assigned to each observation and site. For this example, the sum of the diagonal is 65; however, a typical but arbitrary measure of accuracy is that the maximum probability must be at least 0.50. There are five correct assignments with probability less than 0.50. Applying this rule would yield an estimation accuracy of 0.65.

Now we examine the results obtained when the data is changed to constant sum and the compositional LDA is used. The major difference with the standard LDA is that the data argument to compositional LDA is the isometric log-ratio (ilr) transformation of the constant sum data. Following the prediction, an inverse ilr is made mapping the data back to their original units (those shown in Table 1).

Multivariate Data Analysis in Geosciences, Tools, Table 1 A subset of compositional data; units are percent

Site	Cu	Pb	Zn	Site	Cu	Pb	Zn
A	11.1	5.4	6.1	C	8.6	2.7	6.3
A	9.8	1.8	17.4	C	8.1	5.5	5.0
A	8.9	1.9	13.2	C	8.1	5.0	4.9
B	8.3	2.0	5.8	D	9.2	1.6	9.3
B	7.9	5.9	8.1	D	9.1	2.7	7.9
B	7.9	2.4	9.4	D	8.9	4.7	10.1

Multivariate Data Analysis in Geosciences, Tools, Table 2 A comparison of prior and posterior results using the untransformed data

	Site				
	A	B	C	D	Prior
A	9	2	0	3	14
B	0	24	1	5	30
C	0	5	5	5	15
D	2	3	1	27	33
Posterior	11	34	7	40	92

Multivariate Data Analysis in Geosciences, Tools, Table 3 A comparison of prior and posterior results using the ilr transformed data

	Site				
	A	B	C	D	Prior
A	7	5	1	1	14
B	2	20	2	6	30
C	0	2	10	3	15
D	1	6	3	23	33
Posterior	10	33	16	33	92

The definitions are the same as those in Table 3. However, for this example, the sum of the diagonal is 60, and there are six correct assignments with probability less than 0.50. Applying this rule would yield an accuracy of 0.59.

There are numerous reasons for differences in accuracy between the two approaches that go far beyond the scope of this report. However, it is proper to use the compositional LDA when appropriate.

Summary and Conclusion

This report presents tools for a user to gain insight into multivariate data. They are listed by category, and within each category there are brief discussions of different algorithms. Alternative tools are discussed that are useful when analyzing large complex data sets. Finally, a brief discussion of compositional data followed by an example is presented because its use in the geosciences is common and such data requires a nonstandard statistical approach to analysis.

Bibliography

- Bao F, He X, Zhao F (2012) Applying data mining to the geosciences data. *Phys Procedia* 33:685–689
- Dzwiniel W, Yuen D, Boryczko K, Ben-Zion Y, Yoshioka S, Ito T (2005) Nonlinear multidimensional scaling and visualization of earthquake clusters over space, time and feature space. *Nonlinear Process Geophys* 12(1):117–128
- Grace J (2020) Quantitative analysis using structural equation modeling, USGS, Wetlands and Aquatic Research Center. <https://www.usgs.gov/centers/wetland-and-aquatic-research-center>. Accessed 22 Nov 2020
- James G, Witten D, Hastie T, Tibshirani R (2017) An introduction to statistical learning with applications in R. Springer, New York, 426 p
- Korup O, Stolle A (2014) Landslide prediction from machine learning. *Geol Today* 30:26–33. Wiley Online Library
- Pawlowsky-Glahn V, Egozcue J, Tolosana-Delgado R (2015) Modeling and analysis of compositional data. Wiley, Hoboken. 272 p
- Schuenemeyer J, Drew L (2011) Statistics for earth and environmental scientists. Wiley, Hoboken. 407 p
- Srivastava A, Nemani R, Steinhäuser K (eds) (2017) Large-scale machine learning in the earth sciences. Data mining and knowledge discovery series. Chapman & Hall/CRC, Boca Raton. 226 p
- van den Boogaart K, Tolosana-Delgado R (2013) Analyzing compositional data with R. Springer, Berlin. 258 p

Němec, Václav

Jan Harff¹ and Niichi Nichiwaki²

¹Institute of Marine and Environmental Sciences, University of Szczecin, Szczecin, Poland

²Professor Emeritus of Nara University, Nara, Japan



Fig. 1 Václav Němec, 2019 (Photo by Ekaterina Solntseva)

Biography

Václav Němec is a Czech scientist born on November 2, 1929, in Prague. He has been a founding member of the International Association for Mathematical Geosciences (IAMG) since 1968 and has contributed significantly to the dissemination of scientific knowledge of mathematical geology in Eastern Europe.

He started his study of economics at the Technical University of Prague in 1948. At that time, Czechoslovakia belonged to the so-called Eastern Bloc. In 1951, because of political reasons, he was not allowed to continue with the closing fourth academic year of his studies. The following 26 months, until the end of November 1953, he spent as a miner in a special Czechoslovak army unit for politically unreliable persons. From the end of 1953 to 1990, he was working for the Czechoslovakian geological exploration of mineral deposits service. As distant education student at Charles University in Prague, he completed his studies in the Earth sciences with a diploma in applied geophysics in 1959. He received his RNDr degree in 1967 from Charles University in economic geology with a doctoral thesis on computerized modeling of three mineral deposits exploited by the cement industry in Czechoslovakia. He received a second doctoral degree (analogous to a PhD) in 1987 from the Technical University of Košice (mining sciences) and in 1994 an engineering degree from the University of Economics in Prague as restitution for the injustice received in 1951.

His scientific work was initially focused on mathematical modeling of mineral resources in the cement industry and the investigation of planetary tectonic fault systems (Němec 1970, 1988, 1992). A stay from 1969 to 1970 as visiting research fellow at the Kansas Geological Survey (USA) was dedicated to the latter topic. As Eastern Treasurer of the IAMG, he contributed substantially to mathematical geology in 1968–1980 and 1984–1996. In this role, he organized 19 international conferences on mathematical geology as part of the annual Mining Příbram Symposia. During the East-West separation of the world until 1990, these conferences served as the most important scientific interface for mathematical geology between the politically separated hemispheres.

In addition to mathematical geology, he has devoted himself to geoethics that he introduced in 1991 as a new topic for

the Mining Příbram Symposium inspired by his wife Lidmila Němcová (Němec and Němcová, 2000). He convened five symposia on this topic at the 30th and 34th IGCs and at two outreach symposia at annual meetings of the European Union of Geosciences (EGU) in Vienna (2013–2014). He lectured as invited speaker in about 20 countries on five continents and became recognized as the “Father of Geoethics.”

In recognition of his scientific achievements in mathematical geology and his service to the IAMG, he received the IAMG’s highest award in 1991 – the William Christian Krumbein Medal. He is a member of the Russian Academy of Natural Sciences since 1995 and an active member or officer of various other national and international scientific and cultural NGOs. His scientific competence, his multilingual talent, and his musicality (as a pianist) make him an extraordinary ambassador between the nations not only in science but also in the field of culture.

Bibliography

- Němec V (1970) The law of regular structural pattern, its applications with special regard to mathematical geology. In: Merriam DF (ed) *Geostatistics – a colloquium*. Plenum, New York/London, pp 63–78
- Němec V (1988) Geomathematical models of ore deposits for exploitation purposes. *Sci de la Terre Sér Inf* 27:121–131
- Němec V (1992) Possible ways to decipher spatial distribution of mineral resources. *Math Geol* 24/8:705–709
- Němec V, Němcová L (2000) Geoethical backgrounds for geoenvironmental reclamation. In: Paithankar AG, Jha PK, Agarwal RK (eds) *Geoenvironmental reclamation (International Symposium, Nagpur, India)*. Oxford & IBH Publ. Co., New Delhi, pp 393–396, ISBN 81-204-1457-8

Neural Networks

Vladik Kreinovich

Department of Computer Science, University of Texas at El Paso, El Paso, TX, USA

Definition

A neural network is a general term for machine learning tools that emulate how neurons work in our brains.

Ideally, these tools do what we scientists are supposed to do: We feed them examples of the observed system’s behavior, and hopefully, based on these examples, the tool will predict the future behavior of similar systems. Sometimes they do predict – but in many other cases, the situation is not so simple.

The goal of this entry is to explain what these tools can and cannot do – without going into too many technical details.

How Machine Learning Can Help

What Are the Main Problems of Science and Engineering

The main objectives of science and engineering are:

- To determine the current state of the world.
- To predict the future behavior of different systems and objects.
- If it is possible to affect this behavior, to find the best way to do it.

For example, in geosciences:

- We want to find mineral deposits – oil, gas, water, etc.
- We want to be able to predict potentially dangerous activities such as earthquakes and volcanic eruptions.
- We want to find the best ways to extract minerals that would not lead to undesirable side effects such as pollution or triggered seismic activity.

Let us describe the three classes of general problems in precise terms.

To determine the current state of the world, we can use the measurement results x . Based on these measurement results, we want to describe the actual state y of the system. For example, to find the geological structure y in a certain area – e.g., the density (and/or speed of sound) values at different depths and at different locations, we can use the seismic data x , both passive (measuring seismic waves generated by actual earthquakes) and active (measuring seismic signals generated by specially set explosions or vibrations).

To predict the future state y of the system – or at least the future values y of the quantities of interest – we can use the information x about the current and past state of the system. For example, by using the accurate GPS measurements x , we can find how fast the continents drift, and thus predict their future location y .

To find the best control y , we can use all the known information x about the current state of the system. For example, based on our knowledge of the geological structure x of the area, we would like to find the parameters (e.g., pressure) y of the fracking technique that would leave the pollution below the desired level.

Sometimes We Know the Equations, Sometimes We Do Not

In some cases, we know the equations relating the available information x and the desired quantities y . In some of these cases, the relation is straightforward: For example, simple linear extrapolation formulas enable us to predict the future continent drift. In other cases, the equations are not easy to solve: For example, it is relatively easy, giving the density structure y , to describe how the seismic signals will propagate and what

signals x to expect – but to find y based on x (i.e., to solve the *inverse* problem) is often not easy.

In most cases, however, we do not know the equations relating x and y . In this, geosciences are drastically different from physics – especially fundamental physics – where the corresponding equations are usually known (they may be difficult to solve, as in predicting chemical properties of atoms and molecules from Schroedinger's equation, but they are known).

How Machine Learning Can Help: General Case

Usually, there are cases when we know both the input x and the desired output y . In other words, we know several pairs (x_i, y_i) corresponding to different situations.

Such cases are ubiquitous for prediction problems: For example, if we are trying to predict seismic activity a week ahead, then every time a week has passed, we have a pair (x_i, y_i) consisting of the measurement results x_i obtained before the passed week and the passed week's seismic activity y_i .

Such cases are ubiquitous in control problems: Every time we do something and succeed, we get a pair (x_i, y_i) consisting of the previous situation x_i and of the successful action.

Based on such pairs, machine learning tools produce an algorithm that, given the input x , provides an estimate for the desired output y . For example, for prediction problems, we can use the current values x to get some predictions of the future values y .

Producing such an algorithm usually takes time – but once the algorithm has been produced, it usually works very fast.

How Machine Learning Can Help: Case When We Know Equations

When we know equations, the difficulty is usually in solving the inverse problem – finding y based on x . In contrast, the “forward” problem – finding x based on y – is usually easy to solve. So, what we can do is select several realistic examples $y_1, \dots, y_n$ of y , compute the corresponding x 's $x_1, \dots, x_n$, and feed the resulting pairs $(x_1, y_1), \dots, (x_n, y_n)$ into a machine learning tool.

As a result, we get an algorithm that, given x , produces y – i.e., that solves the inverse problem.

Limitations

There Are Limitations

So far, it may have seemed that machine learning is a panacea that can solve almost all of our problems. But, of course, the reality is not always rosy. To effectively use machine learning, we need to solve three major problems:

- First, the ability of machine learning tools to process multi-D data is limited. In most cases, these tools cannot use *all* the measurement results that form the input x_i and the output y_i . So, we need to come up with a small number of informative characteristics. This selection is up to us, it is difficult, and if we do not select these characteristics correctly, we lose information – and the machine learning program will not be able to learn anything.
- Second, to make accurate predictions, we need to have a large number of pairs: thousands, sometimes millions. Sometimes we have many such pairs – e.g., when we are solving an inverse problem. But in many other important cases – e.g., in predicting volcanic eruptions or strong earthquakes – there are – thankfully – simply not that many such events. Of course, we can add to the days of observed eruptions days when nothing happened, but then the machine learning program will simply always predict “no eruption” – and be accurate in the spectacular 99.99% of the cases!
- Third, training a machine learning program requires a lot of time, so much that often it can only be done on a high-performance computer.

The need to solve these three problems – especially the first two – severely limits the usefulness of machine learning tools.

Let us explain, on a qualitative level, where these problems come from.

Machine Learning Is, in Effect, a Nonlinear Regression

Machine learning is not magic. It is, in effect, a nonlinear regression: Just like linear regression enables us to find the coefficients of a linear dependence based on samples, nonlinear regression enables us to find the parameters of a nonlinear regression. For neural networks, such parameters are known as *weights*.

In many practical situations – especially in geosciences – linear models provide a very crude approximation. So, crudely speaking, in addition to parameters describing linear terms, we also need parameters describing quadratic terms, cubic terms, etc. The more parameters we use, the more accurately we can describe the actual dependence. This is similar to how we can approximate, e.g., an exponential function $\exp(x) = 1 + x + \frac{x^2}{2!} + \dots$ by the first few terms in its Taylor expansion: $\exp(x) = 1 + x + \dots + \frac{x^m}{m!}$ (this is, by the way, how computers actually compute $\exp(x)$): the more terms we use, the more accurate is the result.

We Cannot Use All the Information

And here lies the problem. The numbers of nonlinear terms drastically grow with the number of inputs. For example, if the input x consists of m values $x = (X_1, \dots, X_m)$, then to describe a

generic linear dependencies $Y = c_0 + \sum_{i=1}^m c_i \cdot X_i$, we need $m + 1$ parameters. To describe a generic quadratic dependence $Y = c_0 + \sum_{i=1}^m c_i \cdot X_i + \sum_{i=1}^m \sum_{j=1}^m c_{ij} \cdot X_i \cdot X_j$, we already need $\approx m^2$ parameters. This already leads to a problem. For example, suppose that we are processing images. Describing an image x means describing the intensity X_i at each of its pixels i . A typical image consists of about $1000 \times 1000 = 10^6$ pixels, so here we have $m \approx 10^6$. It is easy to store and process a million values, but finding 10^{12} unknown coefficients – even in the simplest case, when we have a system of 10^{12} linear equations with 10^{12} unknown – is way beyond the abilities of modern computers.

As a result, we cannot simply feed the image into a machine learning tool – and, similarly, we cannot simply feed the seismogram into this tool. We need to select a few important parameters characterizing this image (or this seismogram). And here computers are not much help, we the scientists need to do the selection. This explains the first problem.

It should be mentioned that the situation is not so bad with images. Everyone knows that images can be compressed into a much smaller size without losing much information – e.g., a small-size photo on a webpage is still quite recognizable – and modern machine learning techniques use such methods automatically. However, for seismograms, no such no-information-loss drastic compressions are known.

What About Computation Time?

The more parameters we need, the more computation time we need to find the values of these parameters. Even if the system is linear, to find the values of q parameters – i.e., to solve a system of q equations for q unknowns – we need time $\approx q^3$, at least as much as we need to multiply two $q \times q$ matrices A and B – where to compute each of q^2 elements of the product $c_{ij} = a_{i1} \cdot b_{1j} + \dots + a_{iq} \cdot b_{qj}$, we need q computational steps (and $q^2 \cdot q = q^3$).

Even if we compress the image from a million to $m = 300$ values, if we take the simplest – quadratic – terms into account, we will need $q \approx m^2 = 10^5$ variables, so we need at least $q^3 \approx 10^{15}$ computational steps. On a usual GigaHertz PC that performs 10^9 operations per second, this means 10^6 seconds – about 2 weeks. And we only took into account quadratic terms – and just like an exponential function or a sinusoid do not look like graphs of x^2 , real-life dependencies are not quadratic either. So machine learning requires a lot of computation time – often so much time that only high-performance computers can do it. This explains the third problem.

We Need a Large Number of Samples

And this is not all. The more accurately we want to predict y , the more parameters we need. How many samples do we need? Crudely speaking, each pair (x_i, y_i) with $y_i = (Y_{i1}, \dots, Y_{ir})$ provides r equations $Y_{ij} = f(X_{i1}, \dots, X_{iq})$ for determining the unknown parameters, for some small r . So, we need approximately as many samples as parameters. In the above example of $q \approx 10^5$ unknowns, we need hundreds of thousands of example – and usually, even more. This explains the second problem.

There is an additional reason why we need many pairs. The reason is that it is not enough to just train the tool, we need to test how well the trained machine learning tool works. For that purpose, the usual idea is to divide the original pairs into the training set and the testing set, train the tool on the training set only, and then test the resulted training on the training set.

How do we know that it works well? One correct prediction may be a coincidence. However, if we have several good predictions, this makes us confident that the trained model works well. So, to gain this confidence, we need a significant number of such pairs. And in many important problems, we do not have that many pairs. For example, even if a volcano has been very active, had five eruptions, there is not enough data to confirm the model.

An Additional Problem

All the above arguments assumed that once we arranged an appropriate compression, gathered millions of sample, and rented time on high-performance computer, the machine learning tool will always succeed. Unfortunately, this is not always the case.

It is known that, in general, prediction problems, inverse problems, etc. are NP-hard, which means that (unless $P = NP$, which most computer scientists believe to be impossible) no feasible algorithm is possible that would always solve these problems. In other words, any feasible algorithm – and algorithms implemented in the machine learning tools are feasible – will sometimes not work.

Good news is that computer scientists are constantly inventing new algorithms, new tools. So if you encounter such a situation, a good idea is to team up with such a researcher. Maybe his/her new algorithm will work well when the algorithms from the software package you used did not work.

This brings us to the question of what machine learning tools are available.

What Tools Are Available

Traditional neural networks used three layers of neurons. Corresponding the number of parameters was not high, so while such neural networks can be easily implemented on a usual PC, they are not very accurate. So, if you have a few samples – not enough for more accurate training – you can use traditional neural networks, and get some reasonable (but not very accurate) results.

In the latest decades, the most popular are deep neural networks that have up to several dozen layers, and thus a much larger number of adjustable parameters. Often, they lead to spectacular results and accurate predictions. However, due to the high number of parameters, deep neural networks need a large (sometimes unrealistically large) number of sample pairs. For the same reason, deep neural networks require a lot of time to train – so much that this training is rarely possible on a usual computer. So, if you have a large number of samples – and you know how to compress the original information – it is a good idea to try to use deep learning.

There are also other efficient machine learning tools, such as support vector machine (SVM) – that used to be the most efficient tool until deep learning appeared – but these tools are outside the scope of this article.

In Case You Are Curious

So how do neural networks work? An artificial neural network consists of *neurons*. Each neuron takes several inputs $v_1, \dots, v_k$ and returns an output $u = s(w_0 + w_1 \cdot v_1 + \dots + w_k \cdot v_k)$, where w_i are coefficients (“weights”) that need to be determined during training, and $s(z)$ is an appropriate non-linear function. Traditional neural networks used the so-called sigmoid functions $s(z) = 1/(1 + \exp(-z))$, while deep neural networks use the “rectified linear” function $s(z) = \max(0, z)$.

Neurons usually form layers:

- Neurons from the first layer use the measurement results (or whatever we feed them as x_i) as inputs.
- Neurons from the second layer use outputs of the first layer neurons as inputs, etc.

How are the coefficients w_i trained? In a nutshell, by using *gradient descent* – the very first optimization method that students learn in their numerical methods class. The main idea behind this method is very straightforward: to go down the mountain as fast as possible, you follow the direction in which the descent is the steepest. Neural networks use some clever algorithmic tricks that help find this steepest direction, but they *do* use gradient descent.

Readers interested in technical details are welcome to read (Goodfellow et al. 2016) – at present (2021) the main textbook on deep neural networks.

Examples of Successful Applications

Successful applications of *traditional* (three-layers) neural networks have been summarized in a widely cited book (Dowla and Rogers 1996). There have been many interesting applications since then, especially in petroleum engineering, where even a small improvement in prediction accuracy can lead to multimillion gains. The paper (Thonhauser 2015) provides a survey of such applications. For example, neural networks are efficient in classifying geological layers based on the well log (Bhatt and Helle 2002) – this is one of the cases when we have a large amount of data, sufficient to train a neural network.

Applications of *deep learning* are a rapidly growing research area. There is a large number of papers, there are surveys – e.g., (Bergen et al. 2019). However, new applications and new techniques appear all the time – this research area grows faster than surveys can catch up. Because of this fast growth, this part of the article is more fragmentary than comprehensive – we will just cite examples of typical applications.

Probably the most active application areas are processing images – e.g., satellite images. One of the main reasons why these applications of deep learning are most successful is that, as we have mentioned, for images we can apply efficient almost-no-loss compression and thus, naturally prepare the data for neural processing. A typical recent example of such an application is (Ullo et al. 2019).

One of the most interesting applications of deep learning to satellite images is the possibility to predict volcanic activities based on images produced by satellites equipped with interferometric synthetic aperture radar (InSAR) – which can detect centimeter-scale deformations of earth’s surface (Gaddes et al. 2019).

Many examples of applications are related to solving inverse problems – since, as we have mentioned, in this case, we can easily generate a large number of examples; see, e.g., (Araya-Polo et al. 2018), (Mosser et al. 2018). The use of state-of-the-art machine learning techniques for solving inverse problem has already led to interesting discoveries. For example, it turns out that, contrary to the previously assumed simplified models of seismic activity – according to which only usual (“fast”) earthquakes release the stress, a significant portion of the stress is released through slow slip and slow earthquakes; see, e.g., (Pratt 2019).

Interestingly, sometimes even for the forward problem, neural networks produce results faster than the traditional numerical techniques; see, e.g., (Moseley et al. 2018).

The papers (Magaña-Zook and Ruppert 2017) and (Linville et al. 2019) use deep learning to solve another important problem: How to distinguish earthquakes from explosions based on their seismic waves? This problem was actively developed in the past because of the need to distinguish nuclear weapons tests from earthquakes. Now, the main case study is separating small earthquakes from small explosions like quarry blasts – and in this application, there is a large number of examples, which makes neural networks very successful.

Summary

In a nutshell, neural networks are interpolation and extrapolation tools. When we have a large number (x_i, y_i) of pairs (x, y) of related tuples of quantities, machine learning techniques – such as neural networks – produce a program that, given a generic tuple x , estimates y .

In many cases, neural networks have led to successful applications in geoscience. However, neural networks are not a panacea.

For a neural network application to be successful, we really need a very large number of examples – which is not always possible in geosciences; we need to compress the data without losing information – which is also often difficult in geosciences; we usually need to spend a large amount of computation time; and we need to be patient: Sometimes these techniques work, sometimes they do not.

We hope that this entry will help geoscientists to select problems in which all these conditions are satisfied, and get great results by applying neural networks! (And do not hesitate to collaborate with computer scientists if it does not work the first time around.)

Acknowledgments This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science) and HRD-1242122 (Cyber-SHARE Center of Excellence).

References

- Araya-Polo M, Jennings J, Adler A, Dahlke T (2018) Deep-learning tomography. *Lead Edge* (Tulsa Okla) 37:58–66
- Bergen KJ, Johnson PA, de Hoop MV, Beroza GC (2019) Machine learning for data-driven discovery in solid earth geoscience. *Science* 363(6433) Paper eaau0323
- Bhatt A, Helle HB (2002) Determination of facies from well logs using modular neural networks. *Pet Geosci* 8:217–228
- Dowla FU, Rogers LL (1996) Solving problems in environmental engineering and geo-sciences with artificial neural networks. MIT Press, Cambridge, MA
- Gaddes ME, Hooper A, Bagnard M (2019) Using machine learning to automatically detect volcanic unrest in a time series of interferograms. *J Geophys Res Solid Earth* 124(11):12304–12322

- Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge, MA
- Linville L, Pankow K, Draelos T (2019) Deep learning models augment analyst decisions for event discrimination. *Geophys Res Lett* 46(7): 3643–3651
- Magaña-Zook SA, Ruppert SD (2017) Explosion monitoring with machine learning: A LSTM approach to seismic event discrimination. In: Abstract of the American Geophysical Union (AGU) Fall 2017 Meeting, New Orleans, Louisiana, December 11–15, 2017, Abstract S43A–0834
- Moseley B, Markham A, Nissen-Meyer T (2018) Fast approximate simulation of seismic waves with deep learning. In: Proceedings of the 2018 conference on Neural Information Processing Systems NeurIPS'2018, Montreal, Canada, December 3–8, 2018
- Mosser L, Kimman W, Dramsch J, Purves S, De la Fuente A, Ganssle G (2018) Rapid seismic domain transfer: Seismic velocity inversion and modeling using deep generative neural networks. In: Proceedings of the 80th European Association of Geoscientists and Engineers (EAGE) conference, Copenhagen, Denmark, June 11–14, 2018
- Pratt SE (2019) Machine fault. *Earth & Space Science News* (EoS), December 2019, 28–35
- Thonhauser G (2015) Application of artificial neural networks in geoscience and petroleum industry. In: Craganu C, Luchian H, Breaban ME (eds) Artificial intelligent approaches in petroleum geosciences. Springer, Cham, pp 127–166
- Ulló SL, Langenkamp MS, Oikarinen TP, Del Rosso MP, Sebastianelli A, Piccirillo F, Sica S (2019) Landslide geohazard assessment with convolutional neural networks using Sentinel-2 imagery data. In: Proceedings of the 2019 IEEE International Geoscience and Remote Sensing Symposium IGARSS'2019, Yokohama, Japan, July 28 – August 2, 2019, 9646–9649

Nonlinear Mapping Algorithm

Jie Zhao¹, Wenlei Wang², Qiuming Cheng³ and Yunqing Shao³

¹School of the Earth Sciences and Resources, China University of Geosciences, Beijing, China

²Institute of Geomechanics, Chinese Academy of Geological Sciences, Beijing, China

³China University of Geosciences, Beijing, China

Definition

Nonlinear mapping algorithm, firstly proposed by Sammon (1969), is a nonhierarchical method of cluster analysis for geometric image dimensionality reduction (Howarth 2017). According to the algorithm, complex relationships in high-dimensional space can be transformed and demonstrated in a low-dimensional space without changing relative relationships among internal factors. Approximate images of the relationship between high-dimensional samples can be intuitively seen in the low-dimensional space that will greatly simplify the interpretation (Sammon 1969). Nonlinear mapping algorithm has been widely applied in target classification and recognition within the research fields of geology,

chemistry, and hydrology, especially for distributions and processes with nonlinearity, for examples, applications in flood forecasting, fault sealing discrimination (Smith 1980), etc. The algorithm focuses only on data without considering types and categories, and has been efficiently used in prediction of mineralization and natural disasters.

Introduction

Observational data indicative to natures of the Earth system are often multidimensional (e.g., stream sedimentary geochemical samples recording concentrations of various elements; structure data in a form of vector consisting of length, azimuth, age, strain, stress, etc.). Consequently, researchers working in the Earth sciences, especially mathematical geosciences, often needs to deal with high-dimensional data and sometimes even faces issues of information redundancy. The so-called “curse of dimensionality” is a vivid description of complicated calculation and interpretation brought by high-dimensional data. It may also lead to obstacle of analytical method, for example calculating the inner product. Dimensionality reduction is a good way to solve the problems by transforming original high-dimensional data into a low-dimensional “subspace” where the sample density is increased and distance calculation will become simpler and more convenient. Classic methods including principal component analysis (PCA) and linear discriminant analysis (Belhumeur et al. 1996) with systematic and theoretical foundation have been well practiced in various geological applications, which assume linear data structures are existed in the high-dimensional data. However, inherited from nonlinear geo-systems, relationships among geo-variables often present nonlinearity. Utilization of linear methods to investigate these geo-variables will mislead the interpretation of nonlinear geo-processes. The proposed nonlinear mapping technique proposed by Sammon (1969) has triggered a big progress in understanding the world.

Within a p -dimension space R^p , suppose N samples with P variables can be denoted by $X_i = [x_{i1}, x_{i2}, \dots, x_{ip}]$, $i = 1, 2, \dots, N$. If the N points in the R^p space are nonlinearly mapped in a lower dimensional space R^l ($l < p$): N samples will be $Y_i = [y_{i1}, y_{i2}, \dots, y_{il}]$, $i = 1, 2, \dots, N$. The distances or relations among samples should be approximate in both spaces of R^p and R^l (d_{ij} and d_{ij}^*). A critical variable E , termed as mapping errors, is defined by total sample distances in spaces of R^p and R^l ($\sum d_{ij}$ and $\sum d_{ij}^*$) to constrained the transformation:

$$E = \frac{1}{NF} \sum_{i \in j} w_{ij} (d_{ij}^* - d_{ij})^2 \quad (1)$$

where, $NF = \sum_{i < j} d_{ij}^* \sum_{i=1}^{N-1} \sum_{j=i+1}^N d_{ij}^*$ is the normalized factor and $W_{ij} = 1/d_{ij}^*$ is the weighting factor.

$$d_{ij}^* = \sqrt{\sum_{i=1}^p [x_{ki} - x_{kj}]^2} \quad (2)$$

$$d_{ij} = \sqrt{\sum_{k=1}^l [y_{ki} - y_{kj}]^2} \quad (3)$$

The dimensionality reduction implemented by nonlinear mapping algorithm applies the variable E as the weight coefficient to minimize difference between the $\sum d_{ij}$ and $\sum d_{ij}^*$. The steepest descent method can iteratively reduce the total mapping error through adjusting coordinates of points. In general, a well mapping result will be obtained with less than a hundred iterations. When the iteration results become convergent, variations of errors should not be less than the “mapping tolerance” (often set as 0.000001). A result with less mapping errors can better preserve topological structures of samples in the original space for acquiring geometric configuration of samples in the new space.

Manifold Learning and Basic Algorithms

Manifold learning method proposed by Roweis and Saul (2000) and Tenenbaum et al. (2000) has gradually become a heat point in studies of feature extraction. Manifold learning is another category of dimensionality reduction methods. Manifold is a homeomorphous space with Euclidean space at the local scale. In other words, the Euclidean distance can be used for distance measurement at the local scale. It means when the low-dimensional manifold is embedded in the high-dimensional space, although the distribution of samples in the high-dimensional space seems extremely complicated, the properties of Euclidean space still locally stand true for those samples. Thus, the dimension reduction mapping relations between the high-dimensional and the low-dimensional spaces can be easily established at the local scale. Typical algorithms include spectrum-based algorithms, isometric mapping algorithm (Tenenbaum et al. 2000), local linear embedding algorithm (Rowies et al. 2000), local tangent space alignment (Zhang 2004), kernel principal component analysis (Schölkopf et al. 1997), Laplacian feature mapping (Belkin and Niyogi 2002), Hessian feature mapping (Donoho and Grimes 2003), etc. Here will introduce three famous and popular learning methods.

Isometric Mapping Algorithm (ISOMAP)

The basic assumption of isometric mapping algorithm (Tenenbaum et al. 2000) is when the low-dimensional manifold is embedded in the high-dimensional space, estimated straight-line distances in the high-dimensional space are misleading, since the distances cannot find their equivalence in the embedded low-dimensional manifold. The fundamental thinking of the algorithm is to approximate geodesic distances among samples according to local neighborhood distance; meanwhile, based on systematical derivation of these distances, the problem of coordinates calculation in the embedded low-dimensional manifold can be changed to the issues of eigenvalues of matrices. The implementation of isometric mapping algorithm including three steps: (1) k neighbor (the first k data closest to the data point) or ε neighbor data (all data whose point distance is less than ε) of each sample in the high-dimensional spatial data set $\{x_i\}_{i=1}^N$ need to be defined and connected to construct the weighted neighborhood figure of the high-dimensional data; (2) the shortest distances between each pair of data points within the neighborhood figure are calculated, which are further defined as an approximate geodesic estimation; and (3) the multidimensional scaling algorithm is applied to reduce the dimensionality of the original data set.

The advantages of ISOMAP are obvious. With simple parameter settings, it provides a fast and globally optimal nonlinear mapping by calculating the geodesic distance, rather than the Euclidean distance between points in high-dimensional space. However, searching for the shortest distances between each sample pair is quite time-consuming when the sample volume is too large. The ISOMAP algorithm has been successfully employed in many research fields.

Local Linear Embedding Algorithm (LLE)

Local linear embedding proposed by Roweis and Saul (2000) is another nonlinear dimensionality reduction method based on the think of manifold learning. Suppose the data lies on or near a smooth nonlinear manifold of lower dimensionality, the main idea of the algorithm is to approximate global nonlinear structure with locally linear fits. During this processes, local linear relationships between original neighborhoods are maintained, and the intrinsic local geometric properties remain invariant to the local patches on the manifold. Its specific steps of LLE include:

1. Assign neighbors to each data points, for example using the k nearest neighbors.

2. Calculate the local linear reconstruction weight *matrix* W of the data from its neighbors by solving the following constraint least-square problem:

$$\varepsilon(W) = \sum_i \left| X_i - \sum_j W_{ij} X_j \right|^2 \quad (4)$$

- Then, minimize $\varepsilon(W)$ subject to (a) $\sum_j W_{ij} = 1$, and (b) $W_{ij} = 0$, if X_j is not belong to the set of neighbors of X_i .

3. Take the parameter of local linear representation as the invariant eigenvalue of high-dimensional and low-dimensional data, and calculate the unconstrained optimization problem to obtain the dimensionality reduction result Y :

$$\min_Y \phi(Y) = \sum_i \left| Y_i - \sum_j W_{ij} Y_j \right|^2 \quad (5)$$

The advantage of LLE algorithm is to effectively record the inherent geometric structure of the data by utilizing the local linear relationship. Compared to the isometric mapping algorithm, it seeks for eigenvalue solving of sparse matrix and not requires iteration, which makes it more efficient in calculation.

Rather than the geodesic distance relationship between sample points, the local linear embedding algorithm only maintains the local neighbor relationship, so it is not well descriptive of the manifold structure of data with equidistant characteristics. It should be noticed that this algorithm is more sensitive to noises since it assumes the samples are uniformly distributed.

Kernel Principal Component Analysis Algorithm (KPCA)

Geological processes often present nonlinear characteristics. As a linear algorithm, PCA has been widely used in geosciences, but it may still fail to depict nonlinear structures embraced in geo-datasets, and unable to implement dimensionality reduction of geological characteristics, effectively. As an upgrading version, KPCA proposed by Schölkopf et al. (1997) is a nonlinear form of PCA. By introducing an integral operator kernel functions, KPCA allows to map low-dimensional space to high-dimensional space via nonlinear mapping and then executes PCA in this feature space. The intuitive idea of kernel PCA is to transform linear indivisible

problems in input space to linear divisible problem in feature space.

Suppose we first map the data in an original space R^n into a new feature space F via nonlinear mapping function φ :

$$\varphi : R^n \rightarrow F \quad (6)$$

$$x \rightarrow \varphi(x).$$

The Euclidean distance in the new feature space F can be expressed as:

$$d_n(x_1y) = \sqrt{|p(x) - \rho(y)|} \\ - \sqrt{\phi(x) \cdot \rho(x) - 2\rho(x) \cdot \varphi(y) + \varphi(y) \cdot \varphi(y)} \quad (7)$$

According to the Mercer's theorem of functional analysis, the dot product in the original space can be represented by the Mercer kernel function in the new feature space F :

$$K(x, y) = (\varphi(x) \cdot \varphi(y)) \quad (8)$$

Equations (7) and (8) can be expressed as:

$$d_H(x, y) = \sqrt{K(x, x) - 2K(x, y) + K(y, y)} \quad (9)$$

Obviously, if the kernel function $K(x, y)$ is given, by which the data can be implicitly mapped to a high-dimensional feature space. Consequently, the data are linearly separable in the high-dimensional feature space. Since the inner product in the feature space is given by the kernel function, the calculation of which will no longer be confined by the dimensionality.

The kernel function plays a vital role in the kernel method. As long as the kernel function is selected, a certain mapping is then defined. The only requirement on the kernel function is that it should meet requirements of Mercer's theorem for functional analysis. It means if the kernel function is a continuous kernel of a positive integral operator, then it performs a dot product during the mapping. Commonly used kernel functions are listed as follows, and the results will be varied according to the kernel function selected.

1. Polynomial kernel function:

$$K(x, y) = (a(x \cdot y) + b)^d \quad d > 0, \text{ and } a, b \in R \quad (10)$$

2. Gaussian radial basis kernel function:

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \sigma \in R \quad (11)$$

3. The Sigmoid function:

$$K(x, y) = \tanh(a(x \cdot y) - b) a, b \in R \quad (12)$$

In comparison with other nonlinear algorithms, kernel PCA demonstrates its specific advantages: (1) no nonlinear optimization should be considered; (2) simple as PCA, its calculation is still focusing on solving eigenvalues; and (3) the number of preserved PCs is not necessary to be determined prior to the calculation. Due to the superiority of kernel PCA in feature extraction, it has been applied in various fields, e.g., face recognition, image feature extraction, and fault detection. From its research status, KPCA with mature algorithms employed for nonlinear feature extraction has been ideally applied in many fields.

Summary or Conclusions

Nowadays, research topics relating to nonlinear mapping are still in development. Something worth to note are:

1. Nonlinear mapping algorithm applied to geological data has advantages in removal the information redundancy of high-dimensional data, intuitive expression of data distribution patterns, and statistical significance than that of linear mapping. However, attention should be paid to two-dimensional interpretation which has possibility of failure for cases in high dimensions. Therefore, delicate interpretation and discussion at multidimensions are more appropriate.
2. Almost all algorithms have both advantages and defects, due to their complicated theoretical foundations. Some of them had been further developed, but they may still not be able to interpret the nonlinear data comprehensively. Determination of neighborhood size and algorithm upgrading for reducing noise interference and enhancing robustness will be the potential focuses in the further.

Bibliography

- Belhumeur PN, Hespanha JP, Kriegman DJ (1996) Eigenfaces vs. fisherfaces: recognition using class specific linear projection. In: European conference on computer vision. Springer, Berlin/Heidelberg, pp 43–58
- Belkin M, Niyogi P (2002) Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Advances eural information processing systems, pp 585–591
- David LD, Carrie G (2003) Hessian eigenmaps: new locally linear embedding techniques for high dimensional data, proc. Natl Acad Sci 100:5591–5596
- Donoho DL, Grimes C (2003) Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. Proceedings of the National Academy of Sciences 100(10):5591–5596
- Howarth RJ (2017) Dictionary of mathematical geosciences. Springer, Cham
- Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500):2323–2326

- Sammon JW (1969) A nonlinear mapping for data structure analysis. *IEEE Trans Comput* 100:401–409
- Schölkopf B, Smola A, Müller KR (1997) Kernel principal component analysis. In: *International conference on artificial neural networks*. Springer, Berlin/Heidelberg, pp 583–588
- Smith DA (1980) Sealing and nonsealing faults in Louisiana Gulf Coast salt basin. *AAPG Bull* 64:145–172
- Tenenbaum JB, De Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* 290: 2319–2323
- Zhang Z (2004) Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM J Sci Comput* 26:313–338

Nonlinearity

Wenlei Wang¹, Jie Zhao² and Qiuming Cheng²

¹Institute of Geomechanics, Chinese Academy of Geological Sciences, Beijing, China

²China University of Geosciences (Beijing), Beijing, China

Definition

Nonlinearity is a broad term descriptive of complex natures of the Earth system. These signatures are not an artifact of models, equations, and simplified experiments but have been observed and documented in many investigations and surveys as diverse variables. Formally defined in “*Dictionary of Mathematical Geosciences*” (Howarth 2017), the concept of nonlinearity is “The possession of a nonlinear relationship between two or more variables. In some cases, this may be a theoretical assumption which has to be confirmed by experimental data. The converse is linearity.” In order to better recognize nonlinearity, the term linearity should be referred first, which denotes possession of linear or proportional relationships among variables and can create “a straight line” in Cartesian coordinates. In contrary, nonlinearity denoting disproportional relationships among variables often creates “curves” (not a straight line) in the coordinates, e.g., parabola of a quadratic function relation. However, discrimination on linearity and nonlinearity is not that simple within the context of complex systems and complexity science. Linearity as a special case of nonlinearity merely has a proportional relationship among variables that small causes (inputs) lead to small effects (outputs), while nonlinearity often displays deviation from this simple relationship. Nonlinear relationship can be varied dramatically, and its intrinsic mechanisms are governed by the nonlinear paradigm that one cause (input) can lead to more effects (outputs) and/or one effect (output) may be led by more causes (inputs). Causative factors of nonlinearity in the Earth system are often with interrelated, interacted, and intercoupling effects, by which small variation can result disproportional and even catastrophic influence on results or other variables. Many Earth scientists even stated

that the world is inherently nonlinear, and knowledge on nonlinearity and its causative mechanisms will significantly enhance our recognition to interactions among the lithosphere, hydrosphere, atmosphere, and biosphere that will consequently progress human capability of reasonable development, natural resource utilization, prediction and/or safeguard against serious geological disasters, and environment protection.

Introduction

In the community of the Earth sciences, one of significant missions is to understand natures of the Earth for better living and long-term development. The real world is a complicated system. For solving practical problems, mathematical models should be established and solved for simulating and interpreting natural phenomena in the system, respectively. Constrained by inadequate capability and means of cognition, before advent of nonlinearity, it was generally accepted that the nature could be explicated linearly by linear equations. In order to find the approximate solution to complex systems or geoprocesses, they were routinely simplified by eliding some secondary causes or decomposing into several simple subsystems. The transformation of a nonlinear problem into linear problems is termed as linearization, during which multiple linear approximation methods were often adopted as well. There is generally a parameter by which the nonlinear equation is expanded to obtain several or infinite linear equations. The approximate solution of nonlinear issues can consequently be achieved by solving these expanded linear equations.

However, the method of linear approximation is not always effective, since development of the Earth system accompanies with not only quantitative transition but also dramatic qualitative changes. It is obvious that solution or interpretation by linear approximation works effective for linear or approximately linear phenomena, but the cases with fundamental differences from linear phenomena are often powerless. As mentioned above, nonlinearity is a ubiquitous and objective attribute in the Earth system and its compositions (subsystems). In particular, diverse geological processes including plate movement, tectono-magmatic activities, earthquakes, mineralization, etc. with dramatic variations in temperature, pressure, stress, and strain are disproportionately caused and effected by complex interaction and intercoupling mechanisms. These geological (or formation) processes essentially demonstrate nonlinearity, and it is precisely because the existence of nonlinearity prevents the practical efficiency of classic mathematical methodologies.

Joint research and thinking of the Earth and nonlinear sciences have been developed for 100 years (Ghil 2019). The Earth system and its subsystems spanning wide spatial-temporal scales that generally demonstrate nonlinearity are dynamically developed and often unrepeatable,

unpredictable, and dimensionless. The Earth scientists have proposed some nonlinear ways of thinking about these nonlinear signatures. The complex Earth system pertains to geo-processes that are believed as deterministic and can be modeled with fractal and deterministic chaos (Turcotte 1997). Progressive development in theories of chaos (Lorenz 1963), fractal (Mandelbrot 1983), and dissipative structure has profound implications for simulation, predication, and interpretation of many subsystems of the Earth (Flinders and Clemens 1996; McFadden et al. 1985; Phillips 2003).

Chaos

From observations, it can be found that occurrences of numerous natural phenomena in the deterministic Earth system with nonlinearity seem stochastic. A deterministic system is governed by rules that have no randomness and nothing stochastic. The present state can determine the future state. However, the system can sometimes have aperiodic and apparently unpredictable behaviors, and its development or evolution is not simply repetitive motion to equilibrium. Chaos is a term descriptive of evolution or development with stochastic appearance in deterministic dynamic systems. The chaos theory (Lorenz 1963; Turcotte 1997) mainly analyzes evolutive characteristics of nonlinear systems that is the evolution or development process from disorder, order, to chaos with the increase of nonlinearity. It is the science to explore their causative processes and evolutions originated from nonlinear dynamic systems which are usually described by ordinary differential equation, partial differential equation, or iterative equation. Chaos possesses both ordered periodic and disordered quasi-periodic structures that manifest the inherent randomness of nonlinear dynamic systems. The chaos theory still in development currently focuses on studying the characteristics, orderliness, and evolutionary mechanism of chaotic behaviors in nonlinear dynamic systems.

Chaotic behaviors possess random appearance but are not fundamentally stochastic. During geo-processes, the evolution and development of geo-variables or components of the Earth system demonstrate both repulsion and attraction, simultaneously. The attraction causes these variables clustered in the vicinity of a collection, while the repulsion of the collection makes these variables neither too close nor separate that reach an equilibrium state. Then the repulsion may be further enhanced resulting that these variables will be attracted by other attractors. Iterations of this process constitute complex chaotic behaviors (Fig. 1). These behaviors or geo-processes are sensitive to initial conditions and small perturbations, and initial differences or causes of minor deviations tend to persist and grow over time. From the Lorenz equation (Eq. 1), even under a same initial condition, integral

curves of repeating computation will be gradually separated and become dissimilar as the computation time increases (Fig. 1). The butterfly effect is a figurative expression of this characteristic. Consequently, short-term weather condition can be predictable, but not for long-term forecast (Lorenz 1963).

$$\begin{cases} \frac{dx}{dt} = \sigma(y - x) \\ \frac{dy}{dt} = R_a x - y - xz \\ \frac{dz}{dt} = xy - bz \end{cases} \quad (1)$$

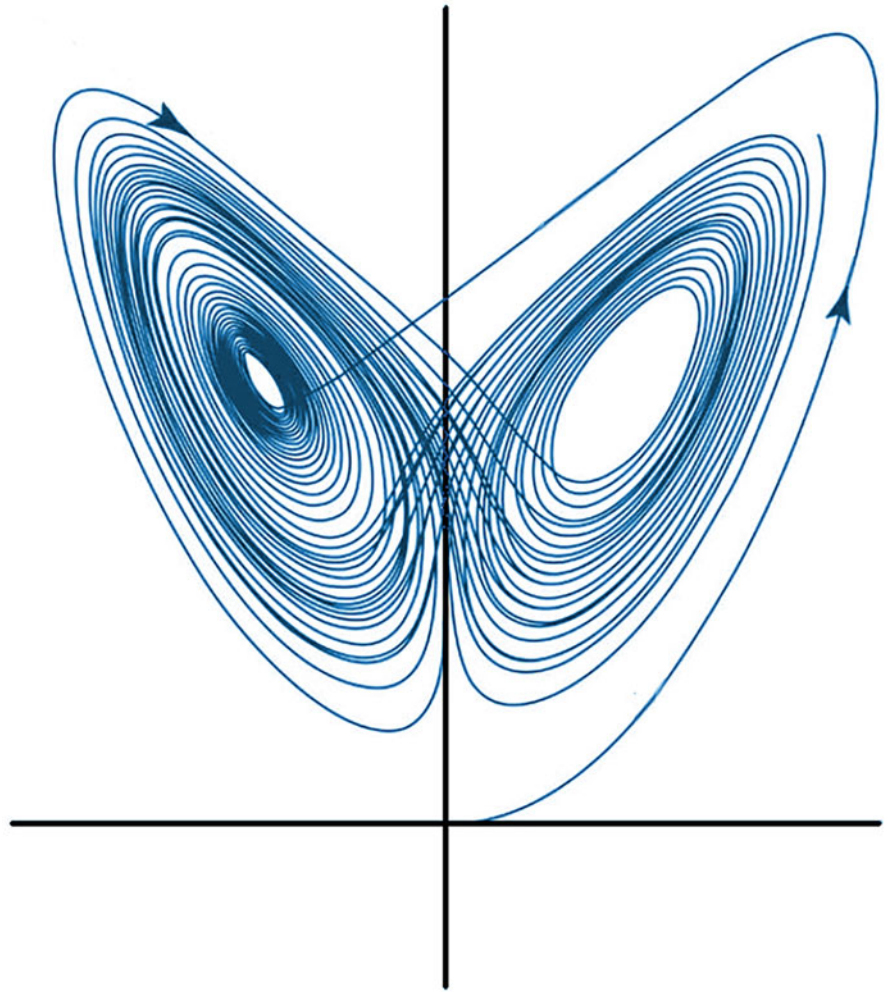
where, x , y , and z are proportional to the intensity of convection motion, the temperature difference between the ascending and descending currents, and the distortion of the vertical temperature profile from linearity, respectively; σ is Prandtl number; b is velocity damping constant; and R_a is Rayleigh number, a natural convection-related dimensionless number.

The application of chaos theory greatly expands human's horizon and thinking of recognition to nonlinearity of geo-processes (Turcotte 1997). Applying to investigate the convection caused by nonuniform heating during the raising of mantle can reveal a dynamic process from horizontal melting, multilayered convection, and turbulence (chaos) indicating that the Earth is an open thermal-dynamic system. The evolution process includes numerous nonlinear dynamic mechanisms that produce irregular (anomalous) and incoherent chaotic but stable structures. In the Earth system, many other chaotic behaviors or phenomena can be found in lithosphere evolution, mantle convection, geomorphic systems, geomagnetic reversal, etc. (Turcotte 1997).

Fractal

Many substances formed or produced in the nonlinear Earth system are often with irregular shapes (e.g., cloud, mountain, and coastline) and heterogeneously distributed quantities (e.g., element concentration, density, magnetism, etc.). Their irregular shapes different from continuously smooth curves and surfaces well described by classical differential geometry can be continuous but unsmooth or coarse with noninteger dimensions. Traditional mathematical methodologies like Euclidean geometry (e.g., 1, 2, and 3 dimensions for point, area, and cube, respectively) cannot measure these abnormal features. They were mentioned in classical mathematical literatures, but mostly treated as special and "weird" cases without delicate discussions. As progressive awareness of spatial structures (e.g., symmetry and self-similarity of crystal

Nonlinearity, Fig. 1 A structure in the chaos – when plotted in three dimension, the solutions to Lorenz equation fell onto a butterfly-shaped set of points



and crystal grown), it turns out that they may represent another geometric form of the real world.

In mathematical sense, fractal patterns or shapes (Fig. 2) can be generated by a simple iterative equation that is a recursion-based feedback system. Using the Koch curve as an example, a straight line is taken as the initial curve. It can then be reconstructed as shown in Fig. 2 that is termed as generator with $4/3$ length of the initial curve. Repeating this process for all line segments can update the curve constituted by line segments with a total length of $(4/3)^2$. Further iteration will produce a curve with an infinite length. The curve is continuous but nowhere can be differentiable (Schroeder 1991).

The fractal theory or fractal geometry proposed by Mandelbrot (1983) is to quantify objects with irregular geometry. Since the Earth substances with irregular shapes are ubiquitous in natural systems, it is so called as “geometry of nature” (Mandelbrot 1983). If a curve is equally divided into N straight-line segments with a length of r , its approximate length $L(r)$ can be evaluated by: $L(r) = Nr$. The estimated $L(r)$

will be a finite limit as the step (r) reducing to zero. For a fractal or more complex curve, the approximate length $L(r)$ will be infinite as r goes to 0; however, there is a critical exponent $D_H > 1$ which can keep the length $L(r) = Nr^{D_H}$ as finite (Schroeder 1991). The exponent D_H is called Hausdorff dimension:

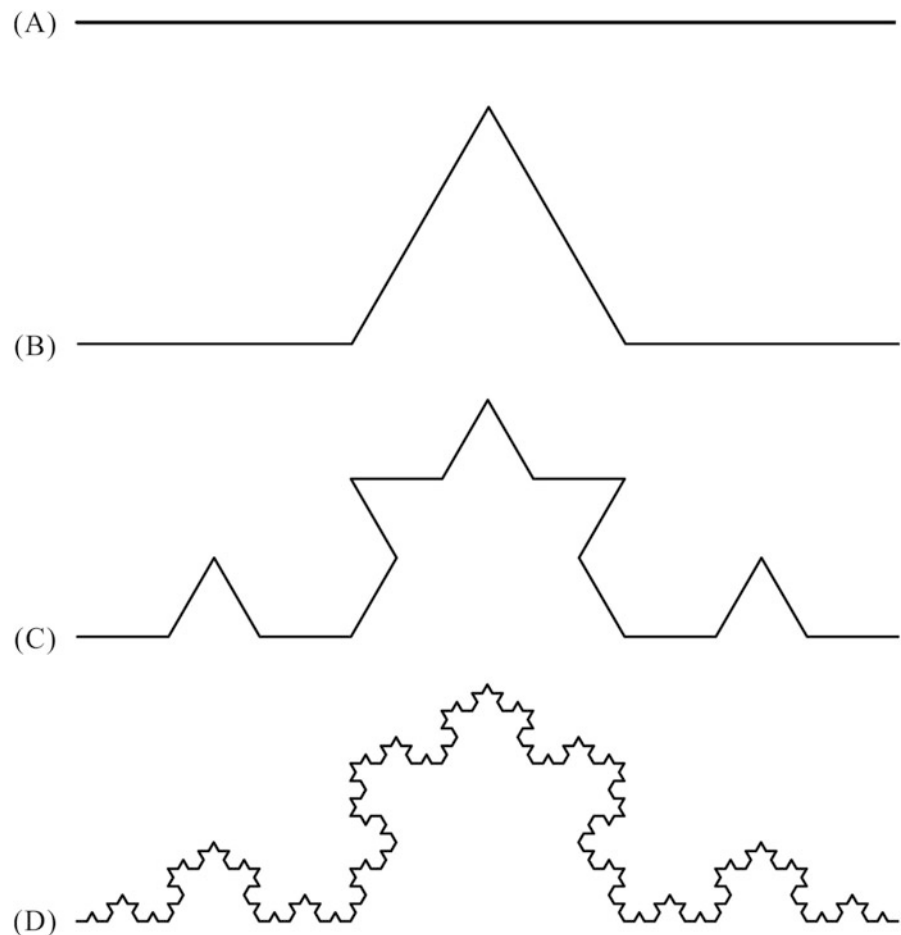
$$D_H = \lim_{r \rightarrow 0} \frac{\log N}{\log(1/r)} \quad (2)$$

When the above approximation is applied to the Koch curve (Fig. 2), choosing $r = r_0/3^n$ for the n^{th} generation, the number of segments N is proportional to 4^n .

$$D_H = \frac{\log 4}{\log 3} = 1.26 \dots \quad (3)$$

The approximation between 1 and 2 well demonstrates that the Koch curve is more irregular and complicated than a smooth curve but does not cover an area ($D_H = 2$). The Hausdorff dimension which can be fraction values is called fractal dimension in fractal geometry. Patterns, forms, or structures of the Earth substances with Hausdorff dimensions

Nonlinearity, Fig. 2 A: The initial curve; B: generator for the Koch curve; C: the next stage in the construction of the Koch curve; and D: high-order approximation to the Koch curve (D). After Schroeder (1991)



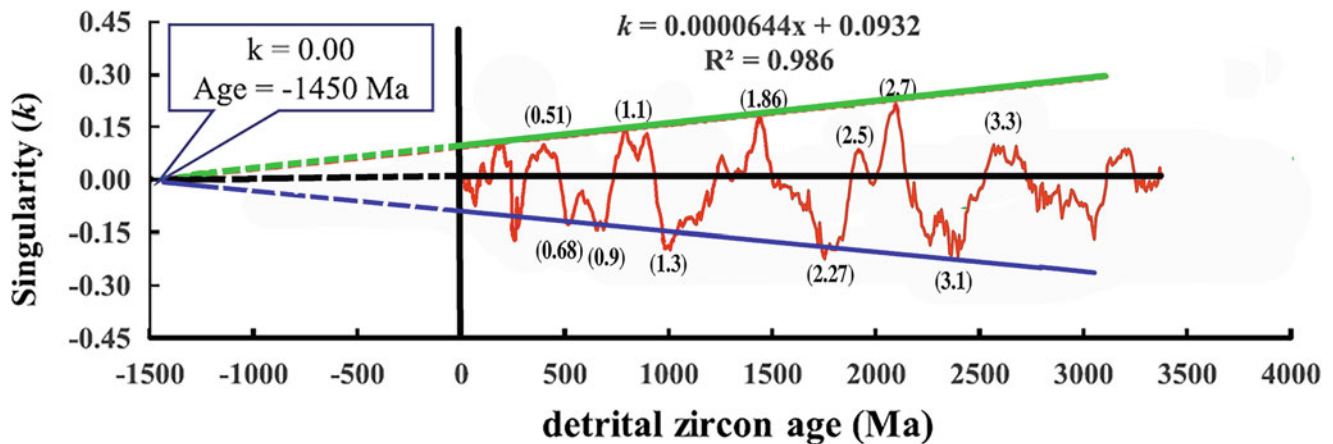
greater than their Euclidian dimensions are fractal, like the above Koch curve with a $D_H = 1.26$ between Euclidian dimensions of planar surface and volume.

Moreover, fractal possesses self-similarity, and its basic structure is scale invariant (Fig. 3). It can exhibit similar patterns as changing scales (Fig. 2). Fractal geometry is the science of complex geometry with self-similar structures to investigate intrinsic mathematical laws beneath the complex surface of nature. In practice, self-similarity is often not strictly similar, but for the case that any parts possess same statistical distributions as the whole is named as statistical self-similarity. Meanwhile, fractal theory mathematically linked to chaos and other forms of nonlinearity has been used as a morphometric descriptor of many forms and patterns produced by complex nonlinear dynamics. It has been broadly applied in the Earth sciences, especially in geology and geophysics (Turcotte 1997). In addition to classic geometry, it opens up a new way to investigate complex systems in support of simulation and explanation of nonlinear geological phenomena. The early applications were mainly to discover fractal structures and measure fractal dimensions of various geological features. Nowadays, it is more focusing on

geological processes and their indicative meanings interpreted from fractal dimension and multifractal spectrum of geological phenomena and geological bodies (Cheng 2017).

Selected Case Studies

The first selected case study of nonlinearity is chaotic processes discovered in magma mingling and mixing processes (Flinders and Clemens 1996). Inspired by the forming process of mixed patterns of two or more distinctly different fluids, the inherent deterministic chaos of magma systems was investigated through the mingling and mixing of enclave magmas with host magmas. During the chaotic process, migration of adjacent enclave blobs may be arbitrary, and these enclaves will experience different evolution. Data (e.g., crystal morphology, chemical features, and isotope) collected from microgranitoid enclaves within individual felsic magmas exhibits significant variations which may discover driving effects of nonlinear dynamics on the evolution of complex magma systems. Another example is the discovery of



Nonlinearity, Fig. 3 Singularity curve calculated from global databases of igneous and detrital zircon U-Pb ages, and the predication of the crust evolution by fractal-based models. Modified from Cheng (2017)

nonlinearity in geodynamo evidenced by palaeomagnetic data (McFadden et al. 1985). From observational palaeomagnetic data, the magnetic field in the core has a substantial and usually nonlinear effect on the velocity field, and the variation of westward drift rates can be interpreted according to nonlinear effects relating to the magnetic field intensity.

The last selected case study is on the predication of the Earth evolution (Cheng 2017). As mentioned above, knowledge on nonlinearity can enhance our capability in predication and explanation. In Cheng (2017), global databases of igneous and detrital zircon U-Pb ages were utilized to investigate episodic evolution of the crust. Fractal-based methodologies were developed to analyze shapes of the age curve with scale invariant fractality and singularity of the age records. It was discovered that the age peaks follow a power law distribution, and exponents estimated from the age curve and age peaks well demonstrate an episodic evolution of the crust. The obtained descending trend (Fig. 3) may be indicative of mantle cooling. Furthermore, following the trend of mantle cooling, it was predicted that the magmatic activities may vanish in 1.45Gyr.

Summary and Conclusions

Nonlinearity inherent from the origin and evolution of the Earth system is significant. After a century of efforts on understanding nonlinearity, theories and methodologies to interpret and explain nonlinear phenomena have greatly progressed. Nowadays, we do not only discover nonlinearity from phenomenological observation but also can reveal nonlinearity from intrinsic relations of data. There are effective

methods such as chaos and fractal to study evolutionary mechanism of nonlinear behaviors and irregular shapes caused by nonlinear processes. As a fundamental scientific issue, prediction of future development or evolution of the Earth system is an important goal of mathematical geosciences. The Earth system is a giant open nonlinear system whose variation is mutual transformation among non-stationary, deterministic, and random processes. For nonlinearity at different spatial-temporal scales, applying chaos, fractal, and other methods to construct new predication theory and methodology will be the main content of future research.

Bibliography

- Cheng Q (2017) Singularity analysis of global zircon U-Pb age series and implication of continental crust evolution. *Gondwana Res* 51: 51–63
- Flinders J, Clemens JD (1996) Non-linear dynamics, chaos, complexity and enclaves in granitoid magmas. *Trans R Soc Edinb Earth Sci* 87: 217–224
- Ghil M (2019) A century of nonlinearity in the geosciences. *Earth Space Sci* 6:1007–1042
- Howarth RJ (2017) *Dictionary of mathematical geosciences*. Springer, Cham
- Lorenz EN (1963) Deterministic nonperiodic flow. *J Atmos Sci* 20: 130–141
- Mandelbrot BB (1983) *The fractal geometry of nature*. Freeman, Francisco
- McFadden PL, Merrill RT, McElhinny MW (1985) Non-linear processes in the geodynamo: palaeomagnetic evidence. *Geophys J Int* 83: 111–126
- Phillips JD (2003) Sources of nonlinearity and complexity in geomorphic systems. *Prog Phys Geogr* 27:1–23
- Schroeder M (1991) *Fractals, chaos, power laws: minutes from an infinite paradise*. Freeman, New York
- Turcotte DL (1997) *Fractals and chaos in geology and geophysics*. New York, Cambridge

Normal Distribution

Jaya Sreevalsan-Nair

Graphics-Visualization-Computing Lab, International
Institute of Information Technology Bangalore, Electronic
City, Bangalore, Karnataka, India

Definition

Normal distributions have been widely used in natural and social sciences to approximate unknown continuous distributions of observed or simulated data. Real world data is real-valued and is handled using random variables which are variables whose values depend on outcomes of random processes. It is a parametric distribution, defined by two parameters, namely, its mean μ and its standard deviation σ . It is represented as $\mathcal{N}(\mu, \sigma)$, whose probability density function is given by

$$\phi(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty,$$

giving the characteristic bell curve shape, as shown in Fig. 1.

Normal distribution is unimodal, i.e., it is a distribution with a unique mode, where mode ($\mathcal{N}(\mu, \sigma)$) = μ .

The standard normal distribution is a special case of $\mathcal{N}(0, 1)$, as shown in Fig. 1. A *non-standard* normal distribution can be converted to its *standard* form, using standardized values called Z-score of the random variable, given by $Z_i = \frac{X_i - \mu}{\sigma}$. Computing probabilities for different intervals of random variables is done by converting them to their z-scores and looking up probabilities of the z-scores from the Standard Normal Distribution table. Computing probabilities of such clearly defined events correspond to direct problems. Computing the value of the random variable using its probability corresponds to inverse problems, which are also solved in the case of normal distributions by using its standardized form and the lookup table.

Normal distributions are significant due to the Central Limit Theorem (CLT). For a set of independent and identically distributed (i.i.d.) random variables in real-valued data, $X_1, X_2, \dots, X_n$, each with the same mean μ and standard deviation σ , CLT states that as n grows, i.e., $n \rightarrow \infty$, the distribution of the average of the set $\bar{X}_n$ converges to the normal distribution $\mathcal{N}(\mu, \sigma^2/n)$. This implies that, for the average value, $\lim_{n \rightarrow \infty} Z_n = \sqrt{n} \cdot \left(\frac{\bar{X}_n - \mu}{\sigma} \right)$ is the limiting case of the standard normal random variable.

Overview

Gaussian distribution is another name for the normal one, as K. F. Gauss has been attributed the origin of the normal

distribution through his work in 1809 (Gauss 1963). However, the history of normal distribution dates back to the eighteenth century (Fendler and Muzaffar 2008; Kim and Shevlyakov 2008; Stigler 1986), starting out as a way to represent binomial probability. If binomial variables represent a special case of the set of random variables, where all X_i have Bernoulli distribution with a parameter p , then, for large n and moderate values of p , ($0.05 \leq p \leq 0.95$), based on CLT,

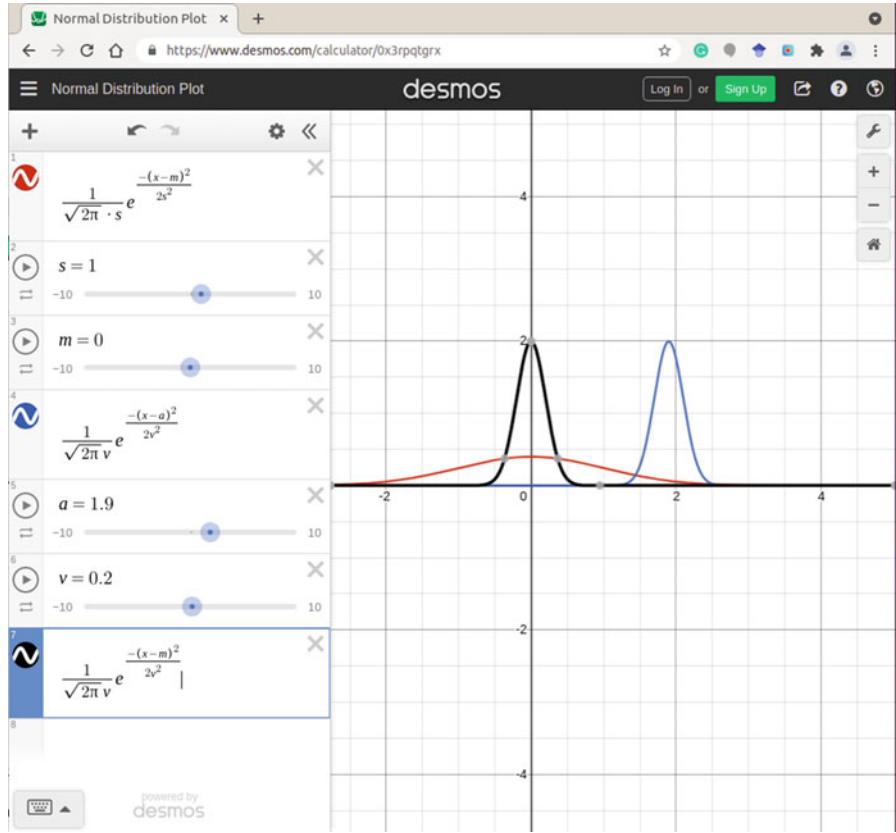
$$\text{Binomial}(n, p) \approx \mathcal{N}\left(\mu = np, \sigma = \sqrt{np(1-p)}\right).$$

In the year 1733, Abraham de Moivre had determined the normal distribution for a special case of $p = 0.5$, as a limit of a coin-tossing distribution (Pearson et al. 1926; Stigler 1986). de Moivre had provided the first treatment of the probability integral. This was improved by P-S Laplace in 1781 by generalizing for $0 < p < 1$ (Laplace 1781; Kim and Shevlyakov 2008). Gauss in 1809 popularized the derivation of the normal distribution, but not as a limiting form of the binomial distribution. He used the normal distribution to explain the distribution of errors, thus leading to the distribution becoming eponymous to him (Gauss 1963). The use of normal distribution for errors made it appropriate for modeling errors in the linear least-squares model (LS) by Gauss. Since LS is solved using a system of equations, called as normal equations, the name “normal” is associated with the distribution itself. In short, it is widely believed that Laplace, Gauss, and a few other scientists had arrived at the normal distribution through their independent work (Stigler 1986).

The study of the theory of errors played a key role in developing statistical theory and improved the applicability of normal distribution in different domains. As the applications of normal distribution shifted from scientific to societal ones, its use began to popularize the notion of the average man in 1846 by Belgian scientist, Adolphe Quetelet, as a representative of the human population. The notion of average man perpetuates the deceptive characterization of the distribution to have stable homogeneity (Fendler and Muzaffar 2008). Thus, despite its varied applications in establishing the universality of the normal distribution, its overuse has slowly led to the need of being scrupulous of its correct formulation and usage across varied applications. Normality tests, including Q-Q plot and goodness-of-fit tests, are used for determining the likelihood that the given set of observations $X_1, X_2, \dots, X_n$ form a normal distribution. There are different types of normality. The normality tests are used to determine exact normality and approximate normality in the data. Another type of normality, asymptotic normality is a property of data, especially for a large sample, where the sample means have an approximately normal distribution, conforming to the CLT, i.e., the distribution of the z-scores of the sample means, $\bar{X}$, converges to a standard normal one as $n \rightarrow \infty$. Asymptotic normality is a key property of the least squares estimators of generalized space-time autoregressive

Normal Distribution,

Fig. 1 Plots for the probability density function in normal distribution with different values for mean μ and standard deviation σ , where the red curve is the *standard normal distribution*. (Image source: an online browser-based application, Desmos, <https://www.desmos.com/calculator/0x3rpqgrx>)



(STAR) models, used widely in ecology and geology (Borovkova et al. 2008). STAR models are applied in geological applications where prior information about spatial dependence, e.g., exact site locations for observations, is available. Asymptotic normality says that the estimator converges to the unknown parameter at a very fast rate of $\frac{1}{\sqrt{n}}$, which emphasizes the quality of the estimator. STAR models and their generalized variants are used for predicting rainfall, air pollution, groundwater levels in basin models, and similar variables in geospatial models.

Apart from its role in the CLT, normal distributions are indispensable in signal processing. Following are some of the significant properties of the normal distribution (Kim and Shevlyakov 2008) related to signal processing:

1. Two normal distributions upon convolution give another normal distribution.
2. The Fourier transformation of a normal distribution is another normal one.
3. The normal distribution maximizes entropy and minimizes Fisher information.

Solving for the variational problem of maximizing Shannon entropy $H(f)$, the distribution f that would discard

information and preserve variance bounded by its expected value $\bar{\sigma}^2$, is the normal distribution.

$$\text{For } H(f) = - \int_{-\infty}^{\infty} f(x) \log(f(x)) dx, \text{ we get } f^*(x) \\ = \arg \max_{f(x)} H(f) = \mathcal{N}(0, \bar{\sigma}).$$

Similarly, the lower bound of the variance of a parameter estimator gives minimum Fisher information. The solution of the distribution that minimizes Fisher information is a normal distribution.

$$\text{For } I(f) = \int_{-\infty}^{\infty} \left(\frac{\partial \log(f(x, \theta))}{\partial \theta} \right)^2 f(x, \theta) dx, \text{ we get } f^*(x) \\ = \arg \min_{f(x)} I(f) = \mathcal{N}(0, \bar{\sigma}).$$

Related Distributions

The normal distribution belongs to the exponential family of probability distributions, which also include log-normal, von Mises, and others.

Log-normal distribution is a close relative of normal distribution in the exponential family. It is a continuous probability distribution where the logarithm of the random variable follows a normal distribution. It is also referred to as Galton's distribution, and is widely used in modeling data in engineering sciences, including others. If the observed values are strictly positive, then the log-normal distribution is preferred for modeling over the normal distribution, which accommodates both positive and negative values (Barnes et al. 2021). Its probability distribution function is given as:

$$\phi(x; \boldsymbol{\mu}, \boldsymbol{\sigma}) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} \text{ for } \mu \in (-\infty, +\infty), \\ \sigma > 0, \text{ and support } x \in (0, +\infty).$$

von Mises distribution is a continuous probability distribution, that can be closely approximated as a *wrapped* or circular normal distribution. Hence, it is also referred to as the circular normal or Tikhonov distribution. Applications of von Mises distribution include any periodic phenomena, e.g., wind direction (Barnes et al. 2021). The parameters of the von Mises distribution are μ and κ , where the latter is the reciprocal parameter of dispersion and is analogous to variance σ^2 . Its probability density function, with $I_0(\kappa)$ as the modified Bessel function of order 0, is given as:

$$\phi(x; \boldsymbol{\mu}, \boldsymbol{\kappa}) = \frac{e^{\kappa \cos(x-\mu)}}{2\pi I_0(\kappa)}, \text{ where } \mu \in \mathbb{R}, \kappa > 0, \\ \text{support } x \in \text{any interval of length } 2\pi.$$

Geostatistical Applications

Normal distributions are widely used in geostatistical modeling. Two different examples, namely, Kriging and statistical parametric mapping (SPM), are explored here to understand the relevance of normal distributions in geoscientific applications.

Data gaussianity is an essential property, where the data has the propensity for normal distributions, in geo- and spatial statistical modeling and experimental data analysis. Kriging is a widely used geostatistical methodology for interpolation that relies on the prior covariances between samples that are assumed to be normally distributed. The Kriging methodology, which is a form of Bayesian inference, involves estimating the posterior distribution by combining the Gaussian prior with a Gaussian likelihood function for each of the observed variables. This posterior distribution also tends to be normal. The covariance function is used to stipulate the properties of the normal distribution. Thus, Kriging represents the variability of the data up to its second moment, i.e., covariance. Kriging has varied interdisciplinary uses in hydrogeology, mining, petroleum geosciences (for oil production), etc.

Statistical parametric mapping (SPM) also demonstrates varied uses of normal or Gaussian distributions in spatial data mining (McKenna 2018). Spatial fields are a common representation of continuous data in geoscientific and environmental applications, e.g., permeability, porosity, mineral content, reflection in satellite imagery, etc. Comparison of such fields over time or other characteristics (e.g., ensemble runs of simulations) require the generation of difference maps, in which anomalies are detected, for better data interpretation, using SPM. SPM involves analyzing each voxel in a volumetric grid or pixel in an image, using any univariate parametric test, and the results of the test statistics are assembled in an image, in the form of a map. Properties of Gaussian or normal distribution of the data in the fields drive the SPM techniques. The pixel values in the statistical parametric map are modeled using a two-dimensional normal/Gaussian distribution function. Usually, this involves transforming a map of t-statistics to that of z-scores. Once the SPM is generated for localized anomaly detection, several post-processing steps of smoothing using Gaussian filtering, reinflation of the variance, image segmentation, and hypothesis testing are implemented in sequence. SPM has been used for anomaly detection in the imagery in satellite images, in transmissivity in groundwater pumping, etc. This specific application places a large emphasis on the Gaussian distribution of the pixels in the difference maps. Given the multivariate nature of data involved in SPM, the modeling involves multi-Gaussian (MG) fields.

Future Scope

Some of the popular geostatistical methodologies, e.g., Kriging, exploit the properties of normal distribution. Hence, to improve the applicability of such a methodology, data transformation is an option. One such method is Gaussian transformation of spatial data (Varouchakis 2021), which is needed to reduce the effect of outliers, improve stationarity of the observed data, and process the data to be apt as the input for ordinary Kriging, which is widely used. Determining such transformations is challenging, as it is highly dependent on the data and its application. The use of the modified Box-Cox technique for Gaussian transformation has been successfully used to improve the normality metrics of datasets, such as detrended rainfall data, groundwater-level data, etc. Thus, this line of work opens up research on appropriate data transformation techniques to improve normality in the data and will continue to keep normal distributions highly relevant to geostatistics.

Given the uptake in the use of neural networks for data analysis, normal distributions and related distributions (e.g., log-normal, von Mises, etc.) play a crucial role in uncertainty

quantification of data via machine learning in geosciences (Barnes et al. 2021). Uncertainty can be added to any neural network regression architecture to locally predict the parameters of a user-specified probability distribution instead of just performing value prediction. Apart from using normal distributions as user-defined ones for fitting the data, uncertainty quantification also involves adding noise which follows either normal or log-normal distribution.

In summary, this article introduces the concept of normal distribution and brushes upon its relevance in geosciences.

Cross-References

- [Kriging](#)
- [Lognormal Distribution](#)
- [Spatial Statistics](#)
- [Spatiotemporal Modeling](#)
- [Stationarity](#)
- [Uncertainty Quantification](#)
- [Variance](#)
- [Z-transform](#)

Bibliography

- Barnes EA, Barnes RJ, Gordillo N (2021) Adding uncertainty to neural network regression tasks in the geosciences. arXiv preprint arXiv:210907250
- Borovkova S, Lopuhaä HP, Ruchjana BN (2008) Consistency and asymptotic normality of least squares estimators in generalized STAR models. *Statistica Neerlandica* 62(4):482–508
- Fendler L, Muzaffar I (2008) The history of the bell curve: sorting and the idea of normal. *Educ Theory* 58(1):63–82
- Gauss KF (1963) *Theoria Motus Corporum Celestium*, Perthes, Hamburg, 1809; English translation, *Theory of the motion of the heavenly bodies moving about the sun in conic sections*. Dover, New York
- Kim K, Shevlyakov G (2008) Why Gaussianity? *IEEE Signal Process Mag* 25(2):102–113
- Laplace PS (1781) Mémoire sur les probabilités. *Mémoires de l'Académie royale des Sciences de Paris* 9:384–485
- McKenna SA (2018) Statistical parametric mapping for geoscience applications. In: *Handbook of mathematical geosciences*. Springer, Cham, pp 277–297
- Pearson K, de Moivre A, Archibald R (1926) A rare pamphlet of Moivre and some of his discoveries. *Isis* 8(4):671–683
- Stigler SM (1986) *The history of statistics: the measurement of uncertainty before 1900*. Harvard University Press, Cambridge
- Varouchakis EA (2021) Gaussian transformation methods for spatial data. *Geosciences* 11(5):196

Object Boundary Analysis

Hannes Thiergärtner

Department of Geosciences, Free University of Berlin, Berlin, Germany

Definition

A linearly ordered set of multivariate geological objects can be segmented into quasi-homogeneous subsets significantly different from directly neighbored subsets. Rodionov (1968) developed a statistical algorithm augmenting the already known heuristic cluster-analytical models to classify ordered objects and to estimate boundaries between different subsets (1. Boundaries between ordered objects). The basic algorithm is supplemented by two algorithms that classify disordered objects into quasi-homogeneous subsets (2. Boundaries between disordered objects) and rank detected boundaries according to their importance (3. Hierarchy of boundaries).

Basic Algorithm

1. *Boundaries Between Ordered Objects.* A linearly ordered set of n vectors of m quantitative attributes $x_n = \{x_{n1}, x_{n2}, \dots, x_{nm}\}$ is to be classified iteratively into two subsets h_1 covering k vectors and h_2 covering $(n-k)$ vectors ($1 \leq k \leq n-1$). The assumed boundary between h_1 and h_2 is subjected to a statistical test. The iterative procedure is starting with $k=1$ and ending with $k=(n-1)$. Mean value vectors are calculated for both subsets. Significance of a difference d between the mean value vectors is subjected to the alternative hypothesis $H_1 (d(r^2_0) \neq \{0; 0; \dots; 0\})$ and tested against the null hypothesis $H_0 (d(r^2_0) = \{0; 0; \dots; 0\})$. The test criterion is

$$v(r^2_0) = \frac{n-1}{n(n-k)k} \sum_{i=1}^m \frac{\left((n-k) \sum_{t=1}^k x_{ti} - k \sum_{t=k+1}^n x_{ti} \right)^2}{\sum_{t=1}^n x_{ti}^2 - \frac{1}{n} \left(\sum_{t=1}^n x_{ti} \right)^2} \quad (1)$$

where x_{ti} is the measured value of attribute i for object t . A statistically significant boundary between both subsets is given at a significance level α if $v(r^2_0) > \chi^2(\alpha; m)$. The value $v(r^2_0)$ is marked. The null hypothesis is accepted if $v(r^2_0) \leq \chi^2(\alpha; m)$ meaning that the assumed boundary is not statistically significant. All significant boundaries are marked by their $v(r^2_0)$ values. Finally, a series is created of 1, 2, or at most $(n-1)$ boundaries each characterized by its $v(r^2_0)$ value. In a second step, the boundary with the maximum test criterion is selected as the most important one. It subdivides the set of ordered objects into two main subsets. Steps 1 and 2 will be repeated successively for all created subsets until new subsets cannot be established.

Step 3 includes elimination of created but unnecessary boundaries which could subdivide homogeneous neighbor vectors. It is based on a homogeneity test for all obtained boundaries between neighbored subsets T_1 and T_2 which are not resulting from the last step. The null hypothesis that a generated boundary is unnecessary will be accepted if $v(T_1; T_2) \leq \chi^2(\alpha; m)$ where

$$v(T_1; T_2) = \frac{n_1 + n_2 - 1}{n_1 n_2 (n_1 + n_2)} \times \sum_{i=1}^m \frac{\left(n_1 \sum_{t \in T_2} x_{ti} - n_2 \sum_{t \in T_1} x_{ti} \right)^2}{\sum_{t \in T_1 \cup T_2} x_{ti}^2 - \frac{1}{n_1 + n_2} \left(\sum_{t \in T_1 \cup T_2} x_{ti} \right)^2} \quad (2)$$

Acceptance of the null hypothesis means that the neighboring subsets are not significantly different from each other and that their boundary was artificially resulting from the preceding iterative procedure.

2. *Boundaries Between Disordered Objects.* This is a similar approach, but it does not get prevalence because the power of the first algorithm regarding the neighborhood is not inserted into the model.

A set of n vectors for m quantitative attributes $x_n = \{x_{n1}, x_{n2}, \dots, x_{ni}, \dots, x_{nm}\}$ which do not show any spatial order is to be classified into h subsets $T_1, T_2, \dots, T_s, \dots, T_h$ with at least two objects each. The calculation starts with the determination of all $v(T_s, T_k)$ between the $\frac{h(h-1)}{2}$ pairs of objects:

$$v(T_s, T_k) = \frac{n_s + n_k - 1}{n_s n_k (n_s + n_k)} \times \sum_{i=1}^m \frac{\left(n_s \sum_{t \in T_k} x_{ti} - n_k \sum_{t \in T_s} x_{ti} \right)^2}{\sum_{t \in T_s \cup T_k} x_{ti}^2 - \frac{1}{n_s + n_k} \left(\sum_{t \in T_s \cup T_k} x_{ti} \right)^2} \quad (3)$$

where T_s and T_k are the first and second subsets with n_s and n_k objects ($n \geq 2$) respectively, and x_{ti} is the value for attribute i for object t . Next, a test for general homogeneity of all objects is conducted. General homogeneity means that there are no significant boundaries between the multivariate described objects involved. The null hypothesis

$$H_0(\text{all } [\Theta_{T_s} - \Theta_{T_k}] = \{0; 0; \dots; 0\}) \quad (4)$$

regarding all objects will be accepted if

$$\min_{s,k} v(T_s, T_k) \leq \chi^2(\alpha; m). \quad (5)$$

At least one object shows a significantly different attribute vector, and at least one boundary exists if the null hypothesis is rejected. Θ_{T_s} refers to the mean value vector of the subset T_s which includes n_s geological objects. The procedure will be continued only if the null hypothesis is rejected. In this case, subsets will be united iteratively into larger quasi-homogenous subsets. At every step, two subsets corresponding to the term $\min_{(s,k)} v(T_s, T_k)$ form a new, larger subset without internal boundary if $v \leq \chi^2(\alpha; m)$. This new subset replaces the previous subsets T_s and T_k . The number of classes is reduced by 1. The tests are continued as before until all possible combinations of subsets have been processed.

3. *Hierarchy of Boundaries.* The problem is testing whether or not significant boundaries b_i characterized by their v -values $v_i (r^2_0)$ are of equal rank. There can arise four possible situations: boundaries are equal, one boundary is greater than or less than another one, or the boundaries are not equal.

A normally distributed test criterion

$$N = \frac{v_1 - v_2}{2\sqrt{v_1 + v_2 - m}} \quad (6)$$

will be applied only if two boundaries b_1 and b_2 are in existence. The null hypothesis $H_0 (b_1 = b_2)$ is preferred if no other hypothesis is accepted; $H_1 (b_1 < b_2)$ is accepted if $N < N(\alpha)$; $H_2 (b_1 > b_2)$ is accepted if $N > N(1-\alpha)$; and $H_3 (b_1 \neq b_2)$ is accepted if $N > N(1-\frac{\alpha}{2})$.

A χ^2 -distributed test criterion

$$\chi^2 = \frac{1}{2(2y - m)} \sum_{i=1}^k (v_i - y)^2 \quad (7)$$

will be applied if g boundaries $b_1, b_2, \dots, b_g$ exist; y is the mean value of all v -values. The null hypothesis $H_0 (b_1 = b_2 = \dots = b_g = b_0)$ that all boundaries are of equal rank is to be accepted if $\chi^2 \leq \chi^2(\alpha; k-2)$. Otherwise, at least one of the boundaries can be distinguished clearly from the other ones. If this is the case, all v -values will be ordered according to their decreasing size: $y_{i1} > y_{i2} \dots > y_{ig}$. Next, boundaries between geological objects will be united within a common class of equivalent boundaries if they do not show significant differences. This procedure will be iteratively repeated until no relation $v \leq \chi^2$ can be found anymore. The end result is a hierarchical set of boundaries classified according to their range.

Applications

Examples of linearly ordered sets of multivariate geological objects are borehole profiles, profile sections, terrestrial or submarine traverses, and time series as well.

Neighboring geological objects not separated by significant boundaries constitute quasi-homogeneous classes. Each class can be described by the estimated mean value and variance of every attribute, by the number of objects included, by its spatial or temporal position and other related information.

Spatial or temporal coordinates of the geological objects are not taken into account. Equidistance of samples or observations is not necessary, but their given linear sequence must be respected. In view of the statistical test criteria belonging to the procedure, a (0;1)-standardization of all input data is recommended so that all attributes are to receive equal weight. More subtle differences between neighbored objects can be detected if more attributes are considered. An increased statistical significance level such as $\alpha = 0.10$ instead of the commonly used $\alpha = 0.05$ diminishes the theoretical χ^2 values and results in a finer subdivision of the set of linearly ordered geological objects.

The boundaries between obtained classes can be interpreted geologically. An application of the testing model “hierarchy of boundaries” is recommended for differentiation between more or less important boundaries. Boundaries of higher order may be tectonic (e.g., thrust faults) or stratigraphic discordances. Lower-order boundaries represent general changes in tectonically undisturbed sequences of sedimentary layers or in the evolving fossil content. A boundary in the chemical composition or in mineral content provides essential information if quasi-homogeneous blocks in mineral deposits should be distinguished and delineated for subsequent mining.

The model “boundaries between disordered objects” represents an extension of the basic model “boundaries between linearly ordered objects.” The principle that a spatial neighborhood of multivariate objects is necessary does not exist anymore. Quasi-homogeneous classes can be generated even between nonneighboring objects. This statistical model competes with the better-known cluster analysis methods. However, it has the advantage of forming spatially interrelated groups of objects. An application can be recommended if a number of multivariate well logs are available for an exploratory target area or for a mineral deposit. The model then results in several closed classes both within every profile and between profiles. This gives the opportunity for cross-correlation between geologically connected profiles. The subset boundaries can be interpreted as stratigraphic, fossiliferous, or lithologic ones, or as delineations between quasi-homogeneous spatial blocks within mineral deposits. Past applications in various geological projects in Germany gave good results.

Relation to Cluster Analysis and Discriminant Analysis

The segmentation of a linearly ordered set of multivariate geological objects belongs to the group of nonsupervised classification methods. It is based on a multivariate statistical comparison of means at a chosen significance level. The user cannot incorporate any other (geo)scientific information. The model developed and introduced by Rodionov respects the neighborly relations between geological objects and enables uniting neighboring objects into a common quasi-homogeneous subset even if their attribute vectors differ to some degree. Thus, it is a generalizing approach.

A subdivision of the same set of objects by cluster analysis is a heuristic model based on one of several multivariate similarity or distance measures. Cluster analysis belongs also to the class of nonsupervised classification models. The user selects a suitable measure and can influence the result using experience and preknowledge. Information about the statistical

uncertainty is not available. Every geological object will be classified into one of several resulting groups regardless of its spatial or temporal position. In this way, the ordered sequence of objects will be subdivided into numerous independent subsets or single objects. The approach is individualizing.

Models of discriminant analysis differ clearly from the preceding models. They are supervised multivariate-statistical approaches. The boundaries between geological objects are determined a priori by geoscientific facts such as lithologic, stratigraphic, or facies allocation. This approach is applied to test the multivariate differences between predetermined classes of objects according to their statistical significance. Not yet classified objects can be assigned to existing classes at a given statistical level of significance, or they result in generating new classes of geological objects.

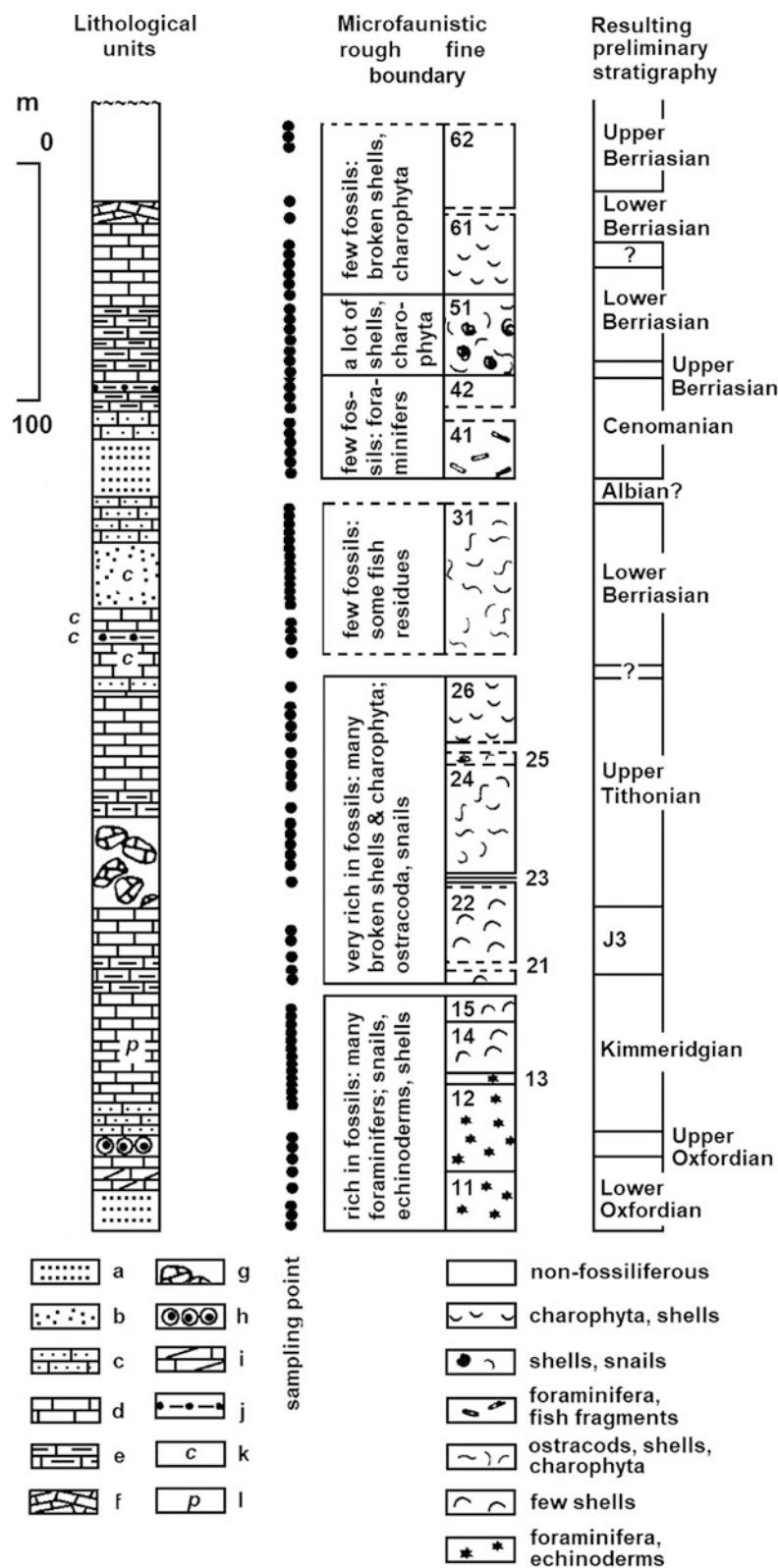
Example

The profile of borehole Herzfelde 4/63 east of Berlin (Germany) is to be classified by its microfauna population. It shows a monotone sedimentary sequence, poor in fossils and without any index fossils (Fig. 1). The series consists of fine to medium-grained sandstone (a), ordinary sandstone (b), calcareous sandstone (c), pelitic limestone (d), clayey limestone (e), limestone breccia (f), limestone conglomerate (g), oolitic limestone (h), dolomite (i), and clayish sandstone (j). Some layers are colored (k) and the deeper limestone is pelitic (l). The sequence covers Mesozoic sediments from the Oxfordian to the Cenomanian. The series was syn-sedimentarily to postsedimentarily disturbed by halokinetic processes (Schudeck and Tessin 2015). The sparse fossil content (residues of ostracods, charophyta, echinoderms, snails, and other types of shells and fish) was determined in 94 samples (column 2 in Fig. 1).

The segmentation of this profile by microfauna components resulted in six classes (column 3 in Fig. 1) for significance level $\alpha = 0.10$. The lower class corresponds to Oxfordian and Kimmeridgian strata and is concordantly overlain by Tithonian and earliest Berriasian followed by a weathered horizon (Early Berriasian). This sequence is overlain by poorly fossiliferous Late Berriasian and Cenomanian sediments (class no. 4, possibly limited by a tectonic fault?). Above an obvious discordance, class no. 5 continues the Early Berriasian sequence which is very rich in fossils. The hanging group follows discordantly and is composed of Early Berriasian and Late Berriasian containing only a few fossils.

A reduction of the significance level to $\alpha = 0.05$ results in altogether 16 microfauna boundaries (column 4 in Fig. 1), but the entire profile is still clearly structured into coherent sections. This is important for the geological interpretation.

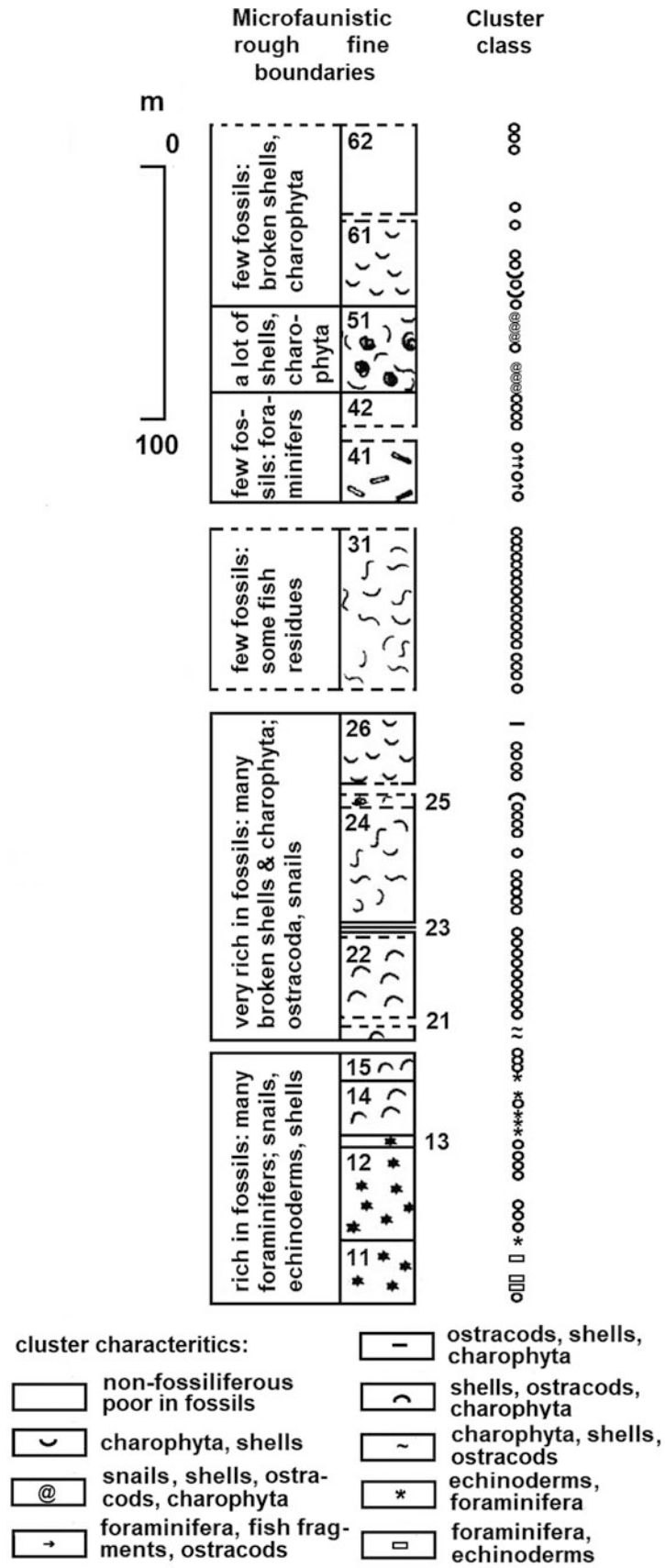
Object Boundary Analysis,
Fig. 1 Boundaries between
 ordered samples in a borehole
 profile



Column 5 in Fig. 1 shows the resulting preliminary stratigraphic classification of the studied Mesozoic profile.

The difference between the generalizing method “boundaries of ordered objects” and the individualizing cluster

analysis is presented in Fig. 2. The same data set was used for both approaches. The cluster-analytic result (column 3 in Fig. 2) shows nine classes similarly composed as the groups of the Rodionov algorithm. It shows partly also closed sequences



Object Boundary Analysis, Fig. 2 Comparison between generalized and individualized boundaries

of a quasi-homogeneous composition but also embedded thin layers characterized by a different fossil content. This comparison between results is suitable to recognize the difference between generalizing and individualizing methods.

Summary

Linearly ordered or not ordered sets of multivariate geological objects can be subdivided into quasi-homogeneous subsets by multivariate statistical methods developed by Rodionov (1968). The resulting boundaries can be tested for their statistical significance and can be interpreted geologically. This *generalizing* approach is distinguished from the *individualizing* cluster analysis by preferential involving neighboring objects into a common class without explicit inclusion of spatial or temporal coordinates. Application is recommended to classify borehole profiles, time series, and related ordered geoscientific objects.

Cross-References

- [Cluster Analysis and Classification](#)
- [Discriminant Analysis](#)
- [Pattern Classification](#)
- [Rodionov, Dmitriy Alekseevich](#)

Bibliography

- Rodionov DA (1968) Statistical methods to mark geological objects by a complex of attributes. Nedra, Moscow. (in Russian)
- Schudeck M, Tessin R (2015) Jurassic. In: Stackebrandt W, Franke D (eds) Geology of Brandenburg. Schweizerbart, Stuttgart, pp 217–240. (in German)

Object-Based Image Analysis

D. Arun Kumar¹, M. Venkatanarayana¹ and V. S. S. Murthy²

¹Department of Electronics and Communication Engineering and Center for Research and Innovation, KSRM College of Engineering, Kadapa, Andhra Pradesh, India

²KSRM College of Engineering, Kadapa, Andhra Pradesh, India

Definition

Object-based image analysis (OBIA) is defined as recognizing the target objects in an image through the processes like image segmentation, image classification, evaluation, and analysis of objects. Image segmentation is the process of dividing an image into homogeneous segments by grouping

the neighboring pixels with similar characteristics like tone, texture, color, and intensity. Image classification is the process of labeling the image objects to the respective classes as they appear in the reality. Evaluation is the process of comparing the predicted object classes to the actual/true classes. The term image objects is a group of pixels with similar characteristics. A pixel is a digital number representing the spectral reflectance from a given area.

Introduction

The earth resources change dynamically over the years due to various reasons like anthropogenic activities, natural processes and other activities. The consumption of resources such as water, forest, minerals, oil, and natural gases has increased from the past years (Lillesand and Kiefer 1994). Also, the land use/land cover changes are frequent due to the reasons such as overflows, drought, cyclones, tsunamis, etc. The percentage of resource utilization by the humans has increased over the recent years (Lillesand and Kiefer 1994). Many studies were implemented scientifically to quantify the changes in the resources like forest, water, and land use/land cover changes (Richards and Jia 2006). Remote sensing is one such type of art, science, and technology of acquiring the information about the resources on the earth surface without any physical contact with the object of interest (Campbell 1987). The images/photographs of the target of interest are obtained remotely from airborne/spaceborne platforms (Richards and Jia 2006). The very first attempt of acquiring the aerial photographs were implemented by airborne platforms like parachutes, air crafts, etc. In the recent years, space-based platforms like satellites are used to acquire the information about the target of interest (Schowengerdt 2006). The information about the target of interest is remotely sensed and represented in the form of images (Jensen 2004).

In the earlier days, the images were obtained in analog format with coarse resolution (Gonzalez and Woods 2008). Identification of object of interest in the images with coarse resolution is difficult task for the manual analysis. Advancement in the sensor technology like development of charged coupled devices generated digital images with fine spatial resolution. The digital images are more informative in comparison with analog images. A digital image consists of finite set of elements called pixels/pels (Bovik 2009). Each pixel in a digital image is indicated with digital number (DN) (Gonzalez and Woods 2008). A DN of a pixel represents the average spectral reflectance sensed by the sensor in the finite area of interest. The digital image consists of digital numbers represented in rows and columns (Bovik 2009). The digital images are processed to extract the information about the area of interest using various processing techniques like image enhancement, image transformation, and image classification (Richards and Jia 2006). The required information from the digital images is acquired with the usage of automated

methods. These automated methods are mathematically implemented using statistics, probability, set theory, graph theory, etc. (Schowengerdt 2006).

The processing and analysis of images is classified as

(i) manual approach and (ii) digital approach. In the manual approach expert committee will recognize the object of interest using the image interpretation keys/elements (Lillesand and Kiefer 1994). In the digital image processing and analysis approach the objects are detected and recognized using automated methods followed by the expert committee analysis. The DNs in an image are processed using two approaches, namely (i) Pixel-based image analysis and (ii) Object-based image analysis (Jensen 2004). The contents of present entry are given as pixel-based image analysis, object-based image analysis, major applications of object-based image analysis in remote sensing, and conclusions.

Pixel-Based Image Analysis

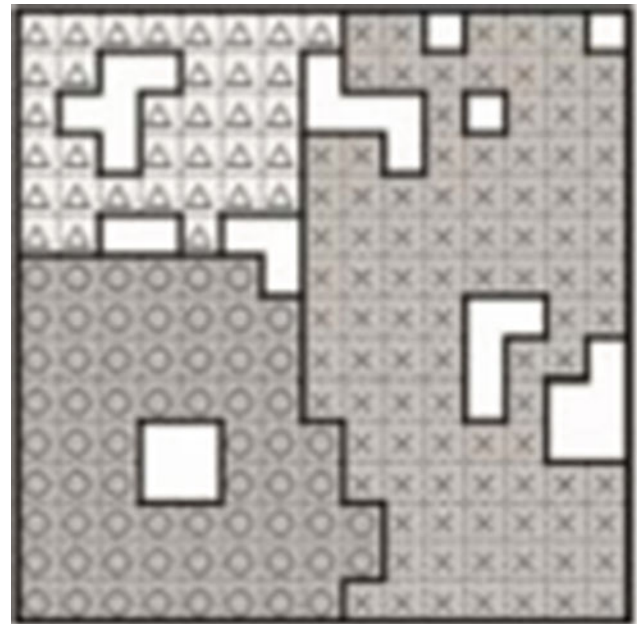
The pixel-based image analysis approach (PBIA) is based on spectral signature of the area of interest. In the pixel-based image analysis the processing techniques are implemented at the finite level of the digital images called as pixels (Veljanovski et al. 2011). Each pixel in the image is processed to acquire the information of the area. Pixel-based classification is one of the key operations performed in PBIA. In pixel-based classification each pixel in the input image is labelled to the corresponding class using various classifiers (Johnson and Ma 2020). The input pixel is considered as a vector in the N-dimensional feature space and the pixels of same class will form a cluster in the N-dimensional feature space. In pixel-based classification approach each pixel is assigned with the class label. There exists various pixel-based classifiers for remote sensing image analysis (Blaschke et al. 2014). K-nearest neighbors classifier assigns the class label to the pixel based on the K-nearest neighbors in the Euclidean space. Minimum distance to mean classifier assigns the class label to the pixel based on the minimum difference between the pixel value and the class mean (Veljanovski et al. 2011). Maximum likelihood classifier assigns the class label to the pixel based on the class conditional probability. The output of pixel-based classification approach is the classified image/classified map. Furthermore, the classified map is analyzed by the expert for quantitative assessment of various classes present in the image (Johnson and Ma 2020).

Pixel-based image analysis approach works well with coarse resolution image where the target object area is lesser than the spatial resolution of the sensor (Blaschke et al. 2014). This approach has limitations when the input image is with finer resolution where the target object area is larger than the spatial resolution of the sensor. Motivated with this reason, object-based image analysis approach was suggested for the

digital images. The object-based image analysis (OBIA) approach works well in recognizing the object of interest in both finer and coarser resolution digital images (Johnson and Ma 2020). The example of PBIA is provided in Fig. 1 and Fig. 2. In Fig. 1, there are three categories of vegetation classes. The input image is classified using PBIA approach and the classified output image is given in Fig. 2.



Object-Based Image Analysis, Fig. 1 Input image with vegetation classes. Reprinted from (Veljanovski et al. 2011)



Object-Based Image Analysis, Fig. 2 Output image of PBIA with vegetation classes represented in three symbols (Image is reprinted from (Veljanovski et al. 2011)). The input image is given in Fig. 1

Object-Based Image Analysis

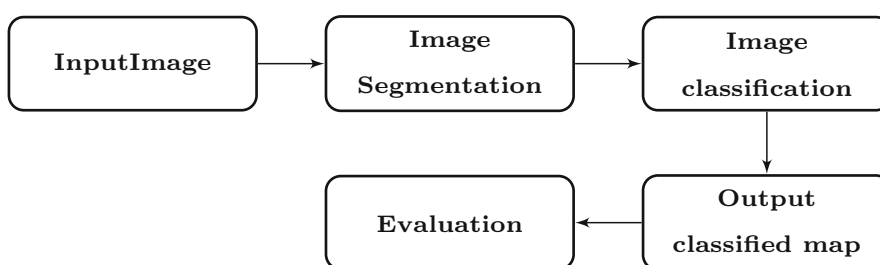
In OBIA, the image pixels are grouped based on the similarity like color, tone, texture, and intensity characters. The group of pixels are combined to generate the image objects (Jensen 2004). The image objects represent the themes like water body, trees, buildings, grounds, etc. (Richards and Jia 2006). In OBIA, the spatial and spectral characteristics of pixels are used to recognize the target objects in the image. The steps implemented in OBIA is given in Fig 3. In the first stage of OBIA, the input image is segmented to obtain the image objects. Image segmentation is the process of dividing the input image into homogeneous regions called image segments. The major image segmentation approaches are given as (i) region based approach and (ii) boundary based approach. In the region based approach, similar type of pixels are detected using the algorithms. In boundary based approach, the algorithm searches for the discontinuity in the image. The boundary based approach is implemented using the segmentation algorithms like edge detection, point detection and line detection (Bovik 2009). The image segmentation techniques are broadly classified as threshold based method, edge-based segmentation, region based segmentation, cluster based segmentation, watershed based method, and artificial neural network based segmentation (Blaschke et al. 2014). The segmented image consists of separated image objects based on shape and size characteristics.

In the second stage of OBIA, the objects in the segmented image are labeled to the respective classes by using various classification methods. The classification algorithms are categorized as (i) parametric classification and (ii) nonparametric classification (Johnson and Ma 2020). In the parametric classification, the statistical parameters like mean, median, mode,

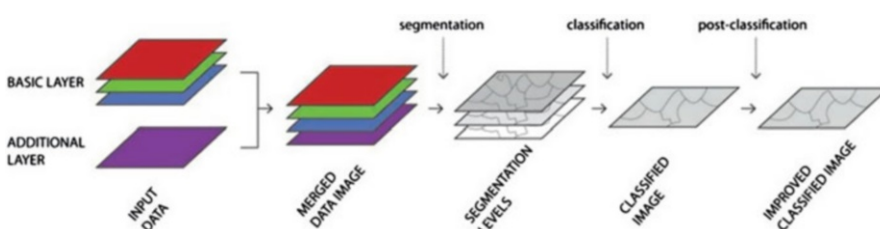
etc. of the input pixel values are considered during the image classification. In the nonparametric classification, the classification of digital images is performed without considering the statistical parameters of the input pixel values (Veljanovski et al. 2011). The major parametric/nonparametric classifier methods include K-means, minimum distance method, maximum probability method, ISODATA, K-nearest neighbor, support vector machines, and parallel-piped method (Johnson and Ma 2020). In addition, there exists neural networks (NNs)-based classifiers such as Kohonen networks, feedforward neural networks (FNNs), recurrent NNs (RNNs), deep neural networks (DNNs), etc. The machine learning-based classifiers such as decision trees and regression trees are used to classify the image objects into respective classes. The output of image classification is a classified map with various classes of land use/land cover classes, natural resources, etc. (Blaschke et al. 2014). The processing steps of OBIA approach is given in Fig. 4. The output of OBIA approach for the input image (Fig. 1) is given in Fig. 5.

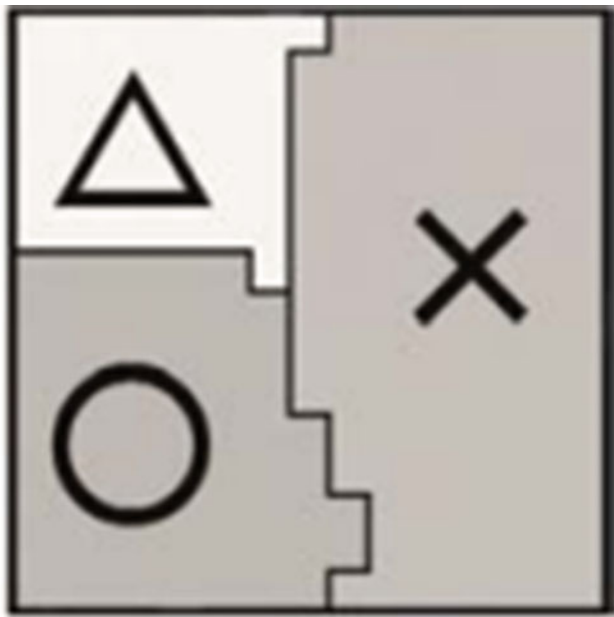
The output classified image of OBIA approach is evaluated using various indices like user's accuracy (UA), producer's accuracy (PA), overall accuracy (OA), dispersion score (DS), and kappa coefficient (KC) (Lillesand and Kiefer 1994). The performance indices are derived from confusion matrix (CM). CM is a table of actual classified output classes assigned by a model. The sum of diagonal elements of the CM divide by total number of pixels is termed as OA (Campbell 1987). The results of OBIA are correlated with the ancillary data and evaluated with the field visits by selecting the locations based on random approach (Jensen 2004). Furthermore, based on the evaluation results, the OBIA is extended to various remote sensing applications.

Object-Based Image Analysis, Fig. 3 Schematic flow diagram of OBIA. The flow diagram consists of steps like passing the input image, segmenting the input image to obtain image objects, classifying the image objects, and evaluation of accuracy



Object-Based Image Analysis, Fig. 4 Processing steps of OBIA. Reprinted from (Veljanovski et al. 2011)





Object-Based Image Analysis, Fig. 5 Output image of OBIA with detailed vegetation classes (Reprinted from (Veljanovski et al. 2011)). The input image is given in Fig. 1

Major Applications of Object-Based Image Analysis in Remote Sensing

The applications of object-based image analysis in remote sensing is broadly classified as land use/land cover classification, forest mapping, mineral mapping, coastal zone studies, crop classification, and urban studies (Blaschke et al. 2014). In land use/land cover classification, the pixels in the image are classified into corresponding land use/land cover class like builtup land, urban dense, urban sparse, scrub land, and deciduous forest (Johnson and Ma 2020). OBIA is used in the forest studies for change detection and quantitative assessment of vegetation classes. OBIA is used to detect the minor and major lineaments in the remote sensing images for mineral exploration (Veljanovski et al. 2011). In coastal zone studies the OBIA is used to study the dynamics of shore line changes. The coastal zone applications also include the change detection in the vegetation like mangroves. The crop classification in remote sensing images consists of mapping and assessment of various type of crops in high-resolution images (Blaschke et al. 2014). In the urban studies, the OBIA is used to study the dynamics of urban sprawl and change detection in the city pattern (Johnson and Ma 2020).

The applications of OBIA exists from coarse resolution to fine resolution remote sensing images (Schowengerdt 2006). The coarse resolution remote sensing images are acquired using the sensors like Linear Imaging Self-Scanner (LISS III), Linear Imaging Self Scanner (LISS II), Multispectral Scanner (MSS), Wide field sensor (WFS), Advanced Wide Field Sensor (AWIFS), etc. (Campbell 1987). The fine resolution remote sensing images are acquired from the

sensors like Optical Sensor Assembly, GeoEye Imaging System, Shortwave Infrared sensor, etc. (Jensen 2004). The coarse resolution images has 15 m to 75 m coverage of ground area and the fine resolution images have 0.5 m to 4.5 m coverage of ground area. The OBIA is efficiently applied on monochromatic, panchromatic, multispectral, and hyper-spectral remote sensing images. OBIA has produced impressive results with the images acquired from 1–350 spectral bands (Gonzalez and Woods 2008).

Conclusions

Advancement of sensor technology produced the images with good spatial resolution. OBIA approach is used for better information extraction from the fine resolution images in comparison with traditional methods. The traditional PBIA approach produces better results with coarse resolution images in comparison with fine resolution remote sensing images. PBIA considers the spectral characteristics of the object of interest during the processing of digital images. The performance of OBIA is efficient as tested with both coarse and fine resolution remote sensing images. OBIA considers the spatial and spectral characteristics of pixels during the object recognition. These advantages of OBIA provide better results in comparison with PBIA. OBIA is used in various type of remote sensing applications like land use/land cover classification, resource monitoring, urban mapping etc., where the spatial resolution of the images varies from 75 m to 0.5 m and the spectral resolution of images varies from single band to hundred bands. OBIA is used in the applications related to earth's atmosphere, earth's surface, and earth's sub-surface. In the recent years, intensive research on OBIA with machine learning is being carried by the scientific community. OBIA with machine learning concepts produced better results in various remote sensing applications. The future scope of OBIA involves the inclusion of temporal and radiometric characteristics of objects along with spatial and spectral characteristics during the processing of digital images.

Acknowledgment The first author is thankful to K. Madan Mohan Reddy, vice-chairman, Kandula Srinivasa Reddy College of Engineering, and A. Mohan, director, Kandula Group of Institutions for establishing Machine Learning group at Kandula Srinivasa Reddy College of Engineering, Kadapa, Andhra Pradesh, India – 516003.

Bibliography

- Blaschke T, Hay GJ, Kelly M, Lang S, Hofmann P, Addink E, Feitosa RQ, der Meer FV, der Werff HV, Coillie FV (2014) Geographic object-based image analysis towards a new paradigm. *ISPRS J Photogramm Remote Sens* 87:180–191
- Bovik AC (2009) The essential guide to image processing. Academic Press

- Campbell JB (1987) Introduction to remote sensing. The Guilford Press
- Gonzalez RC, Woods RE (2008) Digital image processing, 3rd edn. Prentice Hall, New Jersey
- Jensen JR (2004) Introductory digital image processing: a remote sensing perspective, 3rd edn. Prentice Hall
- Johnson BA, Ma L (2020) Image segmentation and object-based image analysis for environmental monitoring: recent areas of interest, researchers views on the future priorities. *Remote Sens* 12(11). <https://doi.org/10.3390/rs12111772>
- Lillesand TM, Kiefer RW (1994) Remote sensing and image interpretation. Wiley
- Richards JA, Jia X (2006) Remote sensing digital image analysis: an introduction, 4th edn. Springer
- Schowengerdt RA (2006) Remote sensing: models and methods for image processing. Elsevier
- Veljanovski T, Kanjir U, Ostir K (2011) Object-based image analysis of remote sensing data. *GEOD VESTN* 55(4):678–688

Olea, Ricardo A.

C. Özgen Karacan
U.S. Geological Survey, Geology, Energy and Minerals
Science Center, Reston, VA, USA



Fig. 1 Ricardo A. Olea, courtesy of Ricardo A. Olea, Jr.

Biography

Dr. Ricardo A. Olea received his B.Sc. degree in Mining Engineering from the University of Chile, Santiago, in 1966, and was awarded the Juan Brüggen Prize by the Instituto de Ingenieros de Minas de Chile for the best graduating Mining Engineer. Following graduation, he started his

professional career as an exploration seismologist with the National Oil Company of Chile (ENAP). Later, he pursued advanced qualifications in the United States. He received an M.Sc. degree in Computer Science from the University of Kansas in 1972 with a thesis on “Application of regionalized variable theory to automatic contouring,” and a Ph.D. from the Chemical and Petroleum Engineering Department of the University of Kansas in 1982 with a dissertation titled “Systematic approach to sampling of spatial functions.” He immigrated to the United States in 1985 to pursue a career as a research scientist with the Kansas Geological Survey in Lawrence, Kansas, where he worked until 2003. Before joining the USGS (United States Geological Survey) in 2006, where he currently works as a Research Mathematical Statistician, he had appointments with the Department of Environmental Sciences and Engineering of the School of Public Health at the University of North Carolina in Chapel Hill; the Marine Geology Section of the Baltic Research Institute of the University of Rostock, Warnemünde, Germany; and the Department of Petroleum Engineering at Stanford University.

Besides authoring and coauthoring more than 250 publications, including 4 books, in quantitative modeling in the earth sciences, energy resources, geostatistics, compositional data modeling, well log analysis, geophysics, and geohydrology, Ricardo made exceptional contributions to the profession of mathematical geosciences and the IAMG (International Association for Mathematical Geosciences). He served as the chair of IAMG Geostatistics Committee (1985–1989), and the IAMG Membership Committee (1989–1992), as IAMG Secretary-General (1992–1996), IAMG President (1996–2000), and IAMG Past-President (2000–2004). He still takes active roles in the IAMG and maintains affiliations with professional organizations. He currently serves as an Associate Editor of the Springer journal *Stochastic Environmental Research and Risk Assessment*. Ricardo was presented with the IAMG’s highest Award, the William Christian Krumbein Medal in 2004, due to his distinguished research, service to the IAMG, and service to the profession in general.

Ricardo’s achievements and professional accomplishments are many and much more than the brief summary highlighted above. As importantly, Ricardo is a very kind and modest person, despite his distinguished career, he is always keen to highlight excellence in others rather than his own. He is an influential and a well-loved mentor for many around the world. He is a great friend and an exceptional colleague to work with. Ricardo is always dependable and democratic, and he seeks a broadly based consensus before making any decisions. He always strives to learn new things and share them with his colleagues with openness. I feel privileged to have been asked to write this contribution about Ricardo in this volume.

Ontology

Torsten Hahmann

Spatial Informatics, School of Computing and Information Science, University of Maine, Orono, ME, USA

Definition

In the mathematical geosciences the term ontology denotes an information artifact that specifies a set of terms, consisting of classes of objects, their relationships and properties, and semantic relationships between them. But even in the information and computing sciences, the word “ontology” can denote vastly different kinds of artifacts, see Fig. 1 and Table 1 for examples. They differ in purpose and scope, ranging from informal conceptual ontologies to formal ontologies and from very broad to narrowly tailored to specific domains or specialized reasoning applications. They also differ in their representation formats, which vary in formality and expressivity, ranging from concept maps and term lists to structured thesauri and further to formal languages. If the semantics of terms are expressed in a formal, that is machine-interpretable, language one speaks of a “formal ontology.” Ontologies specified in less rigorous languages are referred to as conceptual ontologies or models.

Introduction

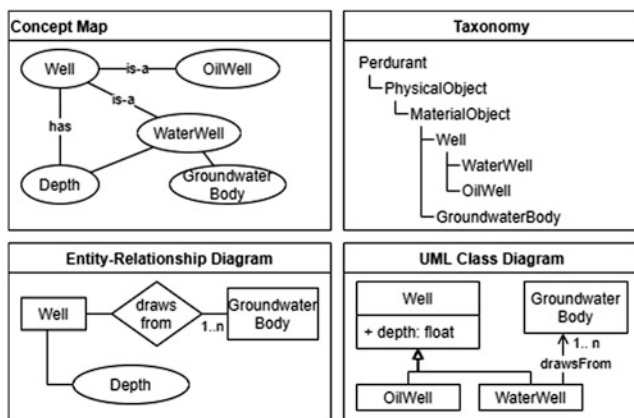
The term ontology originates from the study of the nature of being and the categorization of the things that exist in the world. This philosophical tradition – referred to by the uppercase term *Ontology* – dates back to Aristotle’s work

on metaphysics. Within computing and information sciences, the term *ontology* is used more narrowly to refer to an artifact that result from explicitly modeling a portion of the world or, in the words of Gruber and Studer, an “explicit formalization of a [shared] conceptualization” (Staab and Studer 2009). See Jakus et al. (2013) for additional historical context.

Over the last decades, ontologies of various scopes, formalities, and expressivities have emerged for a wide range of uses. The most obvious distinction between them is how they are represented, which ranges from simple term lists and graphical models to highly sophisticated logic-based languages, with the Resource Description Framework (RDF) (<https://www.w3.org/TR/rdf11-concepts/>) and RDF Schema (RDFS) (<https://www.w3.org/TR/rdf-schema/>) and the Web Ontology Language (OWL2) (<https://www.w3.org/TR/owl2-overview/>) most widely used (Staab and Studer 2009).

Regardless of their differences, all ontologies share two ingredients: (1) a set of terms – its *terminology* or *vocabulary* – and (2) a specification of how the terms are defined or related to one another as a means of specifying the terms’ semantics, that is, how to interpret them. The terms encompass *classes*, sometimes called *concepts*, and *properties*, also referred to as *relations*, *predicates*, or *roles*. Mathematically, classes can be thought of as sets (unary relations), whereas properties have arities of two or greater (Many representation formats, including RDFS and OWL2, only permit binary properties.). As an example, an ontology may include classes such as “Town,” “Well,” and “GroundwaterBody” and binary relations between classes (*object properties* in RDFS and OWL2) such as “contains,” “drawsFrom,” or “connectedTo” as well as attributes (*data properties* in RDFS, OWL2) such as “depth,” “volume,” “latitude,” and “longitude.” By relating terms to one another via generic or specific semantic relationships or via axioms, the terms’ semantics are specified. The most prevalent semantic relations are taxonomic ones (*hyponymy* and *hypernymy*) such as “is a subclass of” (short: “is-a”) and “is a subproperty of.” Other common semantic relations are *meronymy* (part of; mereology in logic) and its opposite *holonymy* as well *synonymy*, *antonymy*, and *similarity*.

The process of developing ontologies is called *ontological engineering*, for which established methodologies, such as TOVE, Methontology, On-To, DILIGENT, or NeOn, are surveyed in the entry “Ontology Engineering Methodology”, Staab and Studer (2009). Development typically is a multi-stage process from knowledge elicitation all the way to ontology verification and validation and involves both knowledge engineers and domain experts. Various tools, including competency questions, ontology design patterns (entry ► “Pattern Recognition”), and formal ontological analysis help at various stages. More recently, the emphasis of ontological



Ontology, Fig. 1 Example snippets of nonformal ontologies loosely modeled after (Hahmann and Stephen 2018; Brodaric et al. 2019)

Ontology, Table 1

Examples snippets of formal ontologies using some of the concepts from Fig. 1

RDFS:
:WaterWell rdfs:subClassOf: Well.
:OilWell rdfs:subClassOf: Well.
:drawsFrom rdfs:domain: WaterWell;
rdfs:range: GroundWaterBody.
OWL2:
:WaterWell rdf:type owl:Class;
rdfs:subClassOf [rdf:type owl:Restriction;
owl:onProperty: drawsFrom;
owl:someValuesFrom: GroundWaterBody];
owl:disjointWith: OilWell;
:drawsWaterFrom rdf:type owl:ObjectProperty;
rdfs:domain: WaterWell;
rdfs:range: GroundWaterBody.
SWRL:
Town(?x1) & WaterWell(?x2) & contains(?x1,?x2) & drawsFrom(?x2,?x3) → townGetsWaterFrom(?x1,?x3)
First-order logic (FOL):
$\forall x [\text{WaterWell}(x) \rightarrow \neg \text{OilWell}(x)]$
$\forall x, y [\text{drawsFrom}(x, y) \rightarrow \text{WaterWell}(x) \wedge \text{GroundWaterBody}(y)]$
$\forall x [\text{WaterWell}(x) \rightarrow \exists y [\text{drawsFrom}(x, y) \wedge \text{GroundWaterBody}(y)]]$
$\forall x [\text{HydroRockBody}(x) \wedge \forall y [\text{GroundWaterBody}(y) \wedge \text{submat}(y, x) \rightarrow \text{WellWaterBody}(y)] \rightarrow \text{WaterWell}(x)]$

engineering has shifted to maintaining and evolving existing ontologies (Kotis et al. 2020).

Types of Ontologies

To provide an overview of the broader ontology landscape, we survey it along three dimensions. The first is an ontology's purpose, that is, how it is intended to be used. The subsequent dimensions – scope and representation format – are partially correlated to the purpose (see Table 3).

Categorizing Ontologies by Purpose

Among the vastly different uses of ontologies of interest to the mathematical geosciences, five common purposes emerge: conceptual modeling, knowledge organization (including information retrieval), terminological disambiguation, standardization, and reasoning. Many ontologies, however, serve combinations or variations of these prototypical purposes.

Conceptual modeling (see Guizzardi 2010) is the process of brainstorming a domain or application in order to identify how the concepts therein relate to one another, creating *topic* or *concept maps* wherein concepts are related via unlabeled (generic) or labeled relationships. *Entity-Relationship (ER) diagrams* or *UML class diagrams* (examples in Fig. 1) – frequently used in software and database design – are more expressive yet still predominantly visual representations. Conceptual models are often integral parts of standards.

Knowledge organization and management (see also Jakus et al. 2013) encompasses various techniques to organize and help retrieve knowledge within knowledge organization systems (KOS), supporting metadata annotations, keyword-

based semantic search, and hierarchical navigation through knowledge. The utilized ontologies typically focus on classes and include formats such as *controlled vocabularies* and *glossaries* (lists of terms without definitions or relationships between them), *taxonomies* (i.e., classes organized using hypernyms), and *thesauri* (e.g., the Getty Thesaurus of Geographic Names, (<https://www.getty.edu/research/tools/vocabularies/tgn/>) which uses taxonomic relations and generic associations). Thesauri can be encoded using the SKOS language or more expressive languages, such as RDFS or OWL2 (<http://www.w3.org/TR/skos-reference>). The latter two are commonly used for sharing and connecting data as knowledge graphs, which support some basic semantic reasoning.

Terminological clarification and disambiguation is about precisely defining terms and, in the process, distinguishing and relating different *word senses* of terms (e.g., “well” for a bore hole to extract liquids/gases vs the space in a building containing a staircase) and nuances in its use (e.g., referring to a water or oil well; referring to the excavated hole only vs to the entire structure that includes container and liquid). A primary tool for disambiguation is natural language (“prose”) definitions. They are core to *dictionaries* but also incorporated into SKOS ontologies via the “skos:definition” attribute and into *WordNets* (Fellbaum 2006) that additionally relate words and word senses using semantic relationships such as synonymy and hyponymy. Ontologies specified in RDFS or OWL2 can also incorporate such definitions in comments (using “rdfs:comment”) but they are not formally interpreted.

Gazetteers, such as the USGS's Geographic Names Information System (GNIS) (<https://geonames.usgs.gov/>), are ontologies that clarify and disambiguate between geographic names using feature types (e.g., a town, county, or river),

meronomic relationships (e.g., the country or state where it is located), and precise locations (e.g., latitude and longitude).

Standardization is agreeing on a shared interpretation of terms for a specific domain to improve semantic interoperability. Because standards primarily guide the people developing such applications, nonformal representations consisting of conceptual models and natural language definitions of terms are most common, as exemplified by OGC's spatial data standards Simple Features (SFA) (<https://www.ogc.org/standards/sfa>) and GML (<https://www.ogc.org/standards/gml>) and their domain-specific standards such as GeoSciML (<https://www.ogc.org/standards/geosciml>) and WaterML (<https://www.ogc.org/standards/waterml>) for geologic and hydrologic data, respectively. Other standards additionally provide reference implementations or machine-interpretable versions, such as the OWL2 versions for Semantic Sensor Networks (SSN) (<https://www.w3.org/TR/vocab-ssn/>), PROV-O (<https://www.w3.org/TR/prov-o/>), and OWLTime (<https://www.w3.org/TR/owl-time/>). Such formalized standards are critical for increasing information integration and exchange across geoscience applications geosciences and, specifically, for constructing spatial data infrastructures (SDI) and knowledge graphs.

Reference ontologies (Menzel 2003) play important standardization roles. *Foundational* (or *upper*) *reference ontologies* (Guizzardi 2010) provide anchor concepts refinement by other ontologies, while *domain reference ontologies* (Hahmann and Stephen 2018) provide domain-specific reference concepts. Both kinds are discussed further in the categorization by scope.

Semantic reasoning is about inferring implicit knowledge and requires expressive formal ontologies – also called **knowledge representation ontologies** (Jakus et al. 2013) – specified using logic-based languages such as OWL2 (or other Description Logics), SWRL (<https://www.w3.org/Submission/SWRL/>) or first-order logic (or Common Logic) (https://standards.iso.org/ittf/PubliclyAvailableStandards/c066249_ISO_IEC_24707_2018.zip). By relating terms flexibly via logical axioms, these languages enable complex automated reasoning about the interaction of classes and relations, such as about the domain or range of objects participating in a relation, or the existence of specific relations for certain types of objects. Querying and reasoning with such ontologies is facilitated by SPARQL (<https://www.w3.org/TR/rdf-sparql-query/>) (for RDFS, OWL2) and GeoSPARQL (<https://www.ogc.org/standards/geosparql>), Description Logic reasoners, or first-order theorem provers and model finders.

Categorizing Ontologies by Scope

Ontologies can also be categorized by scope, that is, how broadly applicable they are, ranging from top-level ontologies, generic ontologies, domain and domain reference ontologies, to highly specialized application ontologies. With narrowing scope the ontologies' depth can increase.

Top-level (or **upper** or **foundational**) **ontologies**, such as BFO or DOLCE and surveyed by Mascardi et al. (2007), define high-level categories of objects according to a certain philosophical perspective. They are primarily used for reference purposes (Menzel 2003) with challenges and limitations discussed by Grüninger et al. (2014). Common distinctions are between *endurants* (also called *continuants*: entities that are entirely present at a given time), *perdurants* (also called *occurents*: entities that are by nature defined over time, such as events or processes) and *qualities* (e.g. a temperature, color, or spatial location). Endurants are sub-categorized into physically present ones, including material (e.g. towns or water bodies) and immaterial ones (a hole), and non-physical endurants such as social or mental objects. DOLCE further distinguishes abstract objects, such as a “space region”, while BFO emphasizes the difference between independent and dependent objects (e.g. boundaries and holes depend on a host object). Perdurants are commonly categorized into events, processes, and states. Foundational relations include *parthood* (temporal or spatial), *constitution* and *dependence* as well as relations linking perdurants to endurants, such as DOLCE's *participation* relation. Ongoing work streamlines the top-level ontologies, including a new modular one (TUpper: Grüninger et al. 2014) within a single ISO standard (Draft standard ISO/IEC: 21838 with Parts for BFO and proposed parts for DOLCE and TUpper.).

Generic ontologies (Grüninger et al. 2014) and **mid-level ontologies** encompass concepts that are used across multiple domains, such as pertaining to space, time, processes, events, observations, or the provenance of information. These generic concepts play critical roles in the mathematical geosciences and the ontologies are intended to be either reused by or used in conjunction with domain and application ontologies. While foundational ontologies also contain generic concepts, purpose-built generic ontologies flesh them out. For example, DOLCE and BFO know “space regions” whereas spatial ontologies like for the FOL formalizations of Simple Features and CODIB define entire hierarchies of different kinds of spatial regions based on dimension, curvature, and configuration (FOL and OWL2 axiomatizations available from http://colore.oor.net/simple_features/ and http://colore.oor.net/multidim_space_codib/).

Domain ontologies cover a specific domain of interest, but vary in domain breadth. For example, GeoSciML covers all kinds of geological formations, while the WaterML parts Hy_Features and GWML2 more narrowly focus on surface and subsurface hydrology, respectively. These are typical geoscience domain ontologies in that they cover a breadth of concepts and attributes for describing data in their respective domains. In comparison, a *domain reference ontology* (variably also called a *reference domain ontology* or *core ontology*), such as the hydro foundational ontology (HyFO; Hahmann and Stephen 2018) or the GeoSciMLBasic module

of GeoSciML, strictly focuses on the high-level concepts needed to integrate ontologies within their domains.

Application ontologies are tailored to specific software systems, for example to describe data models or to enable data retrieval. Their specific nature and ad-hoc development often prevents reuse elsewhere. Examples are ontologies specifically developed to identify rock formations for natural resource extraction, or the data model underlying the National Hydrographic Dataset (<https://www.usgs.gov/core-science-systems/ngp/national-hydrography/national-hydrography-dataset>) that encodes the US stream network and flows.

Categorizing Ontologies by Representation Format

A wide spectrum of representations that differ in (1) formality and (2) expressivity are available (Table 2) for the varied uses

of ontologies (as summarized by Table 3). With respect to formality, we distinguish formal formats from non-formal ones. Expressivity is about how granular and detailed the semantics of terms can be specified.

Formality: The least formal, mostly conceptual, ontologies use graphical notations, such as Entity-Relationship Models or UML class diagrams (examples in Fig. 1), that lack machine-interpretability but are appropriate for communication between people. Semi-formal ontologies include thesauri (e.g., WordNets or SKOS ontologies) that utilizes multiple explicit – though vaguely defined – semantic relationships. On the other end of the spectrum are **formal ontologies** specified using machine-interpretable languages such as RDFS, OWL2 (and other Description Logics), rule languages (SWRL), first- and higher-order logics (examples in Table 1).

Ontology, Table 2 Common types of ontologies by representation format in increasing order of expressivity

Nonformal ontologies	Formal ontologies
Topic Map, Concept Map: <i>Graphical representation of concepts with labeled (Concept Map) or unlabeled (Topic Map) relationships between the concepts</i>	Resource Description Framework Schema (RDFS): <i>XML-based notation for declaring classes, properties, and individuals and relating them via hyponymy relations (subclassOf, subPropertyOf) and domain and range restrictions of properties in terms of classes. Can alternatively be serialized as triples of the form “subject predicate object” in the widely used Turtle or N-Triples notations or in JSON-LD</i>
Entity-Relationship Model: <i>Graphical representation showing relationships between named classes of objects and named properties/relations</i>	
Controlled Vocabulary, Term List, Glossary: <i>A set or list of terms for a domain/application of interest, called a glossary if in alphabetic order</i>	Web Ontology Language (OWL 2): <i>An extension of RDFS that allows more complex class descriptions and constructs for restricting the interaction of properties and classes. For example, complex classes can be constructed using logical connectives (intersection, union, complement), using existential and universal quantification and cardinality restrictions</i>
Semi-formal ontologies	
Dictionary: <i>A set of terms (typically in alphabetic order) with concise natural language definitions</i>	
Taxonomy: <i>A set of classes related hierarchically via “is-a” (hyponymy/subclass) relationships</i>	Semantic Web Rule Language (SWRL): <i>Extends OWL with rules, that is, axioms of the form “antecedent (body) - > consequent (head)” where body and head are conjunctions of statements</i>
Thesaurus (including SKOS and WordNet): <i>A set of classes related via taxonomic and other semantic relationships such as synonymy or meronymy (part-of)</i>	First-order logic (FOL; including Common Logic): <i>Logic language that allows free-from axioms using standard logical connectives (not, and, or, if, iff) over a vocabulary of classes and relations of arbitrary arity. Common Logic provides additional constructs for, e.g., specifying axiom schemata and or to scope axioms to specific classes</i>
UML Class Diagram: <i>Graphical representation that enhances ER models by taxonomic and meronomic relationships.</i>	

Ontology, Table 3

Common representation formats for ontologies and their most common uses

Representation format	Common uses (by purpose)				
	Modeling	KOS	Term. C & D	Standards	Reasoning
Topic Map, Concept Map	X				
Entity-Relationship Model	X				
Controlled Vocabulary, Glossary		X		X	
Dictionary		X	X	X	
Taxonomy	X	X	X	X	
UML Class Diagram	X	X		X	
WordNet (Thesauri)			X		
SKOS (Thesauri)		X	X		
RDF-S		X	X	X	X
OWL2				X	X
SWRL					X
FOL				X	X

Expressivity: Non-formal ontologies range in expressivity from providing a single, unnamed generic semantic relationship as in topic maps to thesauri that incorporate multiple semantic relationships such as WordNet’s synonymy and synonymy and SKOS’ “closeMatch” and “broaderMatch”.

Formal ontologies exhibit more pronounced differences in expressivity, which gradually increases from RDFS to OWL2 and further to SWRL and first-order logic (FOL, Common Logic) and higher-order logics. Because they use a formal logic as representation language, DL and FOL ontologies are called **axiomatic ontologies**, whereas non-axiomatic ones are sometimes called “lightweight ontologies.” The term “axiomatic” emphasizes that they are not reliant upon predefined semantic relations but allow arbitrarily complex logical sentences to be used as axioms to constrain the semantics of terms. This vastly increases expressivity and the kind of queries that can be posed.

Summary

Ontologies are used in a multitude of ways by mathematical geosciences and related scientific communities (see also the survey on geospatial semantics by Kokla and Guilbert 2020): They help conceptualize domains, standardize terminology, enable information organization and retrieval, and support complex semantic reasoning. Ontologies also differ widely in scope and representation formats. Among the available representation formats, formal ontologies specified in languages such as RDFS, OWL2, and more expressive languages experience increased uptake by the community as they afford the greatest semantic clarity and support the overall trend toward semantically enabled, knowledge-intensive science.

Cross-References

- [Graph Mathematical Morphology](#)
- [Metadata](#)

Bibliography

- Brodaric B, Hahmann T, Gruninger M (2019) Water features and their parts. *Appl Ontol* 14(1):1–42. <https://doi.org/10.3233/AO-190205>
- Fellbaum C (2006) WordNet(s). In: *Encyclopedia of language & linguistics*, vol 13, 2nd edn. Elsevier, Amsterdam, pp 665–670
- Grüniger M, Hahmann T, Katsumi M, Chui C (2014) A sideways look at upper ontologies. In: *Formal ontology in information systems*. IOS Press, Amsterdam, pp 9–22
- Guizzardi G (2010) Theoretical foundations and engineering tools for building ontologies as reference conceptual models. *Semantic Web J* 1:3–10

- Hahmann T, Stephen S (2018) Using a hydro reference ontology to provide improved computer-interpretable semantics for the groundwater markup language (GWML2). *Int J Geogr Inf Sci* 32(6): 1138–1171
- Jakus G, Milutinović V, Omerović S, Tomažič S (2013) *Concepts, ontologies, and knowledge representation*. Springer, New York
- Kokla M, Guilbert E (2020) A review of geospatial semantic information modeling and elicitation approaches. *ISPRS Int J Geo Inf* 9(3):146. MDPI
- Kotis K, Vouras G, Spiliotopoulos D (2020) *Ontology engineering methodologies for the evolution of living and reused ontologies: status, trends, findings and recommendations*. *Knowl Eng Rev* 35: 1–34. Cambridge University Press
- Mascardi V, Cordi V, Rosso P (2007) A comparison of upper ontologies. Workshop dagli Oggetti agli Agenti, Genova, Italy, Seneca Edizioni Torino, pp 55–64
- Menzel C (2003) Reference ontologies – application ontologies: either/or or both/and? *Proceedings of the KI2003 Workshop on Reference Ontologies and Application Ontologies*, Hamburg, Germany, September 16, 2003, CEUR Workshop Proceedings 94, <http://ceur-ws.org/Vol-94,CEUR-WS.org>
- Staab S, Studer R (2009) *Handbook on ontologies*, 2nd edn. Springer, Berlin

Open Data

Lucia R. Profeta
Lamont-Doherty Earth Observatory, Columbia University in the City of New York, New York, NY, USA

Definition

The summary definition of *Open Data* is “data that can be freely used, modified, and shared by anyone for any purpose” (OKF 2015).

The acceptance of what qualifies as *open* is under constant revision. The first definition in its modern use was created by the Open Knowledge Foundation in 2005 and its most recent iteration (version 2.1) was established in 2015. Changes to this definition were made using community feedback in order to ensure improved reusability, licensing, and attribution of the data.

Prerequisites of Open Data

The Open Definition (OKF 2015) describes in rigorous detail what elements are required, recommended, or optional to meet the criteria for Open Data. This poses conditions for the data itself (open works) and the license that will be associated with the data.

There are four criteria that the data must satisfy in order to classify as open: (1) it must have an open license, or be freely available in the public domain, (2) it must be accessible at most at a one-time reproduction cost and should be downloadable from the internet, (3) it needs to be compiled in a way

that makes it machine readable, and (4) the format in which the data is shared should be nonproprietary.

The license associated with Open Data must comply with the conditions of open licenses as outlined in Fig. 1. The differences between open licenses are important to understand, as they impose specific rights and responsibilities, both for data creators, as well as users.

Through a general licensing process, Open Data can be freely usable without any monetary compensation to its creator. It can be redistributed or sold, added to collections, and partial or derivative data must be distributed under the same license. Moreover, there must be no discrimination against any person or group, and all uses of the data are acceptable. The aforementioned conditions can only be changed under the following circumstances, which may be imposed by the license agreement:

- (i) Sharing the data requires that the creators will be cited (attribution).
- (ii) If any modifications are made to the data they will be clearly stated, or the name of the dataset will be changed or versioned (integrity).
- (iii) Redistributing the work will be done with the same license type (share-alike).
- (iv) Copyright and license information must be preserved (notice).
- (v) Distribution will be done in a predefined format (source).
- (vi) No technical restrictions will be imposed on distribution.
- (vii) Additional permissions and allowed rights can be granted to the public in cases such as patents (nonaggression).

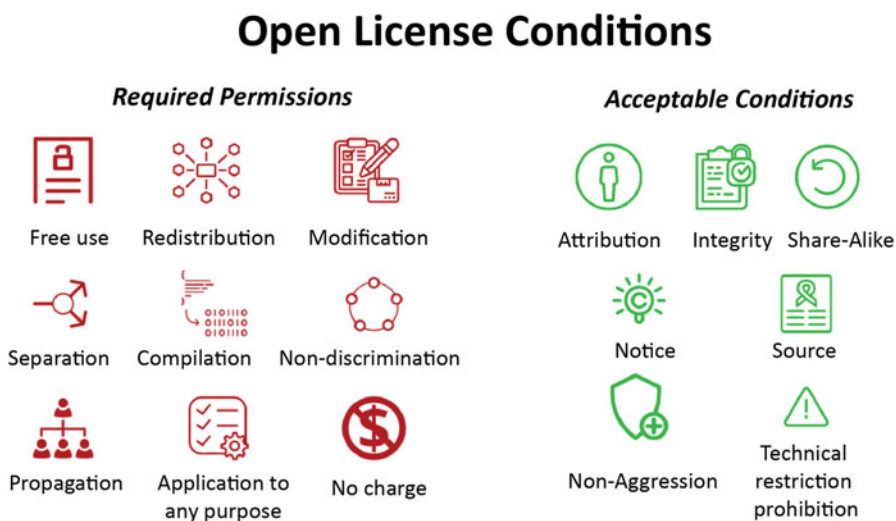
Principles Surrounding Open Data

The Open Data ecosystem incorporates multiple participants throughout the data life cycle at many levels, such as governments, funding agencies, researchers, publishers, and end users. Some guiding principles have emerged, which aim to govern the complex rights and responsibilities of all stakeholders.

The International Open Data Charter (2015) put forth six principles to inform the use of Open Data with an emphasis on how government data should be shared in order to create more transparency and international collaboration. These principles stated that government data should be “*Open By Default, Timely and Comprehensive, Accessible and Usable, Comparable and Interoperable, For Improved Governance & Citizen Engagement, and For Inclusive Development and Innovation.*”

The need for regulating how data is shared was also identified as a priority within the scientific community, which led to the emergence of the FAIR Principles (*Findable, Accessible, Interoperable, Reusable*; Wilkinson et al. 2016). The FAIR principles provided a general framework for how Open Data should be shared and the importance of machine readability. They also recognize the need for data repositories to capture rich **metadata** – information that would accompany the data in order to make it more reusable. To further define the responsibilities of data repositories, a successor to the FAIR Principles was created: the TRUST Principles (*Transparency, Responsibility, User focus, Sustainability, and Technology*; Lin et al. 2020). Another complement to the FAIR principles is represented by the CARE Principles (*Collective Benefit, Authority to control, Responsibility, Ethics*; RDA 2019), a set of recommendations which aim to recognize and protect the value of Indigenous data.

Open Data, Fig. 1 Open License Conditions based on The Open Definition v 2.1 (OKF 2015)



State of Open Data

Since 2016, reports on the State of Open Data (Hyndman 2018) have been generated based on surveys of the scientific community. Researchers have been asked to share their views regarding their perception of Open Data.

The emerging trends show that more researchers are sharing their data, but fewer of them are reusing openly available data. There has been a declined interest among those that have not yet reused Open Data to start doing so. Of those researchers that have shared their data openly, almost half did not know under what license their data was shared. There is an increasing desire for more guidance and regulations to be put in place. The majority of researchers that were surveyed indicated that a national mandate for making primary research openly available is necessary (Fig. 2).

There are many organizations and collaborative projects that are working toward establishing more robust standards and best practices for data sharing and attribution. Initiatives such as Data Together (GO FAIR 2020) combine expertise from international organizations to ensure that the Open Data research ecosystem is governed, implemented, and maintained in a sustainable way.

Linked Open Data

The concept of Linked Data was introduced by Tim Berners-Lee (2006) in the W3C community as a way to regulate data sharing through the Semantic Web: Linked Data should have

a Uniform Resource Identifier (URI) that uses HTTP protocol, links to other URIs, and it should use standardized querying languages such as SPARQL or the Resource Description Framework (RDF). If the Linked Data uses open sources, it becomes Linked Open Data (LOD).

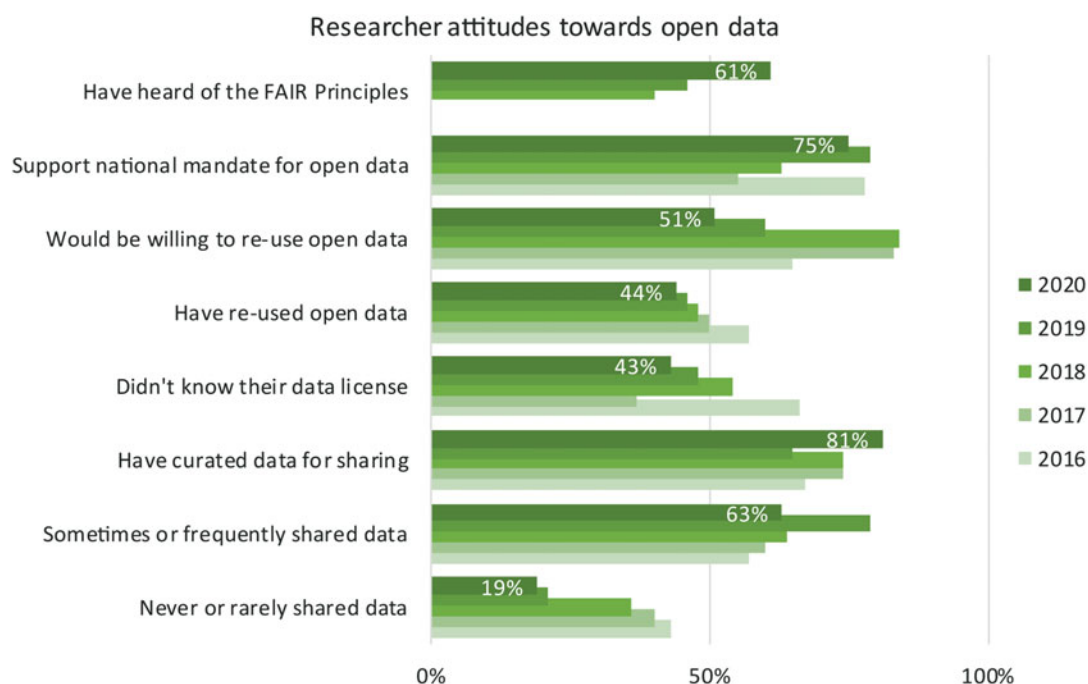
LOD adds value and increases the trustworthiness of data making it very powerful. When combined with open interfaces (APIs) to facilitate automation, LOD allows for better integration, discoverability, and access to complex data.

Open Data in Geosciences

From a data perspective, Earth and Space Sciences are made up of disparate long tail subdomains – data is primarily shared in discreet amounts, with not enough attention being given to interoperability or reusability.

Both technical infrastructure and governance initiatives regarding data stewardship in the geosciences have been gaining more momentum over the last decade. A variety of Open Data providers, tools, and platforms has emerged and is being maintained by large public funding agencies (see Ma 2018 for review). Requirements for newly published data to be open and interoperable by default, as well as legacy data rescue efforts are leading to large amounts of Open Data.

Technological advancements such as cloud computing and high-performance computing (HPC) are enabling large scale reuse of Open Data. Software deployments such as the Pangeo Environment (Odaka et al. 2020) offer solutions for



Open Data, Fig. 2 Trends emerging from surveys of researcher attitudes toward Open Data between 2016 and 2020. (Data from Hyndman 2018)

big data geoscience research that rely on, and concurrently generate, Open Data. Public access to these types of technologies is promoting open science on scales never seen before, facilitating cross-domain research and international collaboration.

Summary or Conclusions

Open Data is the most important building block of open science and scholarship. Big data studies and cloud computing rely on having a steady supply of high quality, metadata-rich, freely accessible, and machine-readable data. This has been historically a burden on data creators, but we are witnessing a perception shift, where it is understood that this is a shared responsibility. Better practices regarding data dissemination are being implemented on every level, from education to publication. The integrity of our data is the key to our shared knowledge – there is an inherent duty to preserve it, and ensure that it is open to all.

Cross-References

- [Data Life Cycle](#)
- [FAIR Data Principles](#)

Bibliography

- Berners-Lee T (2006) Linked data – design issues, world wide web consortium (W3C), London 2006. <https://www.w3.org/DesignIssues/LinkedData.html>. Accessed 05 Nov 2020
- GO FAIR (2020) Data together statement. <https://www.go-fair.org/2020/03/30/data-together-statement/>. Accessed 05 Nov 2020
- Hyndman A (2018) State of open data. <https://doi.org/10.6084/m9.figshare.c.4046897.v4>. Accessed 28 Nov 2020
- International Open Data Charter (2015) Principles. <https://opendatacharter.net/principles/>. Accessed 6 Dec 2020
- Lin D, Crabtree J, Dillo I et al (2020) The TRUST principles for digital repositories. *Nat Sci Data* 7:144. <https://doi.org/10.1038/s41597-020-0486-7>
- Ma X (2018) Data science for geoscience: leveraging mathematical geosciences with semantics and open data. In: *Handbook of mathematical geosciences: fifty years of IAMG*. Springer International Publishing, Cham, pp 687–702
- Odaka TE, Banihirwe A, Eynard-Bontemps G, Ponte A, Maze G, Paul K, Baker J, Abernathy R (2020) The Pangeo ecosystem: interactive computing tools for the geosciences: benchmarking on HPC. In: *Communications in computer and information science*. Springer, pp 190–204
- Open Knowledge Foundation – OKF (2015) Open Data Handbook. <https://opendatahandbook.org/>. Accessed 11 Oct 2020
- Research Data Alliance International Indigenous Data Sovereignty Interest Group (2019) CARE principles for indigenous data governance. The Global Indigenous Data Alliance. [GIDA-global.org](https://gida-global.org)
- Wilkinson M, Dumontier M, Aalbersberg I et al (2016) The FAIR guiding principles for scientific data management and stewardship. *Nat Sci Data* 3:160018. <https://doi.org/10.1038/sdata.2016.18>

Optimization in Geosciences

Ilyas Ahmad Huqqani and Lea Tien Tay
School of Electrical and Electronics Engineering, Universiti Sains Malaysia, Penang, Malaysia

Abbreviations

ACO	Ant colony optimization
ANN	Artificial neural network
BBO	Biogeography-based optimization
DE	Differential evolution
EBF	Evidential belief function
FPAB	Flower pollination by artificial bee
FR	Frequency ratio
GA	Genetic algorithm
GIS	Geographical information system
GS	Geoscience
GWO	Gray wolf optimization
LR	Logistic regression
P	class Polynomial time class
PSO	Particle swarm optimization
NP	Nondeterministic polynomial time
SVM	Support vector machine
SVR	Support vector regression
RF	Random forest
WoE	Weight of evidence

Definitions

Geoscience: The study of earth is called as geoscience. It investigates the processes that shape the earth's surface, the natural resources that we utilize, and how water and ecosystems are interrelated. Geoscience encompasses so much more than rocks, debris, and volcanoes.

Optimization: A process that takes the action of making the best or most effective use of a situation or resource.

Optimization problem: Maximizing or minimizing a specific function in relation to a set, which frequently represents a set of possibilities in a given circumstance. The function makes it possible to compare several possibilities that could be the “best.”

Background

The world is confronted with enormous problems that have ramifications for global socioeconomic performance and quality of life. To prevent and overcome the chaos created by global change, the new solutions are of utmost needed. In order to achieve this aim, better solutions need to be

developed to optimize the management of natural resources, identify new ways to exploit geophysical processes and characteristics, and apply geoscience knowledge to explore new and optimized methods of utilizing terrestrial resources. One of the most challenging tasks in modelling of geophysical and geological processes is to address large data and achieve specified objectives. Therefore, it can be considered as an optimization problem, which allows the use of several new effective optimization techniques. This article provides insight into how optimization can be utilized in geoscience study and how optimization and geoscience should coexist in the future.

What is Optimization?

Optimization is a search process which makes an object or system as useful or effective as feasible. It refers to determine the optimal solutions to meet a given objective subject to certain conditions. The purpose of optimization is to find the best possible solutions while taking into account the problem constraints. Optimization approaches are used to address the complicated scientific and engineering problems. The use of optimization techniques is vital due to time consumption and complexity of existing methods. Various complex engineering problems that demand computational precision in short time cannot be solved using traditional approaches. In this scenario, the optimization techniques are most effective approaches to find the best possible solution of such problems.

Initially, the heuristic optimization techniques explore randomly in the problem search space for the optimal solution with limited computation time and memory without requiring any complex derivatives (Osman and Laporte 1996). The term “heuristic” originated from the Greek word “heuriskein,” which means “to find.” In computing, the heuristic technique works on a rule-of-thumb for providing a solution to the problem despite of considering the repetitive application of the method. However, this method only searches for an estimated solution and does not require any mathematical convergence proof.

The most common and advanced heuristic optimization technique utilized in the context of addressing search and optimization issues is the metaheuristic optimization algorithm that was first proposed by Glover (1986). This method denotes a technique that employs one or more heuristics and hence possesses all the aforementioned features of heuristic method. As a result, a metaheuristic technique has many properties such as it tries to discover a near optimal solution rather than attempting to find the exact ideal solution, typically lacks a formal demonstration of convergence to the optimal solution, and lastly, it requires less time to compute than other exhaustive searches. These approaches are iterative in nature, and frequently utilize stochastic processes to alter

one or more original candidate solutions during the search process (usually generated by random sampling of the search space). Due to the inherent practicality of many real-world optimization issues, classical optimization algorithms may not always be relevant and may not perform well in tackling such problems in a pragmatic manner. Despite disregarding the importance of classical algorithms in the optimization field, many researchers and optimization practitioners realized that the metaheuristic methods are able to obtain a near-optimal solution in a computationally tractable manner. Therefore, the metaheuristic techniques have become the most popular optimization technique in recent years due to their capacity to handle a wide range of practical problems and produce an authentic and reasonable solution.

The most common inspirations for metaheuristic techniques are natural, physical, or biological principles that can be imitated at a fundamental level using a variety of operators. The balance between exploration and exploitation is a recurring subject in all metaheuristics. The term “exploration” relates to how successfully operators diversify solutions in the search space. This feature provides the metaheuristic with a global search behavior. Moreover, the ability of the operators to use the knowledge provided from earlier iterations to increase search is referred as exploitation. The metaheuristic gains a local search feature as a result of this intensification. Some metaheuristics are more explorative than exploitative, whereas others are the inverse. The fundamental approach of randomly selecting answers for a set number of iterations, for example, constitutes a totally exploratory search. Hill climbing, on the other hand, is an example of totally exploitative search in which the present answer is progressively modified until it improves. Metaheuristics allow the user to balance between diversification and intensification via operator settings. Based on these characteristics, metaheuristic optimization algorithms have various advantages as compared to the classical optimization algorithms:

- Metaheuristic algorithms can deal with the P class issues easily under significant complex inputs which may be a challenge for other conventional techniques.
- Metaheuristics are useful in solving difficult NP problems which cannot be solved by any known algorithm in an acceptable time period.
- Metaheuristics, unlike most traditional techniques, do not require gradient knowledge and may thus be utilized with non-analytic, black box, or simulation-based objective functions.
- Due to the inherent stochasticity and deterministic behaviors, many metaheuristic algorithms have the ability to recover from local optima.
- Metaheuristics can also handle uncertainties occurred in objectives in a better way as compared to the classical

optimization methods because of the ability to recover from local optima.

- Lastly, most metaheuristics can accommodate multiple objectives with relatively minor algorithmic modifications.

Classification of Metaheuristic Optimization Algorithms

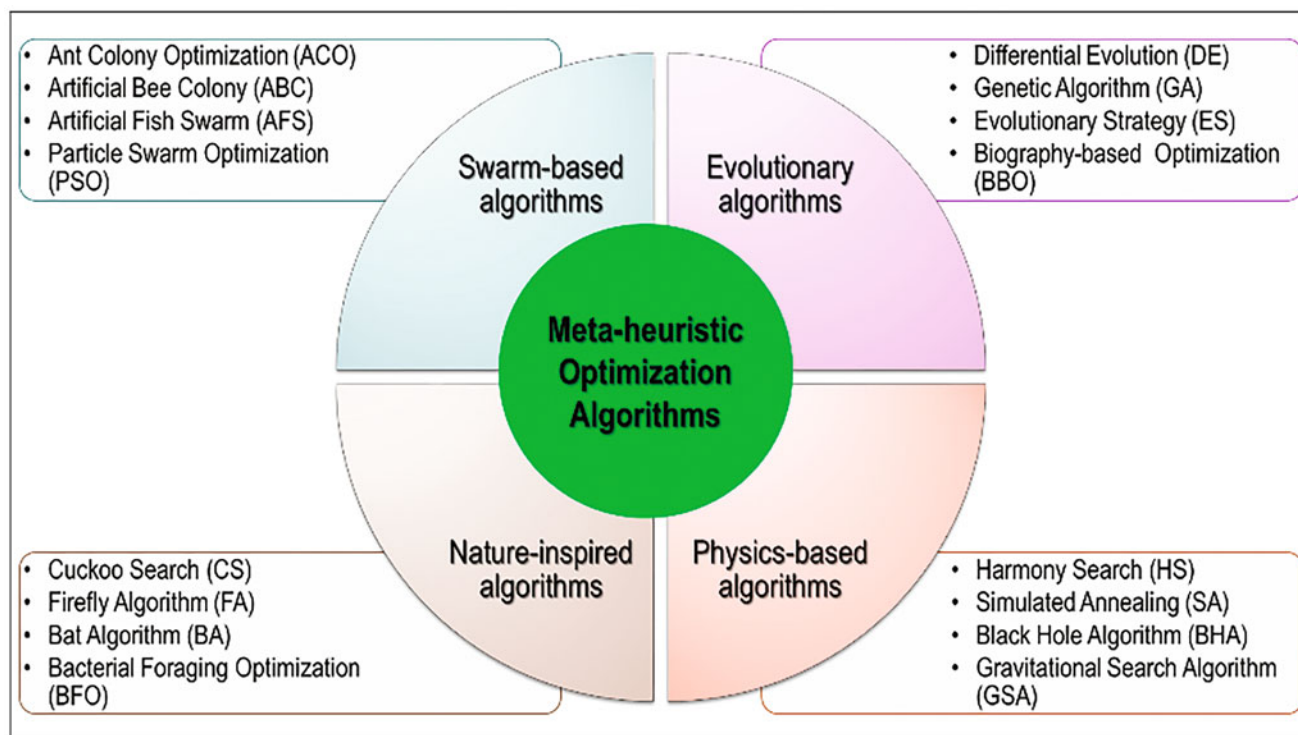
The most popular method of categorizing metaheuristic methods is based on the number of original solutions modified in subsequent rounds. Single-solution metaheuristics begin with a single starting solution that is iteratively modified. It is important to note that while the modification process may involve more than one solution, only one solution is employed in each subsequent iteration. On the other hand, population-based metaheuristics employ more than one starting solution. Multiple solutions are modified in each iteration, and some of them make it to the next iteration. The operators are used to modify solutions, which frequently employ particular statistical features of the population.

Metaheuristic optimization algorithms can also be classified on the basis of the domain in which they operate. These algorithms are commonly referred as “bioinspired” or “nature-inspired” umbrella words. They can, however, be

further subdivided into four categories (Fig. 1): evolutionary algorithms, swarm intelligence algorithms, physical phenomena algorithms, and nature-inspired algorithms. The evolutionary algorithms simulate many elements of natural evolution, such as survival of the fittest, reproduction, and genetic mutation. Swarm intelligence and nature-based algorithms replicate the group behavior and/or interactions of live species (such as ants, bees, birds, fireflies, glowworms, fish, white blood cells, bacteria, and so on) (Huqqani et al. 2022) as well as nonliving objects (like water drops, river systems, masses under gravity). The remaining metaheuristics, classified as physics-based algorithms, imitate different physical processes such as metal annealing, musical aesthetics (harmony), and so on.

Practical Applications of Optimization Algorithms in Geoscience

The optimization algorithms have been widely used in various practical applications related to geoscience such as mineral and water exploration, geological and geo-structural mapping, natural hazards analysis, earthquake prediction and control, and many more. A few of the applications are discussed in the section.



Optimization in Geosciences, Fig. 1 Classification of metaheuristic optimization algorithms

Land Cover Feature Extraction

The term “land cover feature extraction” refers to the process of excerpting features from images of land in diverse locations. This is useful for a variety of different difficulties, such as detecting ground water, locating a path in mountainous terrain, and so on. For the aim of land cover feature extraction in the Alwar area of Rajasthan, methods such as fuzzy set, rough-fuzzy tie-up, membrane computing, and BBO were used. The pixels in the image are classified into five categories: rocky, water, urban barren, and flora. On the same Alwar dataset, hybrid versions of several methods have also been implemented. The algorithms included hybrid of ACO-BBO, hybrid of ACO-PSO-BBO, hybrid of FPAB-BBO, and hybrid of biogeography-based optimization and geoscience (BBO-GS) (Goel et al. 2013). As a result, BBO-GS has shown to be the top performer for the task among all the implemented algorithms.

Groundwater Mapping Analysis

One of the most important case studies of geoscience applications is the groundwater analysis. Water resources, including surface water and groundwater, are vital for life on our planet that are regenerated via evaporation, precipitation, and surface runoff. According to recent climate change estimations, the water cycle will become more spatially and temporally heterogeneous, resulting in water demand exceeding supply (*Intergovernmental panel on climate change* 2014). Groundwater is described as saturated water that fills the pore spaces between mineral grains, cavities, and fractured rocks in a rock mass. According to the World Economic Forum, scarcity of water will be a worldwide issue in the future. Consequently, about 20% of water utilized by humans is sourced from groundwater, and this proportion is expected to rise over the next few decades (Biswas et al. 2020). The extensive use of groundwater in agriculture, industry, and daily life causes several challenges to its management.

Generally, the hydrological testing, field surveys, and geophysical techniques are commonly used in groundwater research projects. These courses of action are time consuming, costly, and necessitate the use of skilled personnel. As a result, techniques for analyzing groundwater that may illustrate the hydrological connection with groundwater, such as the use of data-specific capacity, transmissivity, and yield, must be developed. Groundwater potential is often described as optimum zones for groundwater development or the likelihood of groundwater presence (Díaz-Alcaide and Martínez-Santos 2019). The likelihood of groundwater occurrence in a particular region is estimated using groundwater potential mapping which entails statistical analysis of many forms of field data. Remote sensing data acquiring techniques increase spatial coverage, thus improving the variety and availability

of data. GIS technology may be utilized to analyze vast regions at lower cost. Preliminary GIS studies utilizing machine learning and statistical techniques, among other things, might be beneficial for analyzing groundwater availability based on topographical and geographical parameters. As a result, prospective ground-water wells may be plotted to aid in groundwater detection.

With the rapid growing and complexity of data in GIS, a trustworthy model was necessary to assist in groundwater issue. Several models have been proposed to assess groundwater potential, including FR, WoE, and EBF models, as well as machine learning models such as ANN, RF, LR, and SVM. Some of these studies still have limitations in their predictions, which are impacted by the quality of the data collection and the internal structure of the model (Chen et al. 2019). Fadhillah built the groundwater potential mapping using a machine learning technique based on SVR with training data obtained from a hydraulic transmissivity dataset (Fadhillah et al. 2021). The learning process and the consistency of model outcomes can be influenced by SVR parameter settings. As a result, defining operational parameters becomes a problem in achieving the desired outcomes. To overcome this problem, he applied two metaheuristic optimization algorithms: GWO and PSO. GWO is an optimization method that has been frequently employed in GIS applications to optimize models. Due to its outstanding features above other swarm intelligence methods, it has been widely adapted for a broad variety of optimization problems (Faris et al. 2018). It has relatively few parameters, and no derivation information is necessary in the first search. Similarly, PSO is regarded as dependable in the computational process for model optimization since it provides a gentle computation and rapid convergence. It is believed that by employing the optimization technique, the groundwater model's performance would increase. By adjusting the SVR machine learning parameters, the metaheuristic optimization technique has the potential to improve model prediction accuracy.

Oil and Gas Reservoir Exploration

In resolving geophysical optimization issues, optimization algorithms have been utilized in two ways: either by performing the optimization or by optimizing parameters of other techniques (e.g., neural networks) used in various problems.

In a study on reservoir model optimization to match previous petroleum production data, a few optimization techniques are compared to PSO (Hajizadeh et al. 2011). In a case study involving two petroleum reservoirs, ACO, DE, PSO, and the neighborhood algorithm are combined in a Bayesian framework to assess the uncertainty of each algorithm's predictions. Ahmadi predicted reservoir permeability

using a soft sensor based on a feed-forward artificial neural network, which was subsequently improved using a hybrid GA and PSO approach (Ali Ahmadi et al. 2013). A multi-objective evolutionary method discovers optimum solutions and surpasses a standard weighted-sum technique.

In a research of oil and gas fields, Onwunalu and Durlofsky used PSO to find the best well type and location (Onwunalu and Durlofsky 2010). The comparisons with a GA across several runs of both algorithms reveal that PSO surpasses the GA on average, although the benefits of using PSO over GA may vary depending on the scenario. Well placement problems were also considered by Nwankwor using a hybrid PSO-DE method (Nwankwor et al. 2013). Hybridized metaheuristic optimization algorithms are potentially beneficial in reservoir engineering problem.

Summary

In summary, there are a lot of optimization problems in geosciences, and they can be solved using metaheuristic optimization or hybridized metaheuristic optimization algorithms. Besides, optimization algorithms also play a vital role in determine the best parameter for other machine learning techniques.

Bibliography

- Ali Ahmadi M, Zendeboudi S, Lohi A, Elkamel A, Chatzis I (2013) Reservoir permeability prediction by neural networks combined with hybrid genetic algorithm and particle swarm optimization. *Geophys Prospect* 61(3):582–598
- Biswas S, Mukhopadhyay BP, Bera A (2020) Delineating groundwater potential zones of agriculture dominated landscapes using GIS based AHP techniques: a case study from Uttar Dinajpur district, West Bengal. *Environ Earth Sci* 79(12):302
- Chen W, Tsangaratos P, Ilia I, Duan Z, Chen X (2019) Groundwater spring potential mapping using population-based evolutionary algorithms and data mining methods. *Sci Total Environ* 684:31–49
- Díaz-Alcaide S, Martínez-Santos P (2019) Review: advances in groundwater potential mapping. *Hydrogeol J* 27(7):2307–2324
- Fadhillah MF, Lee S, Lee CW, Park YC (2021) Application of support vector regression and metaheuristic optimization algorithms for groundwater potential mapping in Gangneung-si, South Korea, Remote Sensing. 13(6)
- Faris H, Aljarah I, Al-Betar MA, Mirjalili S (2018) Grey wolf optimizer: a review of recent variants and applications. *Neural Comput & Applic* 30(2):413–435
- Glover F (1986) Future paths for integer programming and links to artificial intelligence. *Comput Oper Res* 13(5):533–549
- Goel L, Gupta D, Panchal VK (2013) Biogeography and geo-sciences based land cover feature extraction. *Appl Soft Comput* 13(10): 4194–4208
- Hajizadeh Y, Demyanov V, Mohamed L, Christie M (2011) Comparison of evolutionary and swarm intelligence methods for history matching and uncertainty quantification in petroleum reservoir models. In: Köppen M, Schaefer G, Abraham A (eds) *Intelligent computational optimization in engineering: techniques and applications*. Springer, Berlin Heidelberg, pp 209–240
- Huqqani IA, Tay LT, Mohamad-Saleh J (2022) Spatial landslide susceptibility modelling using metaheuristic-based machine learning algorithms. *Eng Comput*
- (2014) Intergovernmental panel on climate change. In: *Climate change 2013 – the physical science basis: working group I contribution to the fifth assessment report of the intergovernmental panel on climate change*. Cambridge University Press, Cambridge
- Nwankwor E, Nagar AK, Reid DC (2013) Hybrid differential evolution and particle swarm optimization for optimal well placement. *Comput Geosci* 17(2):249–268
- Onwunalu JE, Durlofsky LJ (2010) Application of a particle swarm optimization algorithm for determining optimum well location and type. *Comput Geosci* 14(1):183–198
- Osman IH, Laporte G (1996) Metaheuristics: a bibliography. *Ann Oper Res* 63(5):511–623

Ordinary Differential Equations

R. N. Singh¹ and Ajay Manglik²

¹Discipline of Earth Sciences, Indian Institute of Technology, Gandhinagar, Palaj, Gandhinagar, India

²CSIR-National Geophysical Research Institute, Hyderabad, India

Definition

Ordinary differential equation – a differential equation that contains unknown functions of one independent variable and their ordinary derivatives.

Introduction

An ordinary differential equation (referred as ODE in literature) is an equation that contains the unknown variable y and its ordinary derivatives with respect to time (or space) t , which in general form can be expressed as

$$F(t, y, y', y'' \dots, y^{(n)}) = 0, \quad (1)$$

where $y' = dy/dt$, $\dots$, $y^{(n)} = d^n y/dt^n$. Here, n represents the highest order of the derivative and hence the order of the ODE. For a unique solution of this equation, we need n conditions. When these can be prescribed at a value of t , it is called initial value problem, and when these are prescribed at two values of t , it is called two-point boundary value problem. This equation can be written as a system of first-order differential equation as $dy/dt = f(y, t)$, where y is a n -dimensional vector.

In initial value problem, the values of y are given at $t = t_0$, $y(t_0)$. Initial value problem is well posed as its solution exists,

is unique, and depends continually on the given data in case both functions f and df/dy are continuous (Coddington 1989). The literature on this subject is vast, but in this entry, we focus on some ODEs used in geosciences and show their applications in solving problems relevant to understanding earth's evolution and processes.

Geosciences deal with the structure, processes, and evolution of earth, the knowledge of which is constructed by using physical, chemical, and biological laws and interpreting observations made above, on, and within the earth. ODEs play a very important role in improving our understanding of the earth, especially when long-term behavior of some aspect of the earth is to be studied, e.g., temporal evolution of the earth, because other spatial variations can be averaged to reduce a generalized model, represented by a partial differential equation (PDE) (see ► “Partial Differential Equations”), to an ODE. These are so-called box models which pervade earth studies, say in biogeochemical cycles. ODEs also arise when PDEs involved in spatiotemporal studies of the earth are decomposed into a set of ordinary differential equations. Although one can have a generalized n th order ODE (Eq. 1), most of geosciences problems can be described by first- and second-order ODEs. Therefore, we discuss linear first- and second-order ODEs in the next section. We then cover nonlinear ODEs which are used in system's thinking. This is followed by discussion on various solution methods. Finally, we illustrate several examples of ODEs used in geosciences.

First- and Second-Order Linear ODEs

First-Order ODEs

We frequently encounter growth and decay problems. Such problems are solutions of the first-order ODEs. We get the following first-order inhomogeneous equation in general case, for instance, in Newton's cooling problem used in the study of cooling of earth and geological bodies:

$$\frac{dy}{dt} = f(t)y + g(t), \quad (2)$$

where, $f(t)$ and $g(t)$ are time-dependent coefficients in general form.

The solution of this equation is obtained by first multiplying both sides by $e^{-\int f(t')dt'}$ and then rearranging the equation as

$$\frac{d}{dt} \left(e^{-\int f(t')dt'} y \right) = e^{-\int f(t')dt'} g(t), \quad (3)$$

which has the general solution as

$$y(t) = C e^{\int f(t')dt'} + e^{\int f(t')dt'} \int e^{-\int f(t')dt'} g(t'') dt''. \quad (4)$$

The value of the constant of integration C depends on the choice of an initial condition, i.e., $y|_t = 0$.

Second-Order ODEs

Second-order ODEs are more prevalent in geosciences as most physical laws are described by second-order equations, e.g., oscillatory phenomena, equations for which are expressed as

$$\frac{d^2 y}{dt^2} + a \frac{dy}{dt} + by = f(t), y(0) = c, \frac{dy}{dt}(t=0) = d, \quad (5)$$

where the coefficients a and b are taken as constants and $f(t)$ is the source function.

The solution of inhomogeneous equation comprises of general solution given by the homogeneous part, also called complementary function, and particular solution of the equation. The solution for the homogeneous part of the equation can be written in terms of elementary functions. The solution is assumed as

$$y(t) = C e^{mt}, \quad (6)$$

where C and m are constants. Substituting this equation in Eq. (5), we get the following polynomial equation

$$m^2 + am + b = 0. \quad (7)$$

This has two roots m_1 and m_2 , which can be real or complex conjugate. Thus, the solution is written as

$$y(t) = C_1 e^{m_1 t} + C_2 e^{m_2 t}. \quad (8)$$

When both roots are the same ($m_1 = m_2$), the solution is given as

$$y(t) = (C_1 + C_2 t) e^{m_1 t}. \quad (9)$$

For complex roots such as $m_1 \pm m_2$, the solution is given by

$$y(t) = e^{m_1 t} (C_1 \cos m_2 t + C_2 \sin m_2 t). \quad (10)$$

The particular solution is obtained by using the method of trial and error in simple cases when the right-hand side is made of power, exponential or trigonometric functions. Some general forms of particular solution taken are C , $Ct + D$, $Ct^2 + Dt + E$, Ce^{pt} , $C \cos pt + D \sin pt$.

In general case, the method of variation of parameters is used to get a particular solution. The particular solution is assumed as linear combination of solutions of homogeneous equation

$$y_p = C_1(t)y_1 + C_2(t)y_2. \quad (11)$$

This is substituted in the defining equation and resulting equation for $C_1(t)$, and $C_2(t)$ is solved. The values of constants are determined by the initial conditions.

Nonlinear Equations

Three such equations under this category are the logistic equation, the predator-prey equations, and the Lorenz equation. In the first two cases, analytical solution can be written whereas in the last case no analytical solution has been found so far.

The logistic equation, used in all population studies, is written as

$$\frac{dy}{dt} = ry\left(1 - \frac{y}{K}\right), \quad y(0) = y_0. \quad (12)$$

Here, K is called carrying capacity, which is the maximum value of the population growing with maximum rate of growth as r , called Malthusian parameter. The solution of this equation is

$$y = \frac{K}{1 + \left(\frac{K}{y_0} - 1\right)e^{-rt}}. \quad (13)$$

The predator (y)–prey (x) equations, also known as the Volterra-Lotka equations, are written as, p , q , r , and s being constants,

$$\frac{dx}{dt} = px - qxy, \quad \frac{dy}{dt} = -ry + sxy. \quad (14)$$

No general analytical solution has been obtained for these equations except for special cases of coefficients. However, an implicit solution can be obtained by combining these equations in the phase plane as $dy/dx = (-ry + sxy)/(px - qxy)$, and the solution is

$$sx - r \log x = p \log y - qy + C; C - \text{a constant} \quad (15)$$

This is a closed curve in phase plane (x , y). C is determined by initial conditions for (x , y). In this way, the nature of solution can be obtained.

Lorenz equations, constituted of three ODEs, also occur frequently in the study of chaos in geosciences. These are, σ , ρ , and β being constants,

$$\frac{dx}{dt} = \sigma(y - x), \quad \frac{dy}{dt} = x(\rho - z) - y, \quad \frac{dz}{dt} = xy - \beta z. \quad (16)$$

No exact analytical solution to these equations has been found. However, the nature of solution can be deciphered

using qualitative methods initially developed by Poincare for studying the stability of the solar/planetary systems.

Solution Strategies

Two-Point Boundary Value Problem

These types of problems are posed as, with y -a vector

$$\frac{dy}{dt} = f(y, t), \quad a < t < b. \quad (17)$$

The boundary conditions at both ends are given either values of y or $\frac{dy}{dt}$, written as

$$B(y(a), y(b)) = 0. \quad (18)$$

The process of solving boundary value problem is to reduce them to initial value problem. In shooting method, we assume missing initial value and solve the problem as initial value problem, $y(a) = x$. The solution is substituted in the boundary condition as

$$B(x, y(b; x)) = 0. \quad (19)$$

This is a nonlinear function, and its root can be found. Simple shooting which is done from the boundary can be extended to multiple shooting by dividing the whole interval into subintervals and then shooting from one end of subinterval and using continuity of values at the nodes.

Eigenvalue problems are also boundary value problems. Here, the aim is to find values of coefficients in the differential equation such that both equation and boundary value are satisfied. The problem is posed, here, y – a vector

$$\frac{dy}{dt} = \lambda f(y), \quad a < t < b; y(a) = 0 = y(b). \quad (20)$$

Functions satisfying boundary conditions are substituted in the equation to get an algebraic equation for eigenvalue. The roots of the equation are eigenvalues, and corresponding orthogonal functions are eigen functions (see ► [“Eigenvalues and Eigenvectors”](#)). In a simple case

$$\frac{d^2y}{dt^2} = -\lambda y; y(0) = 0 = y(b). \quad (21)$$

We have the solution of this equation as

$$y(t) = A \sin \sqrt{\lambda} t + B \cos \sqrt{\lambda} t, \quad (22)$$

where $y(0) = 0$ shows that constant $B = 0$. Thus, nontrivial solutions are possible if $\sin \sqrt{\lambda} b = 0$. The solution of this

equation gives eigenvalues of the $\lambda_n = (n\pi/b)^2$; $n = 1, 2, \dots$, and the solution of boundary value problem is $y_n = A \sin n\pi t/b$. After normalization, the eigen functions are $y_n = \sqrt{2/b} \sin n\pi t/b$.

Green's Function Approach

The Green's function is the solution of a differential equation when the inhomogeneous term is given by Dirac delta function. Thus, for the Green's function $G(t, t_0)$ of second-order differential equation, we have the defining equation as

$$\frac{d^2 G(t, t_0)}{dt^2} + p(t) \frac{dG(t, t_0)}{dt} + q(t)G(t, t_0) = \delta(t - t_0), \quad (23)$$

with the boundary conditions as

$$G(0, t_0) = 0 \text{ and } G(d, t_0) = 0. \quad (24)$$

By integrating the defining equation around $t = t_0$, the following two conditions should be satisfied

$$G(t, t_0)|_{t_0+} - G(t, t_0)|_{t_0-} = 0, \quad (25a)$$

$$G'(t, t_0)|_{t_0+} - G'(t, t_0)|_{t_0-} = 1. \quad (25b)$$

Thus, the expression for the Green's function is obtained from two solutions of the homogeneous equation, y_1 , satisfying boundary condition at $t = 0$, and y_2 , satisfying boundary condition at $t = d$, as

$$G(t, t_0) = \begin{cases} \frac{y_1(t)y_2(t_0)}{W(t_0)} & 0 < t < t_0 \\ \frac{y_2(t)y_1(t_0)}{W(t_0)} & t_0 < t < d \end{cases} \quad (26)$$

This method helps to solve inhomogeneous ODEs.

Transform Methods

Initial value problem of ODEs can be converted to solving an algebraic equation after taking Laplace transformation (see ► “Laplace Transform”). After obtaining solution of this equation, inverse Laplace takes solution in transform domain to original domain. Laplace and inverse Laplace transform are defined as

$$Y(p) = L \{y(t)\} = \int_0^\infty e^{-pt} y(t) dt, \quad (27)$$

$$y(t) = L^{-1}Y(p) = \frac{1}{2\pi i} \int_{\sigma-i\epsilon}^{\sigma+i\epsilon} e^{pt} Y(p) dp; \operatorname{Re}(p) = \sigma. \quad (28)$$

Applying this to a simple first-order ODE

$$\frac{dy}{dt} = -ay, y(t=0) = y_0, \quad (29)$$

the result is

$$pY - y_0 = -aY. \quad (30)$$

Solving for Y and taking inverse Laplace transformation give

$$y(t) = y_0 e^{-at}. \quad (31)$$

This approach can be used for any order of ordinary differential equations.

Linear Systems

With the advent of computers, such equations are written in matrix form as

$$\frac{d}{dt} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ b & a \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} 0 \\ f(t) \end{bmatrix}. \quad (32)$$

This is further written as

$$\frac{dz}{dt} = Mz + g. \quad (33)$$

This is an example of linear dynamical system. The solution is written by mimicking first-order equation as

$$z = Ce^{Mt} + \int e^{M(t-t')} f(t') dt \quad (34)$$

There are large numbers of ways to calculate the exponential of a matrix. Best way is to use the method of singular value decomposition of a matrix. This matrix approach can be generalized to N number of equations. Constants are determined using the initial/boundary conditions (for finite intervals in t).

Numerical Solution

The ordinary differential equations, such as y -scale (vector)

$$\frac{dy}{dt} = f(y, t), y(t_0) = y_0, \quad (35)$$

are approximated by, e.g., the Euler method as

$$\frac{dy}{dt} \approx (y(t + \Delta t) - y(t)) / \Delta t, \quad (36a)$$

$$y(t + \Delta t) = y(t) + \Delta t f(y(t), t) + O(\Delta t^2), \quad (36b)$$

or more generally, with higher order schemes, say the Runge-Kutta, the fourth order, as

$$k_1 = \Delta t f(y_n, t_n), \quad (37a)$$

$$k_2 = \Delta t f\left(y_n + \frac{k_1}{2}, t_n + \Delta t/2\right), \quad (37b)$$

$$k_3 = \Delta t f\left(y_n + \frac{k_2}{2}, t_n + \Delta t/2\right), \quad (37c)$$

$$k_4 = \Delta t f(y_n + k_3, t_n + \Delta t), \quad (37d)$$

$$y(t_n + \Delta t) = y(t_n) + (k_1 + 2k_2 + 2k_3 + k_4)/6 + O(\Delta t^5). \quad (37e)$$

Now, powerful computer codes are available to solve system of ODEs, both nonstiff and stiff, such as `odeint()` method from `scipy.integrate` python module.

Special Functions

In simple cases, the solutions are made of elementary functions such as exponential, logarithmic, trigonometric, or hyperbolic equations. But when the coefficients of the second-order equations are a function of time, i.e.,

$$\frac{d^2 y}{dt^2} + p(t) \frac{dy}{dt} + q(t)y = 0, \quad (38)$$

it is no longer possible to write solution in terms of elementary equations. There are a large number of special functions worked out in various applications. The solutions are written in series form as

$$y = \sum_{n=0}^{\infty} a_n t^n. \quad (39)$$

This is substituted in the equation, and by equating the coefficients of t^n , the values of a_n 's are determined. If coefficients are singular, the expansion is taken around those points under some regularity conditions. Some frequently occurring special function equations are as follows (Arfken et al. 2015):

Bessel's Equation: $x^2 y'' + x y' + (x^2 - p^2)y = 0$

Legendre's Equation: $(1 - x^2)y'' - 2xy' + p(p + 1)y = 0$

Hermite Equation: $y'' - 2xy' + py = 0$

Gauss Hypergeometric Equation: $x(1 - x)y'' + [c - (a + b + 1)x]y' - aby = 0$

Laguerre's Equations: $xy'' + (1 - x)y' + py = 0$

Chebyshev's Equation: $(1 - x^2)y'' - xy' + p^2 y = 0$

Airy's Equation: $y'' \pm p^2 xy = 0$

In the above, p , a , b , and c are constants. All these function values are tabulated in the literature. These functions appear in describing wave and diffusion phenomena in the earth. As we need to compute the special functions, now it is preferable to use numerical solution of ordinary differential equations.

Qualitative Methods

Sometimes, it is not possible to find the solution of nonlinear problems. In such cases, it is still possible to find the solution in a qualitative form. This method was developed in the beginning of this century when Poincare' studied the problem of planetary system stability. Instead of determining the trajectory of planetary system, he devised methods to know whether such a system is stable or unstable. For instance, we are given

$$\frac{dy}{dt} = f(y). \quad (40)$$

We can find the equilibrium points of this problem by solving $f(y_e) = 0$. We can also find the stability of the equilibrium point y_e by finding the Jacobian $\partial f / \partial y$ at $y = y_e$. The nature of eigenvalues gives the nature of stability/instability. Such studies have been done in studying earth systems (Kaper and Engler 2013).

This can be demonstrated by a simple case

$$\frac{dy}{dt} = \mu - y^2, \quad (41)$$

which has equilibrium points at $y_e = \pm\sqrt{\mu}$. The stability of the system is deciphered from the sign of the following expression evaluated at equilibrium points

$$\frac{df}{dy} = -2y \text{ at } y = y_e. \quad (42)$$

Thus, $y_e = \sqrt{\mu}(-\sqrt{\mu})$ is stable (unstable). This is called pitchfork bifurcation.

Example of ODEs in Geoscience

Geochronology

Estimation of the age of rocks is an important study area in geosciences to reconstruct the geological processes in time and evolutionary history of the earth. The problem of decay of radiogenic elements $N(t)$ with time, t , can be expressed in terms of a first-order ODE

$$\frac{dN(t)}{dt} = -\lambda N(t), \quad (43)$$

where λ is the decay constant. For prescribed initial condition $N|_{t=0} = N_0$, the solution is obtained as

$$N(t) = N_0 \exp(-\lambda t). \quad (44)$$

Since our interest is in finding age, the above solution gives

$$t = -\frac{1}{\lambda} \ln\left(\frac{N}{N_0}\right). \quad (45)$$

Here, we need the values of two unknown parameters λ and N_0 to get the age. λ can be obtained in terms of half-life ($t_{1/2}$) of the radioactive element, defined as the time when its initial concentration N_0 is reduced to half. From this, we get $\lambda = \ln 2/t_{1/2}$. However, N_0 is not known. To circumvent this unknown, concentration of the stable decay product, called the daughter product, is used which is related to the parent element as

$$D = N_0 - N = N(e^{\lambda t} - 1). \quad (46)$$

Thus, time is given in terms of the ratio, D/N , as

$$t = \frac{1}{\lambda} \ln\left(\frac{D}{N} + 1\right). \quad (47)$$

Many times, there is a decay series, daughter product decaying to its own daughter products. For the case of the daughter product decaying to its subproduct with the decay constant λ_1 , we get

$$\frac{dD}{dt} = \lambda N - \lambda_1 D = \lambda N_0 e^{-\lambda t} - \lambda_1 D. \quad (48)$$

The solution for D is

$$D = \frac{\lambda}{\lambda_1 - \lambda} N_0 (e^{-\lambda t} - e^{-\lambda_1 t}). \quad (49)$$

In case $\lambda \ll \lambda_1$, this equation reduces to the condition called secular equilibrium.

$$\lambda_1 D \approx \lambda N. \quad (50)$$

Energy Balance at the Earth's Surface

Earth surface receives short wave radiation from the Sun and emits long wavelength radiation. This is modified by the surface albedo, atmospheric greenhouse gases, and planetary heat capacity. Using black body radiation law, the energy balance equation at the earth's surface is written as (Kaper and Engler 2013)

$$C \frac{dT}{dt} = Q(1 - \alpha) - \sigma T^4. \quad (51)$$

The surface temperature, time, heat capacity, incident solar radiation, albedo, and Stefan-Boltzmann constant are denoted respectively by T , t , C , Q , α , and σ . Albedo is also a function of temperature. Thus, this nonlinear equation needs to be solved by numerical method. However, for constant albedo and Budykov approximation for outgoing radiation as linear function of T (in $^\circ\text{C}$) as $(A + BT)$, this is reduced to linear ODE given by

$$C \frac{dT}{dt} = Q(1 - \alpha) - (A + BT). \quad (52)$$

Such equation has been used to find changes in the earth when solar luminosity changes due to solar processes.

Evolution of Mantle Temperature

Earth's mantle cools via loss of initial heat and thermal convection. The ordinary differential equation for mantle average temperature is obtained by equating the rate of mantle temperature change with difference of heat due to exponentially decaying radiogenic heat $h_0 e^{-\lambda t}$, and cooling at the earth surface (q) (Korenaga 2016)

$$C \rho D \frac{dT}{dt} = \rho D h_0 e^{-\lambda t} - q(T). \quad (53)$$

Here C , D , and ρ are heat capacity, thickness, and density of mantle, respectively. In general, $q(T)$ is a nonlinear function; however, this is linearized as

$$q = q_0 + \frac{dq}{dT} (T - T_o). \quad (54)$$

Thus, we get a linear differential equation for the evolution of mantle temperature which can be easily solved.

Overland Flow on a Hillslope

First-order ODEs also describe the overland flow in one horizontal direction (x) on a hillslope (see ► “Earth Surface Processes”), as given by (Anderson and Anderson 2010)

$$\frac{dQ_x}{dx} = R - I. \quad (55)$$

Here Q_x , R , and I are water flux, rainfall, and infiltration, respectively. Using the following relation between water flux, velocity, U , and thickness of overland flow, h ,

$$Q_x = Uh, U = \frac{1}{f} \sqrt{ghS}, \quad (56)$$

where g and S are gravity acceleration and hillslope, respectively; the expression for the thickness of the overland flow with distance from the hill top is expressed as

$$h(x) = \left(\frac{f(R - I)}{gS} \right)^{2/3} x^{2/3}. \quad (57)$$

Crustal Geotherm

The defining equation for the steady-state temperature distribution with depth in the continental crust is obtained by using corresponding special case of the heat diffusion equation given by (Turcotte and Schubert, 2010)

$$K \frac{d^2 T}{dz^2} = -A(z). \quad (58)$$

Here, temperature, depth, thermal conductivity, and radiogenic heat source concentration are denoted by T , K , z , and $A(z)$, respectively. The boundary conditions given at the earth surface ($z = 0$) and at the base of the crust ($z = l$) are

$$T|_{z=0} = T_0 \text{ and } K \frac{dT}{dz} \Big|_{z=l} = q_l, \quad (59)$$

where T_0 is the surface temperature and q_l is the heat flux at the base of the crust. The distribution of the radiogenic heat sources is given by

$$A(z) = A_0 \exp\left(-\frac{z}{d}\right), \quad (60)$$

where A_0 and d are the concentration of radiogenic sources at the surface and the characteristic depth. The solution of the problem is

$$T = T_0 + \frac{q_l}{K} z - \frac{A_0 d^2}{K} (1 - e^{-z/d}). \quad (61)$$

This expression has been used extensively to derive crustal geotherms.

Diffusion of Magnetic Field in Crust

The magnetic field due to external electric currents diffuses inside the earth's crust and the skin depth of diffusion is determined in terms of electrical conductivity of the rocks. This distribution is governed by the Maxwell's equations. For the horizontal magnetic field component $b\hat{x}$ diffusing with depth, denoted by z positive downward, and having time variation as $e^{i\omega t}$, a special case of Maxwell's equations is obtained as

$$\frac{d^2 b}{dz^2} = -i\omega\mu_0\sigma b. \quad (62)$$

Here, electrical conductivity and permeability are denoted, respectively, by σ and μ_0 , and ω is the angular frequency of the signal. The part of the solution which decreases with depth is given by

$$b = b_0 e^{-(1+i)z/z_0}; z_0 = \sqrt{\frac{2}{\omega\mu_0\sigma}}, \quad (63)$$

where b_0 is the amplitude of the magnetic field at the earth's surface and z_0 is called the skin depth that represents the length scale of the diffusion.

Channel Flow in Orogens

In orogens, sometimes, deep rocks are exposed at the earth's surface, and to explain it channel flow is postulated. The equation for variation in horizontal velocity $u(z)$ with depth z in a viscous (μ – viscosity) channel of thickness d due to applied pressure gradient dp/dx is approximated from the Navier-Stokes equation as (Turcotte and Schubert 2014)

$$\mu \frac{d^2 u}{dz^2} = \frac{dp}{dx}. \quad (64)$$

With top velocity as u_0 and bottom velocity as 0, the solution of this equation is

$$u(z) = u_0 \left(1 - \frac{z}{d}\right) + \frac{1}{2\mu} \frac{dp}{dx} (z^2 - dz). \quad (65)$$

Such solutions have been used to explain exposure of deeper rock in the Himalayan orogen.

Hillslope Forms

Hillslopes have convex shapes, and their weathered products diffuse toward their base. In the presence of tectonic uplift (U), the steady state height of hillslope (denoted by $z(x)$, with diffusion coefficient as D) connected to a river at horizontal distance $x = L$ from its peak at $x = 0$, is expressed, along with boundary conditions, as (Anderson and Anderson 2010)

$$\frac{d}{dx} \left(D \frac{dz}{dx} \right) + U = 0, \quad \frac{dz}{dx}(x=0) = 0, \quad z(L) = 0. \quad (66)$$

The solution of the above equation is

$$z(x) = \frac{U}{2D} (L^2 - x^2). \quad (67)$$

How much sediment is being put into the river can be determined by deriving the value of the flux at the river associated with uplift.

Groundwater Mound

Steady state water table ($h(x)$, x – horizontal axis) in an aquifer (conductivity denoted by k) bounded by rivers on both sides ($h(x=0) = h_1$, $h(x=b) = h_2$) under recharge I is governed by second-order differential equation (see ► “Flow in Porous Media”) as (Anderson 2008)

$$\frac{d}{dx} \left(kh \frac{dh}{dx} \right) + I = 0. \quad (68)$$

The solution is easily written for constant recharge as

$$h = \left(h_0^2 + \frac{I}{k} (bx - x^2) + \frac{x}{b} (h_1^2 - h_0^2) \right)^{1/2}. \quad (69)$$

The expression can be used to calculate the amount of water exchange taking place between river and aquifer due to changes in recharge and river levels.

Carbon Cycle

The most important part of carbon cycle at shorter timescale is about how much carbon is being put into the ocean. This problem can be understood by making box model of the coupled ocean and atmosphere and using conservation of mass to find out stocks and fluxes from both boxes. In a simplest model, we have f for anthropogenic input to atmosphere, with fluxes from atmosphere (ocean) to ocean (atmosphere) $aC_A(bC_O)$

$$\frac{dC_A}{dt} = -aC_A + bC_O + f, \quad \frac{dC_O}{dt} = aC_A - bC_O. \quad (70)$$

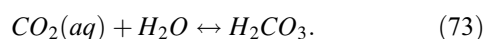
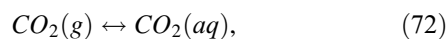
In steady state, we have $C_A/C_O = b/a$. This equation in matrix form can be written as

$$\frac{d}{dt} \begin{bmatrix} C_A \\ C_O \end{bmatrix} = \begin{bmatrix} -a & b \\ a & -b \end{bmatrix} \begin{bmatrix} C_A \\ C_O \end{bmatrix} + \begin{bmatrix} f(t) \\ 0 \end{bmatrix}. \quad (71)$$

This matrix equation can be solved by using linear system method.

Geochemical Reactions

Most common reactions, such as dissolution, hydrolysis, hydration, and oxidation, occur near the earth's surface due to its interaction with the atmosphere. On the one hand, these provide chemical elements for life processes, and on the other, these also lead to environmental hazard. In a simple case, dissolution of carbon dioxide in water is written as



Such reactions can be described by a system of ordinary differential equations which would describe the changes in the concentration of various elements. We can write reaction equation for dissolution of carbon dioxide in water as

$$\begin{aligned} dx/dt &= -k_1x + k_{-1}y, & dy/dt \\ &= k_1x - k_{-1}y - k_2y + k_{-2}p, \end{aligned} \quad (74a)$$

$$dz/dt = -k_2y + k_{-2}p, \quad dp/dt = k_2y - k_{-2}p, \quad (74b)$$

where x , y , z , and p are concentration of $CO_2(g)$, $CO_2(aq)$, H_2O , and H_2CO_3 , respectively. Reaction constants for forward and backward reactions are denoted by k 's. These equations can be written in matrix form and solved by using matrix methods.

Geophysical Systems

Geophysical problems are represented by a set of partial differential equations. These problems are reduced to a set of ODEs after discretizing the spatial domain. Method of lines is a well-developed tool to take this approach. We consider a simple advection-diffusion PDE which occurs quite frequently in geoscience

$$\frac{\partial u}{\partial t} = a \frac{\partial u}{\partial x} + b \frac{\partial^2 u}{\partial x^2}. \quad (75)$$

The spatial derivatives are discretized for the domain ($1 < i < M$), which gives a set of ODEs as

$$\frac{du_i}{dt} = a \frac{u_{i+1} - u_i}{\Delta x} + b \frac{u_{i+1} - 2u_i + u_{i-1}}{(\Delta x)^2}; \quad 1 < i < M. \quad (76)$$

With requisite boundary and initial conditions, a solution can be obtained analytically or numerically. Similar approach is used to discretize space part of second-order diffusion and wave equations.

Summary

Ordinary differential equations (ODEs) pervade geosciences. When studying long-term behavior of earth, box models are used which are described by a set of ODEs. ODEs also occur when using system's thinking about earth. Also, when focus is on steady state and one space-dimensional problem, again ODEs are used. Further, numerical solutions of PDEs often are obtained by decomposing them into a set of coupled ODEs. Both initial and boundary value problems for ODEs have been developed along with eigenvalue/eigenvectors and

Green's function method. For variable coefficients, some solutions are written in terms of special functions, or recourse to numerical methods is taken using powerful computer programs like highly popular Python's `odeint()` method. This entry has described the ODEs as applicable to the earth problems.

Cross-References

- [Earth Surface Processes](#)
- [Eigenvalues and Eigenvectors](#)
- [Flow in Porous Media](#)
- [Laplace Transform](#)
- [Partial Differential Equations](#)

Acknowledgments AM carried out this work under the project MLP6404-28 (AM) with CSIR-NGRI contribution number NGRI/Lib/2021/Pub-03.

Bibliography

- Anderson RS (2008) The little book of geomorphology: exercising the principle of conservation. Electronic textbook. <http://instaar.colorado.edu/~andersrs/publications.html#littlebook>, p 133
- Anderson RS, Anderson SP (2010) Geomorphology: the mechanics and chemistry of landscapes. Cambridge University Press, Cambridge, UK
- Arfken GB, Weber HJ, Harris FE (2015) Mathematical methods for physicists. Academic Press, San Diego
- Coddington EA (1989) An introduction to ordinary differential equations. Dover Publications, New York
- Kaper H, Engler H (2013) Mathematics and climate. SIAM, Philadelphia
- Korenaga J (2016) Can mantle convection be self-regulated? *Sci Adv* 2(8):e1601168. <https://doi.org/10.1126/sciadv.1601168>
- Turcotte D, Schubert G (2014) Geodynamics, 3rd edn. Cambridge University Press, Cambridge, UK

Ordinary Least Squares

Christopher Kotsakis
Department of Geodesy and Surveying, School of Rural and Surveying Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece

Definition

Ordinary least squares. A mathematical framework for analyzing erroneous observations in order to extract information about physical systems based on simple modeling hypotheses.

Introduction

An important aspect of the geosciences is to make inferences about physical systems from imperfect data. For this purpose, it is often required to work with mathematical models linking a set of unknown parameters with other observable quantities through known equations. Least squares theory originated from the aforesaid necessity and provides a standard framework to obtain information about the physical world on the basis of inferences drawn from modeled observations in the presence of additive errors.

Initially motivated by estimation problems in astronomy and geodesy (Nievergelt 2000), least squares theory was developed independently by Gauss and Legendre around 1795–1820, about 40 years after the advent of robust L_1 estimation. Its historical tracing has been studied by several scientists (e.g., Seal 1967; Plackett 1972; Sheynin 1979; Stigler 1991), and it constitutes one of the most interesting disputes in the history of science. The remarkable monograph by Gauss (1823) already contains almost the complete modern theory of least squares including elements of the theory of probability distributions, the definition and properties of the Gaussian distribution, and a discussion of the Gauss-Markov theorem on the statistical properties of the least squares estimator. In this entry, we give a brief overview of *ordinary least squares (OLS)* which is the simplest manifestation of least squares theory for analyzing noisy data with deterministic parametric models.

General Setting of the Problem

Consider a known vector $\mathbf{l} = [l_1, l_2, \dots, l_n]^T$ of n observations. From physical, mathematical or empirical considerations, it is expected that these observations can be explained by a smaller set of m parameters $\mathbf{x} = [x_1, x_2, \dots, x_m]^T$ through a known predictive model

$$\mathbf{f} : \begin{cases} \mathbb{R}^m \rightarrow \mathbb{R}^n & (n > m) \\ \mathbf{x} \mapsto \mathbf{f}(\mathbf{x}) \end{cases} \quad (1)$$

which is represented by a nonlinear over-determined system of n equations with m unknowns

$$\mathbf{l} = \mathbf{f}(\mathbf{x}) \quad (2)$$

The fundamental problem to be discussed herein is the inference of the model parameters from the information contained in the observations through the inversion of the above system (Menke 2015). In most applications of physical sciences and engineering, such a system does not have an exact unique solution, that is, no parameter vector $\mathbf{x} \in \mathbb{R}^m$ exists that can reproduce the given data $\mathbf{l} \in \mathbb{R}^n$. This setback

originates from the unavoidable presence of errors in the redundant observations (and perhaps from hidden deficiencies in the model choice itself). Hence, to a certain degree of simplification, the above problem may be viewed as a quest for a unique solution among the infinite ones that satisfy the under-determined system of observation equations

$$\mathbf{l} = \mathbf{f}(\mathbf{x}) + \mathbf{v} \quad (3)$$

where the residual vector $\mathbf{v}$ contains the unknown observation noise and other possible errors due to imperfect data modeling. In this context, *OLS is an optimal method for solving the above system in order to obtain a best estimate of the model parameters based on algebraic or statistical principles.*

OLS: Algebraic Aspects

From an algebraic perspective, the rationale of OLS is to provide an estimator for the model parameters, denoted hereafter by $\hat{\mathbf{x}}^{OLS}$, that minimizes the Euclidean (squared) distance between the observations and their model-based prediction (Menke 1989; Rao and Toutenburg 1999)

$$\begin{aligned} \hat{\mathbf{x}}^{OLS} &= \arg \min_{\mathbf{x} \in \mathbb{R}^m} \left(\|\mathbf{l} - \mathbf{f}(\mathbf{x})\|^2 \right) \\ &= \arg \min_{\mathbf{x} \in \mathbb{R}^m} \left((\mathbf{l} - \mathbf{f}(\mathbf{x}))^T (\mathbf{l} - \mathbf{f}(\mathbf{x})) \right) \end{aligned} \quad (4)$$

The above solution does not rely on any stochastic or statistical considerations about the observation errors. It is merely a best-fitting solution in the sense of minimizing the residual vector in Eq. (3) or more precisely the sum of the squared values of its elements (hence the term “ordinary least squares”).

Model Linearization

To explicitly compute $\hat{\mathbf{x}}^{OLS}$ in practice, one first has to linearize the predictive model around known approximate values for the parameters $\mathbf{x}_o$ and then apply the least squares criterion to its linearized version in order to solve for the first-order corrections $\delta\mathbf{x} = \mathbf{x} - \mathbf{x}_o$. Of course, this step needs to be properly iterated (i.e., the solution of each step is used as initial approximation in the next step) until sufficient convergence is reached in the estimated parameters. As an alternative, instead of applying the OLS principle to the linearized model, one may directly solve the nonlinear optimization problem of Eq. (4) by using some well-known iterative technique (e.g., Gauss-Newton method, steepest descent gradient method). We prefer to follow the first approach here, since it is the one that is commonly used in least squares problems with nonlinear parametric models. For more details, see, e.g., Sen and Srivastava (1990, pp. 298–316).

In the following, it is assumed that the linearization of the predictive model $\mathbf{l} = \mathbf{f}(\mathbf{x})$ was applied beforehand. Moreover, for the sake of notation simplicity, we shall retain the same symbols in both sides of the linearized model which is expressed hereafter as $\mathbf{l} = \mathbf{A}\mathbf{x}$, instead of the lengthened form $\mathbf{l} - \mathbf{f}(\mathbf{x}_o) = \mathbf{A}(\mathbf{x} - \mathbf{x}_o)$. The Jacobian matrix $\mathbf{A} = \partial\mathbf{f}/\partial\mathbf{x}$, known as the *design matrix*, is computed with the help of approximate parameter values, and it needs to be updated during the iterative implementation of OLS estimation in nonlinear models.

Estimation of Model Parameters

Based on the previous linear(ized) framework, the objective function to be minimized in the optimal criterion of Eq. (4) takes the form

$$\begin{aligned} \|\mathbf{l} - \mathbf{f}(\mathbf{x})\|^2 &= (\mathbf{l} - \mathbf{A}\mathbf{x})^T (\mathbf{l} - \mathbf{A}\mathbf{x}) \\ &= \mathbf{l}^T \mathbf{l} - 2\mathbf{l}^T \mathbf{A}\mathbf{x} + \mathbf{x}^T \mathbf{A}^T \mathbf{A}\mathbf{x} \end{aligned} \quad (5)$$

A necessary condition to minimize the last expression is the nullification of its partial derivative

$$\frac{\partial \|\mathbf{l} - \mathbf{f}(\mathbf{x})\|^2}{\partial \mathbf{x}} = -2\mathbf{l}^T \mathbf{A} + 2\mathbf{x}^T \mathbf{A}^T \mathbf{A} = \mathbf{0} \quad (6)$$

which leads to the so-called normal equations

$$(\mathbf{A}^T \mathbf{A}) \mathbf{x} = \mathbf{A}^T \mathbf{l} \Leftrightarrow \mathbf{N} \mathbf{x} = \mathbf{b} \quad (7)$$

where $\mathbf{N} = \mathbf{A}^T \mathbf{A}$ is called the *normal matrix*. If the design matrix $\mathbf{A}$ has full rank, then the normal matrix is invertible and the normal equations admit a unique solution which is the OLS estimator of the model parameters (Koch 1999; Rao and Toutenburg 1999)

$$\hat{\mathbf{x}}^{OLS} = \mathbf{N}^{-1} \mathbf{b} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{l} \quad (8)$$

Model-Based Prediction

In practice, the user’s interest may be confined on $\hat{\mathbf{x}}^{OLS}$ or it may extend to its predictive ability for other quantities that depend on the model parameters – such quantities are called *response variables* in the terminology of regression analysis (Draper and Smith 1998; Sen and Srivastava 1990). In fact, the observations that were used to estimate the model parameters can be predicted anew as follows:

$$\hat{\mathbf{l}}^{OLS} = \mathbf{A} \hat{\mathbf{x}}^{OLS} = \mathbf{A} (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{l} = \mathbf{H} \mathbf{l} \quad (9)$$

where $\mathbf{H} = \mathbf{A} (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$ is an important matrix, called the *hat matrix*, which executes the orthogonal projection from the observation space $\mathbb{R}^n$ onto the column space of the design

matrix. The latter is denoted as $\text{Im}\{\mathbf{A}\}$, and it corresponds to a particular subspace of $\mathbb{R}^n$ spanning all possible predictions of the observables via the adopted model, that is, $\text{Im}\{\mathbf{A}\} = \{\mathbf{A}\mathbf{x}, \mathbf{x} \in \mathbb{R}^m\}$. The hat matrix is symmetric and idempotent while its trace is equal to the number of model parameters ($\text{tr} \mathbf{H} = m$). The predicted vector in Eq. (9) upgrades the original observations in the sense of reducing (not eliminating) their random errors to get new values that are compatible with the parametric model – this replacement of $\mathbf{l}$ by $\hat{\mathbf{l}}^{OLS}$ is generally termed *least squares adjustment* (Teunissen 2000).

Residuals

The OLS residuals reflect the difference between the original observations and their model-based predictions, and they are given by the equation

$$\begin{aligned} \hat{\mathbf{v}}^{OLS} &= \mathbf{l} - \mathbf{A} \hat{\mathbf{x}}^{OLS} = \mathbf{l} - \mathbf{A} (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{l} \\ &= (\mathbf{I} - \mathbf{H}) \mathbf{l} \end{aligned} \quad (10)$$

where $\mathbf{I}$ is the $n \times n$ unit matrix and $\mathbf{I} - \mathbf{H}$ corresponds to the orthogonal projector from the observation space $\mathbb{R}^n$ onto the orthogonal complement of $\text{Im}\{\mathbf{A}\}$. This matrix is also symmetric and idempotent, while its trace is equal to the *degrees of freedom* of the estimation problem at hand ($\text{tr}(\mathbf{I} - \mathbf{H}) = n - m$). The OLS residuals give a useful metric for the effectiveness of the solution, but they cannot be exploited in more comprehensive quality analyses or statistical testing procedures without the aid of auxiliary hypotheses for the observation errors.

Geometrical Interpretation of OLS

The OLS method provides a best estimate of the model parameters by splitting the observations into two orthogonal components: a model-based predicted part and a residual part

which is attributed to measurement and/or modeling errors. This decomposition is algebraically expressed as

$$\mathbf{l} = (\mathbf{I} - \mathbf{H}) \mathbf{l} + \mathbf{H} \mathbf{l} = \hat{\mathbf{v}}^{OLS} + \hat{\mathbf{l}}^{OLS} \quad (11)$$

whereas the orthogonality condition $(\hat{\mathbf{l}}^{OLS})^T (\hat{\mathbf{v}}^{OLS}) = 0$ is easily verified from the properties of the hat matrix. The above splitting allows a simple geometrical interpretation of OLS which is depicted in Fig. 1.

OLS: Statistical Aspects

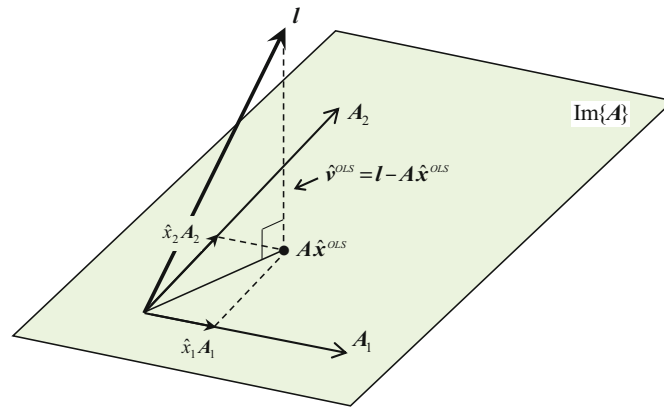
The algebraic formulation of the OLS method does not account for the nature of the residuals in the augmented system of observation equations. Its justification is based on the presumption that their values should be small (hence the reasoning of $\|\mathbf{v}\| \rightarrow \min$), a fact that implies the availability of a “good” dataset and the usage of a “correct” parametric model. The switch to a statistical formulation allows us to formalize those two assumptions while it offers the necessary framework to test their validity in practice and to assess the quality of the OLS solution.

Gauss-Markov Conditions

In the statistical context of OLS, the residuals in Eq. (3) are modeled as zero-mean random variables with common variance and zero covariances (Rao and Toutenburg 1999):

$$E\{\mathbf{v}\} = \mathbf{0}, \mathbf{D}\{\mathbf{v}\} = E\{\mathbf{v}\mathbf{v}^T\} = \sigma^2 \mathbf{I} \quad (12)$$

where $E\{\cdot\}$ is the expectation operator from probability theory, $\mathbf{D}\{\mathbf{v}\}$ is the covariance matrix (also called dispersion matrix) of the residual vector, and σ^2 is the common variance



Ordinary Least Squares, Fig. 1 The OLS solution is an orthogonal projection from the “observation space” $\mathbb{R}^n$ onto the “model space” $\text{Im}\{\mathbf{A}\}$. The columns of the design matrix form a (non-orthogonal) basis in the model space while the estimated parameters are the “coordinates” of

the model-predicted observables with respect to that basis. For easier visualization, it is assumed that the model contains only two parameters which correspond to columns $\mathbf{A}_1$ and $\mathbf{A}_2$ of the design matrix

of its elements. The above assumptions may also be expressed in the observation domain:

$$\begin{aligned} E\{I\} &= A x, \\ D\{I\} &= E\{(I - A x)(I - A x)^T\} = \sigma^2 I \end{aligned} \quad (13)$$

and they are known as *Gauss-Markov (G-M) conditions* which theoretically assure the statistical goodness of the OLS solution (Koch 1999). In simple words, these conditions dictate that the observations are free of any systematic errors, outliers, or blunders, and they are influenced only by (uncorrelated) random errors at the same accuracy level. Moreover, in the absence of such errors, the observations are expected to conform to a linear model which is specified through a full-rank design matrix of error-free elements (A) and a parameter vector with fixed but unknown values (x) that need to be estimated from a given ensemble of noisy observations.

Statistical Optimality of OLS Solution

If the G-M conditions hold true, then the OLS solution admits optimal properties that justify its use in real-life applications. Specifically, it becomes a linear unbiased estimator of the model parameters, that is

$$\begin{aligned} E\{\hat{x}^{OLS}\} &= E\{(A^T A)^{-1} A^T I\} \\ &= (A^T A)^{-1} A^T E\{I\} \\ &= (A^T A)^{-1} A^T A x \\ &= x \end{aligned} \quad (14)$$

whereas its covariance (dispersion) matrix is

$$\begin{aligned} D\{\hat{x}^{OLS}\} &= D\{(A^T A)^{-1} A^T I\} \\ &= (A^T A)^{-1} A^T D\{I\} (A^T A)^{-1} \\ &= \sigma^2 (A^T A)^{-1} A^T A (A^T A)^{-1} \\ &= \sigma^2 (A^T A)^{-1} \end{aligned} \quad (15)$$

and it has the smallest trace among any other linear unbiased estimator of x from the same data I (Rao and Toutenburg 1999; Koch 1999). The square roots of its diagonal elements reflect the (internal) accuracy of the estimated model parameters – also called standard errors – and they are used to describe the statistical goodness of the OLS solution at the parameter level. For example, the standard error of $\hat{x}_i$ denoted by $\sigma_{\hat{x}_i}$ is the square root of the i th diagonal element of $\sigma^2 (A^T A)^{-1}$, and it gives a measure of its likely offset from the true value of the respective parameter. For a more realistic assessment of the estimation accuracy, the standard errors are

usually amplified by a factor of 3 or 4 in order to increase their confidence level.

In a nutshell, the OLS solution minimizes the propagated observation errors to the estimated model parameters. The estimator $\hat{x}^{OLS}$ is the *best linear unbiased estimator* which offers optimal accuracy (minimum variance) in the presence of the G-M conditions. What is more, the statistical optimality extends to all model-based predictions that will use the OLS estimator to infer other quantities in a linear(-ized) context, that is, if $y = q^T x$, then its estimate $\hat{y}^{OLS} = q^T \hat{x}^{OLS}$ is unbiased and it has minimum variance among any other linear unbiased estimator of y from the same data. For more details, see Koch (1999) and the references given therein.

If the G-M conditions are invalid in practice, then the OLS solution does not necessarily become useless, yet its results lose their statistical optimality and may be affected by hidden biases. In such cases, the user needs to make appropriate changes to the data and/or the parametric model which would cause approximate compliance with the G-M conditions.

Other Statistical Properties

Based on Eqs. (9, 10), and after considering the properties of the hat matrix, one easily obtains the covariance matrix of the model-predicted observations

$$\begin{aligned} D\{\hat{l}^{OLS}\} &= D\{H I\} \\ &= H D\{I\} H^T \\ &= \sigma^2 H \end{aligned} \quad (16)$$

and also the covariance matrix of the OLS residuals

$$\begin{aligned} D\{\hat{v}^{OLS}\} &= D\{(I - H) I\} \\ &= (I - H) D\{I\} (I - H)^T \\ &= \sigma^2 (I - H) \end{aligned} \quad (17)$$

In contrast to $D\{I\}$, the covariance matrices $D\{\hat{v}^{OLS}\}$ and $D\{\hat{l}^{OLS}\}$ are always singular, and their diagonal elements have a key role in the statistical testing of observations regarding the presence of outliers. The last two equations imply the decomposition

$$D\{I\} = D\{\hat{v}^{OLS}\} + D\{\hat{l}^{OLS}\} \quad (18)$$

which is a manifestation of the orthogonality property of OLS in the stochastic domain. The model-predicted observations and the estimated residuals are actually uncorrelated with each other (Draper and Smith 1998)

$$E\left\{\hat{\mathbf{v}}^{OLS} \left(\hat{\mathbf{l}}^{OLS}\right)^T\right\} = \mathbf{0} \quad (19)$$

and, therefore, the sum of their covariance matrices will be equal to the covariance matrix of the observations. Note that the expected values of these quantities shall comply with the G-M conditions, that is, $E\left\{\hat{\mathbf{l}}^{OLS}\right\} = \mathbf{A} \mathbf{x}$ and $E\left\{\hat{\mathbf{v}}^{OLS}\right\} = \mathbf{0}$.

OLS and Maximum Likelihood Estimation

The OLS estimator does not require the probability distribution function of the observation errors. In fact, the only necessary stochastic elements that need to be specified through the G-M conditions are the first- and second-order moments of the observation vector $\mathbf{l}$. However, in case of normally distributed errors, the OLS estimator becomes equivalent with the maximum likelihood estimator which is inherently related to *Bayesian inference* under the hypothesis of uniform prior distribution for the model parameters (Rao and Toutenburg 1999).

Goodness of Fit

The evaluation of model fit is an important task that often needs to be performed in the context of OLS estimation. In general, a successful solution is characterized by small residuals which indicate a good fit of the observations to the parametric model (or vice versa). A number of relevant measures to quantify the goodness of fit of the solution are briefly described below. For a more detailed discussion, see the respective chapters in Draper and Smith (1998) and Sen and Srivastava (1990).

Residual Sum of Squares (RSS)

It corresponds to the (squared) Euclidean length of the OLS residual vector. Its value may be computed by three equivalent expressions as follows:

$$RSS = \left(\hat{\mathbf{v}}^{OLS}\right)^T \hat{\mathbf{v}}^{OLS} = \mathbf{l}^T (\mathbf{I} - \mathbf{H}) \mathbf{l} = \mathbf{l}^T \hat{\mathbf{v}}^{OLS} \quad (20)$$

The disadvantage of this quantity is its dependence on the units in which the observations are expressed, thus making it difficult to exploit the RSS for objective assessment of model fit performance. If the above value is divided by the degrees of freedom of the problem at hand, then we obtain an unbiased estimate of the variance of the (true) residuals in the respective model

$$s^2 = \frac{RSS}{n - m}, \quad E\{s^2\} = \sigma^2 \quad (21)$$

The above estimate (often called *residual mean square – RMS*) is useful if the data noise level is unknown, as it allows us to replace σ^2 with s^2 in order to infer the statistical accuracy of the results, e.g.

$$\mathbf{D}\left\{\hat{\mathbf{x}}^{OLS}\right\} = \sigma^2 (\mathbf{A}^T \mathbf{A})^{-1} \rightarrow \hat{\mathbf{D}}\left\{\hat{\mathbf{x}}^{OLS}\right\} = s^2 (\mathbf{A}^T \mathbf{A})^{-1} \quad (22)$$

The knowledge of noise level is not required in the numerical computation of the estimated parameters, but it is necessary for computing their associated covariance matrix. Obviously, for s^2 to be a meaningful estimate of the noise variance, it needs to be ensured that no “problematic” observations (in violation of the G-M conditions) were used in the OLS solution.

Coefficient of Multiple Determination

This is a unitless measure which is commonly used to evaluate the model fit in OLS problems (yet it requires caution when comparing different models with the same or several datasets). It is defined as

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^n \left(\hat{l}_i^{OLS} - \bar{l}\right)^2}{\sum_{i=1}^n \left(l_i - \bar{l}\right)^2} = \frac{\left(\hat{\mathbf{l}}^{OLS} - \bar{\mathbf{l}}\right)^T \left(\hat{\mathbf{l}}^{OLS} - \bar{\mathbf{l}}\right)}{\left(\mathbf{l} - \bar{\mathbf{l}}\right)^T \left(\mathbf{l} - \bar{\mathbf{l}}\right)} \\ &= 1 - \frac{RSS}{\left(\mathbf{l} - \bar{\mathbf{l}}\right)^T \left(\mathbf{l} - \bar{\mathbf{l}}\right)} \end{aligned} \quad (23)$$

where $\bar{l}$ denotes the mean value of the observations and $\bar{\mathbf{l}}$ is an auxiliary vector whose elements are all equal to $\bar{l}$. The value of R^2 ranges between 0 and 1, and it reflects the squared correlation coefficient between the observations and their model-predicted part. A perfect fit to the available data would give $R^2 = 1$ (since $RSS = 0$) which is a highly unlikely event in practice. In general, the closer its value gets to 1, the better is the model fit to the noisy observations.

The previous definition of R^2 relies on the assumption that the parametric model includes a constant term, which means that the design matrix $\mathbf{A}$ should contain a column full of ones (a constant term in a linear model is usually called the “intercept” in statistical literature). As a matter of fact, an alternative interpretation of R^2 is that it measures the usefulness of the extra terms other than the intercept in the parametric model (Draper and Smith 1998). Its value is not comparable between entirely different models, but it can be used to compare the fitting performance of nested models. As a result, parametric models which contain an intercept cannot be compared with those without intercept on the basis of R^2 .

Sometimes it is preferable to use an *adjusted* R^2 , denoted by R_a^2 , which is given by the formula (Sen and Srivastava 1990)

$$R_a^2 = 1 - \frac{\sum_{i=1}^n (\hat{l}_i^{OLS} - \bar{l})^2 / (n - m)}{\sum_{i=1}^n (l_i - \bar{l})^2 / (n - 1)} = 1 - (1 - R^2) \left(\frac{n - 1}{n - m} \right) \quad (24)$$

The above coefficient accounts for the sample size (i.e., number of available observations) since it is often felt that small sample sizes tend to inflate the value of R^2 . This revised metric can be used to compare different parametric models that may be fitted not only to a specific dataset but also to two, or more, different sets of observations. Note that R_a^2 may take negative values, and it practically serves as a gross indicator for the model fit in OLS problems.

Residual Analysis

The observations to be analyzed by least squares methods may be affected by outliers or blunders due to improper conditions in data collection or other external disturbances. Depending on the magnitude of such errors and the importance of the affected observables, the OLS solution may be seriously harmed even by a single bad observation. The analysis of residuals aims to identify bad observations which should be excluded from the estimation procedure as they violate the first of the G-M conditions in Eq. (13). One of the most important tips to remember here is the following: the OLS residuals do not correspond to the true observation errors but they are related to them through the orthogonal projection

$$\hat{\mathbf{v}}^{OLS} = (\mathbf{I} - \mathbf{H}) \mathbf{v} \quad (25)$$

Although the above equation does not have practical value, it shows that the OLS residuals always give a “blurred” snapshot of the true observation errors. In particular, a gross error in a single observation is partly “transferred” to the residuals of other observations, thus causing difficulties to isolate problematic observations solely on the basis of $\hat{\mathbf{v}}^{OLS}$.

Of special importance in residual analysis are the diagonal elements $\{H_{ii}\}$ of the hat matrix $\mathbf{H}$, which are linked with the variances of the model-predicted observations as follows

$$\text{var}\{\hat{l}_i^{OLS}\} = \sigma^2 H_{ii} \quad (26)$$

and they represent the so-called *leverages* of the respective observables (Belsley et al. 1980). Their values are always positive numbers between 0 and 1, and their importance should be mainly considered in a relative context: if an

observation has much higher leverage than the others, then it is poorly predicted by the model parameters, and its prediction is likely to absorb the largest part of a hidden outlier or blunder. The sum of the leverages is equal to $\text{tr } \mathbf{H} = m$ which means that their average value is $\bar{H}_{ii} = m/n$. Therefore, an alarm for very influential observations which may be outliers could be set if $\bar{H}_{ii} > 2m/n$ or maybe $\bar{H}_{ii} > 3m/n$ (Draper and Smith 1998). High-leverage observations usually originate from poor experimental design, and they pose a threat to the integrity of OLS solutions.

For the purpose of detecting bad observations that do not comply with the G-M conditions, more valuable than the ordinary residuals $\hat{\mathbf{v}}^{OLS}$ are the so-called standardized or (internally) studentized residuals given as

$$e_i = \frac{\hat{v}_i^{OLS}}{\left(\text{var}\{\hat{v}_i^{OLS}\}\right)^{1/2}} = \frac{\hat{v}_i^{OLS}}{s \sqrt{1 - H_{ii}}} \quad (27)$$

whose variance is always equal to 1. A modified version of the above residuals

$$e'_i = \frac{e_i \sqrt{n - m - 1}}{\sqrt{n - m - e_i^2}} \quad (28)$$

known as (*externally*) *studentized residuals* is mostly used in practice (Draper and Smith 1998). These residuals follow the t -distribution with $n - m - 1$ degrees of freedom under the normality assumption for the observation errors, and their values are used by most software packages as test statistics in single outlier detection.

Summary

Least squares theory carries a long history dating back to the late eighteenth century, and it offers a powerful framework for analyzing noisy data in a broad range of applications in physical sciences and engineering. The estimation problems that can be treated by the least squares methodology involve a variety of modeling setups with assorted complexity levels, which extend well beyond the simple OLS case that was covered in this entry. Most of these problems are of high relevance to various branches of mathematical geosciences, including the inversion of rank-deficient linear models with or without the presence of prior information, the least squares approximation of continuous signals from discrete observations, the handling of correlated data noise with regular or singular covariance matrix, the recursive estimation of time-varying parameters in dynamic systems (Kalman filtering), and the parameter estimation in the so-called errors-in-

variables models. On the other hand, the OLS methodology and its associated diagnostic tools remain an attractive option for handling parameter estimation and model-fitting problems with homogeneous datasets, having as major advantage the straightforward and easy-to-implement mathematical framework that was outlined in this entry.

Cross-References

- [Best Linear Unbiased Estimation](#)
- [Iterative Weighted Least Squares](#)
- [Least Median of Squares](#)
- [Least Squares](#)
- [Maximum Likelihood](#)
- [Multiple Correlation Coefficient](#)
- [Multivariate Analysis](#)
- [Normal Distribution](#)
- [Optimization in Geosciences](#)
- [Random Variable](#)
- [Rao, C. R.](#)
- [Regression](#)
- [Statistical Outliers](#)
- [Statistical Quality Control](#)

Bibliography

- Belsley DA, Kuh E, Welsch RE (1980) Regression diagnostics. Wiley
- Draper NR, Smith H (1998) Applied regression analysis. Wiley
- Gauss CF (1823) *Theoria combinationis observationum erroribus minimis obnoxia: pars prior, pars posterior*. *Commentat Soc Regiae Sci Gott Recent* 5:1823
- Koch K-R (1999) Parameter estimation and hypothesis testing in linear models. Springer, Berlin/Heidelberg
- Menke W (1989) Geophysical data analysis: discrete inverse theory. Academic, San Diego
- Menke W (2015) Review of the generalized least squares method. *Surv Geophys* 36:1–25
- Nievergelt Y (2000) A tutorial history of least squares with applications to astronomy and geodesy. *J Comput Appl Math* 121:37–72
- Plackett RL (1972) The discovery of the method of least squares. *Biometrika* 59(2):239–251
- Rao CR, Toutenburg H (1999) Linear models, least squares and alternatives. Springer, New York
- Seal HL (1967) The historical development of the Gauss linear model. *Biometrika* 54(1–2):1–24
- Sen A, Srivastava M (1990) Regression analysis: theory, methods and applications. Springer, New York
- Sheynin OB (1979) C.F. Gauss and the theory of errors. *Arch Hist Exact Sci* 20:21–72
- Stigler SM (1991) Gauss and the invention of least squares. *Ann Stat* 9: 465–474
- Teunissen PJG (2000) Adjustment theory, an introduction. Delft Academic Press

Partial Differential Equations

R. N. Singh¹ and Ajay Manglik²

¹Discipline of Earth Sciences, Indian Institute of Technology, Gandhinagar, Palaj, Gandhinagar, India

²CSIR-National Geophysical Research Institute, Hyderabad, India

Definition

Partial differential equation: A differential equation that contains unknown functions of two or more independent variables and their partial derivatives

Introduction

Geosciences use physical laws to derive information about the properties and processes in Earth based on observations of physical fields. Physical laws are those which govern mechanical, thermal, electromagnetic, and chemical behaviors of continuous media. These consist of conservation laws and constitutive relations. This leads to laws as a set of partial differential equations (PDEs) (Officer 1974; Kennett and Bunge 2008). PDEs in general form are written as a function of space, time, function, and its derivatives, i.e.,

$$f(x, y, z, t, \phi, \phi_x, \phi_y, \phi_z, \phi_{xx}, \phi_{yy}, \phi_{zz}) = 0, \text{ where } \phi_x \equiv \frac{\partial \phi}{\partial x}. \quad (1)$$

Most of the physicochemical processes in Earth are governed by second-order PDEs. Three classic examples of linear homogeneous PDEs, elliptic, parabolic, and hyperbolic types, respectively, are (Morse and Feshbach 1953):

1. Potential field equation governing static gravity, magnetic, and electrostatic fields, written as $\nabla^2 \phi = 0$, where the

Laplacian is written as $\nabla^2 \equiv \phi_{xx} + \phi_{yy} + \phi_{zz}$ in the Cartesian coordinate system.

2. Diffusion equation governing transport of heat, chemical concentration, low-frequency electromagnetism, stress diffusion in earthquake generation, groundwater flow, geomorphology, and many more, written as $\nabla^2 \phi = \partial \phi / \partial t$.
3. Wave phenomenon governing propagation of acoustic, elastic, and water waves due to sudden disturbances in the Earth system, written as $\nabla^2 \phi = \partial^2 \phi / \partial t^2$.

For time harmonic waves or fields, diffusion and wave equations are reduced to the Helmholtz equation $\nabla^2 \phi = -k^2 \phi$. PDEs describing mantle and core convection involved in quantifying plate tectonic processes and origin of the magnetic field are also reduced to the above elliptic, parabolic, and hyperbolic equations under various approximations of flow conditions likely to be met in Earth. For various problems in geoscience, these equations are written in the Cartesian, cylindrical, or spherical coordinates.

The above equations are supplemented by boundary and initial conditions to solve problems related to a vast area of geosciences. There are three kinds of boundary conditions, namely, the Dirichlet, the Neumann, and the Robin boundary conditions. These are:

- Dirichlet condition: ϕ is prescribed on the boundary.
- Neumann condition: $\nabla \phi$ is prescribed on the boundary.
- Robin condition: $a\phi + b\nabla \phi$ is prescribed on the boundary.

Mathematical investigations of the PDEs reveal that not all solutions of these equations may be useful for interpreting observations. Hadamard (1902) proposed a set of criteria which a well-posed problem should fulfill. These are as follows: (i) the solution should exist, (ii) the solution should be unique, and (iii) the solution should depend on the given data continuously, i.e., it should be stable. The equations should have proper initial and boundary conditions. Like for heat equation, we should have one initial condition along with

boundary conditions. And for second order wave equation we should have two initial conditions along with boundary conditions. Those problems which are not well posed in the sense of Hadamard criteria are called ill-posed problems. In some cases, solutions may not exist, for some it may not be unique, and for some others it may be unstable with respect to changes in given data. There are regularization methods to make such problems useful for interpreting data.

The basic strategy to solve PDEs is to split the problem into a series of initial value problems in ordinary differential equations (ODEs) (► [Ordinary Differential Equations](#)) or system of linear algebraic equations. For simple geometries in linear problems, the solution can be obtained by using well-known method of separation of variables, method of Green's function, and integral transform methods. For nonlinear and highly heterogenous problems, numerical methods such as finite difference or finite elements are used which involve ultimately solving a set of algebraic equations. These methods are briefly outlined below.

Method of Separation of Variables

This method is discussed here for the potential field equation in two-dimensional Cartesian coordinate system to describe temperature distribution, $T(x, z)$, in the crust. The domain extends in the horizontal (x) and vertical (z) directions as $-l \leq x \leq l$ and $0 \leq z \leq d$, respectively, with z positive downward. The PDE is written as

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial z^2} = 0, \quad (2)$$

with the boundary conditions as

$$T(x, 0) = 0, T(x, d) = f(x), \quad (3a)$$

$$\frac{\partial T}{\partial y}(-l, z) = 0, \frac{\partial T}{\partial y}(l, z) = 0. \quad (3b)$$

Assuming that the variables are separable, the solution can be written as $T(x, y) = X(x)Z(z)$, which transforms Eq. 2 to a set of two ODEs:

$$\frac{1}{X} \frac{d^2 X}{dx^2} = -\frac{1}{Z} \frac{d^2 Z}{dz^2} = -\lambda^2, \quad (4)$$

with respective boundary conditions as

$$Z(0) = 0 = Z(d), \frac{dX}{dx}(-l) = 0 = \frac{dX}{dx}(l). \quad (5)$$

The solutions to this eigenvalue problem for X are orthogonal functions $X_n(x) = \cos(\lambda_n x)$, $\lambda_n = n\pi/l$, and the solution for Z satisfying boundary condition at $z = 0$ is $Z = \sinh(n\pi z/l)$. Therefore, the solution of T can be expressed as

$$T(x, y) = \sum_{n=0}^{\infty} c_n \cos(n\pi x/l) \sinh(n\pi z/l). \quad (6)$$

Applying boundary condition at $z = d$ yields

$$f(x) = \sum_{n=0}^{\infty} c_n \cos(n\pi x/l) \sinh(n\pi d/l). \quad (7)$$

Using the orthogonality property of $\sin(n\pi/l)$, we get expression for c_n as

$$c_n = \frac{2}{\sinh(n\pi d/l)} \int_0^d f(x) \cos(n\pi x/l). \quad (8)$$

Thus, for the given function $f(x)$, the solution can be obtained by evaluating this integral.

This method can also be applied to time-dependent heat conduction problem (κ – diffusivity) applied to cooling of a hot geological body, posed as

$$\frac{\partial T}{\partial t} - \kappa \frac{\partial^2 T}{\partial z^2} = 0, 0 \leq z < d, 0 \leq t < \infty. \quad (9)$$

The initial and boundary conditions are

$$T(z, 0) = f(z), \quad (10a)$$

$$u(0, t) = 0 = u(d, t). \quad (10b)$$

Using the method of separation of variables, we get two ordinary differential equations as

$$\frac{dT(t)}{dt} = -\kappa \lambda^2 T(t), \frac{d^2 Z(z)}{dz^2} = -\lambda^2 Z(z). \quad (11)$$

The differential equation for Z is posed as an eigenvalue problem with the solution $Z_n = \sin(\lambda_n z)$, where $\lambda_n = (n\pi/d)$. Thus, the solution is

$$u(z, t) = \sum_{n=1}^{\infty} C_n \sin\left(\frac{n\pi z}{d}\right) e^{-\lambda_n^2 \kappa t}. \quad (12)$$

The constants C_n are obtained by using the initial condition, which gives

$$u(z, 0) = f(z) = \sum_{n=1}^{\infty} C_n \sin\left(\frac{n\pi z}{d}\right). \quad (13)$$

Using the orthogonality of eigenfunctions, we get expression for constants C_n as

$$C_n = 2 \int_0^d f(x) \sin(n\pi z/d) dz. \quad (14)$$

Thus, the solution is given by

$$u(z, t) = 2 \sum_{n=1}^{\infty} \sin\left(\frac{n\pi z}{d}\right) \int_0^d f(z') \sin(n\pi z'/d) dz' e^{-\lambda_n^2 \kappa t}. \quad (15)$$

This method is also useful for solving wave equation problem posed as (here, u denotes displacement)

$$\frac{\partial^2 u}{\partial t^2} = \frac{1}{c^2} \frac{\partial^2 u}{\partial z^2}, 0 < t < \infty, 0 \leq z \leq d, \quad (16a)$$

$$u(0, t) = 0 = u(d, t), \quad (16b)$$

$$u(x, 0) = f(x), \frac{\partial u}{\partial t}(x, 0) = g(x). \quad (16c)$$

Using the method of separation of variables, $u = T(t)Z(z)$, two ordinary differential equations are

$$\frac{d^2 T(t)}{dt^2} = -\lambda^2 c^2 T(t), \frac{d^2 Z(z)}{dz^2} = -\lambda^2 Z(z), Z(0) = 0 = Z(d). \quad (17)$$

The differential equation for Z poses an eigenvalue problem with the solution $Z_n = \sin(\lambda_n z)$, where $\lambda_n = (n\pi/d)$. The solution of the PDE (Eq. 16a) is

$$u(z, t) = \sum_{n=1}^{\infty} \sin\left(\frac{n\pi z}{d}\right) \left(C_n \sin\left(\frac{n\pi c t}{d}\right) + D_n \cos\left(\frac{n\pi c t}{d}\right) \right). \quad (18)$$

Constants C_n, D_n are found by using the orthogonality property of the eigenfunctions, i.e.,

$$C_n = \frac{1}{d} \int_0^d f(z) \sin\left(\frac{n\pi z}{d}\right), \quad (19)$$

$$D_n = \frac{1}{n\pi c} \int_0^d g(z) \sin\left(\frac{n\pi z}{d}\right). \quad (20)$$

For given values of $f(z), g(z)$, we can get the values of these constants.

We have presented solutions in the Cartesian geometry. However, many problems in seismology and geomagnetism need spherical model of Earth. In such cases, the governing equation is written in the spherical coordinate system. Separated ODEs made use of special functions like spherical harmonics for latitude and longitude dependencies and Bessel functions to represent radial variations.

Method of Eigenvalue: Eigenfunction Expansion

This method can be used to solve more general heat/chemical/stress diffusion problems with inhomogeneous source and boundary conditions, such as

$$\frac{\partial T}{\partial t} = \kappa \frac{\partial^2 T}{\partial z^2} + g(z, t), 0 < t < \infty, 0 \leq z \leq d, \quad (21a)$$

$$T(0, t) = a(t), T(d, t) = b(t), T(z, 0) = f(z). \quad (21b)$$

The solution is written in two parts such that the boundary conditions become homogeneous, i.e.,

$$T(z, t) = u(z, t) + v(z, t). \quad (22)$$

The function $v(z, t)$ is chosen as $v(z, t) = a(t) + z(b(t) - a(t))/d$ so that the equation for $u(z, t)$ becomes

$$\frac{\partial u}{\partial t} = \kappa \frac{\partial^2 u}{\partial z^2} + g'(x, t), 0 < t < \infty, 0 \leq z \leq d, \quad (23)$$

where $g'(z, t) = g(z, t) - \partial v/\partial t$. The boundary and initial conditions of $u(z, t)$ become $u(0, t) = 0 = u(d, t)$ and $u(z, 0) = f(x) - v(z, 0)$, respectively.

The solution for $u(z, t)$ is thus obtained by using the following eigenvalues and eigenfunctions (► [Eigenvalues and Eigenvectors](#)):

$$\lambda_n = \left(\frac{n\pi}{d}\right), u_n = \sin(\lambda_n z). \quad (24)$$

The method described for wave equation in previous section can be followed to get the solution:

$$T(z, t) = v(z, t) + \sum_{i=1}^n \left(\int_0^d e^{\kappa \lambda_n^2 (t-\tau)} g'_n(\tau) d\tau + e^{\kappa \lambda_n^2 t} T_n(0) \right) \sin \lambda_n z. \quad (25)$$

Here, $T_n(0)$ is given by

$$T_n(0) = \frac{2}{d} \int_0^d (f(x) - v(z, 0)) \sin(\lambda_n z) dz. \quad (26)$$

The above formulation used the Dirichlet boundary condition. Similar approach can be used to find solution for the Neumann and the Robin boundary conditions.

The above formulation can also help in solving following more general advection-diffusion equation which occurs in several transport processes like heat, groundwater flow (► [Flow in Porous Media](#)), and dispersion of chemical substance in the underground:

$$\frac{\partial T}{\partial t} + v \frac{\partial T}{\partial z} = \kappa \frac{\partial^2 T}{\partial z^2}, \quad (27a)$$

$$T(0, t) = 0, T(d, t) = 0, T(x, 0) = f(z). \quad (27b)$$

We can use the following transformation to the variable $T(z, t)$

$$T(z, t) = T'(z, t) e^{ax+bt}, \quad (28)$$

to get

$$\frac{\partial T'}{\partial t} + (b + va - a^2\kappa) \kappa T' + (v - 2a\kappa) \frac{\partial T'}{\partial z} = \kappa \frac{\partial^2 T'}{\partial z^2}. \quad (29)$$

We assume $b + va - a^2\kappa = 0$ and $v - 2a\kappa = 0$, which gives $a = v/2\kappa$ and $b = -v^2/4\kappa$. With this, advection diffusion and initial/boundary conditions reduce to

$$\frac{\partial T'}{\partial t} = \kappa \frac{\partial^2 T'}{\partial z^2}, \quad (30a)$$

$$T'(0, t) = 0, T'(d, t) = 0, T(x, 0) = f(z) e^{-vx/2\kappa}. \quad (30b)$$

This can then be solved by using eigenvalue-eigenvector expansion method discussed above.

Method of Integral Transforms

Integral transforms reduce PDEs to ODEs. The choice of integral transforms depends on the nature of boundary condition. A popular method for solving time-dependent problems is the Laplace transform method (► [Laplace Transform](#)). Laplace transform method, in respect of time $0 < t < \infty$, is defined as

$$T(z, p) = L \{T(z, t)\} = \int_0^\infty e^{-pt} T(z, t) dt, \quad (31)$$

with the inverse Laplace transform given by

$$T(z, t) = L^{-1} T(z, p) = \frac{1}{2\pi i} \int_{\sigma - i\epsilon}^{\sigma + i\epsilon} e^{pt} T(z, p) dp; \operatorname{Re}(p) = \sigma. \quad (32)$$

For a large number of functions, Laplace transform pairs have been obtained over the years and are well tabulated. We illustrate this method by estimating the temperature in half-space for prescribed initial condition, posed as

$$\frac{\partial T}{\partial t} - \kappa \frac{\partial^2 T}{\partial z^2} = 0, 0 < z < \infty, 0 < t < \infty, \quad (33a)$$

$$T(z, 0) = A, T(0, t) = 0. \quad (33b)$$

Taking Laplace transform of Eq. 33a, we get

$$pT(z, p) - A = \kappa \frac{d^2 T(z, p)}{dz^2}, \quad (34)$$

the solution of which is

$$T(z, p) = \left(C e^{z\sqrt{\frac{p}{\kappa}}} + D e^{-z\sqrt{\frac{p}{\kappa}}} \right) + \frac{A}{p}. \quad (35)$$

As the solution needs to be finite as z tends to infinity, the solution takes the form as

$$T(z, p) = \left(D e^{-z\sqrt{\frac{p}{\kappa}}} \right) + \frac{A}{p}. \quad (36)$$

The value of D is found by using surface boundary condition as $D = -A/p$. Thus, the solution in transform domain is

$$T(z, p) = \frac{A}{p} \left(1 - e^{-z\sqrt{\frac{p}{\kappa}}} \right). \quad (37)$$

The solution in time domain is obtained by performing inverse Laplace transformation. The solution is

$$T(z, t) = A \left(1 - \operatorname{erfc} \left(\frac{z}{\sqrt{4\kappa t}} \right) \right). \quad (38)$$

Similarly, Laplace transform can be used to solve first-order PDE occurring in geomorphology, h , representing elevation in river long profile under uplift denoted by b (► [Earth-Surface Processes](#)):

$$a \frac{\partial h}{\partial x} + \frac{\partial h}{\partial t} = b, 0 < t < \infty, 0 < x < \infty, \quad (39a)$$

$$h(x, 0) = 0, h(0, t) = 0. \quad (39b)$$

Taking Laplace transform of this equation, we get

$$a \frac{dh(x, p)}{dx} + ph(x, p) = b/p. \quad (40)$$

The solution of the equation is

$$h = \frac{b}{ap} + Ce^{-xp/a}. \quad (41)$$

Using the boundary condition, we get $C = -b/(ap)$, and the solution in transform domain is

$$h(x, p) = \frac{b}{ap} (1 - e^{-xp/a}) = \frac{b}{a} \left(\frac{1}{p} - \frac{e^{-xp/a}}{p} \right). \quad (42)$$

Taking inverse Laplace transformation, we get

$$h(x, t) = \frac{b}{a} (1 - H(t - x/a)). \quad (43)$$

There is a large class of integral transforms and their inverse, and these have been tabulated. Some of these transforms are (Carslaw and Jaeger 1959):

$$\text{Fourier } (-\infty < x < \infty) : F(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-ikx} dx$$

$$\text{Inverse Fourier} : f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(k) e^{-ikx} dk$$

$$\text{Fourier sine } (0 < x < \infty) : F(k) = \frac{2}{\pi} \int_0^{\infty} f(x) \sin(kx) dx;$$

Dirichlet boundary condition

$$\text{Inverse Fourier sine} : f(k) = \int_0^{\infty} F(k) \sin(kx) dx$$

$$\text{Fourier cosine } (0 < x < \infty) : F(k) = \frac{2}{\pi} \int_0^{\infty} f(x) \cos(kx) dx;$$

Neumann boundary condition

$$\text{Inverse Fourier cosine} : f(k) = \int_0^{\infty} F(k) \sin(kx) dx$$

$$\text{Finite sine } (0 < x < L) : F_n = \frac{2}{\pi} \int_0^L f(x) \sin(n\pi x/L) dx;$$

Dirichlet boundary condition

$$\text{Inverse finite sine } (0 < x < L) : f(x) = \sum_{n=1}^{\infty} F_n \sin(n\pi x/L)$$

$$\text{Finite cosine } (0 < x < L) : F_n = \frac{2}{\pi} \int_0^L f(x) \cos(n\pi x/L) dx;$$

Neumann boundary condition

$$\text{Inverse finite cosine } (0 < x < L) : f(x)$$

$$= \frac{F_0}{2} + \sum_{n=1}^{\infty} F_n \cos(n\pi x/L)$$

These transforms are helpful in solving PDEs with applicable boundary conditions.

Green's Function Approach

Green's function arises in geoscience problems when attention is focused on generation of fields by sources. For example, in case of earthquakes, the generated displace field $\mathbf{u}$, expressed in terms of potential ϕ as $\mathbf{u} = \nabla \phi$, due to seismic sources $f(r, t)$, can be obtained by solving (Aki and Richards 2002)

$$\nabla^2 \phi - \frac{1}{\alpha^2} \frac{\partial^2 \phi}{\partial t^2} = -f(r, t). \quad (44)$$

Green's function, $G(r, t, r_0, t_0)$, corresponding to the above problem is given by the solution of

$$\nabla^2 G - \frac{1}{\alpha^2} \frac{\partial^2 G}{\partial t^2} = -\delta(r - r_0) \delta(t - t_0), \quad (45)$$

representing point and instantaneous source. Taking Fourier transform (► [Fourier Transform](#)) of this equation, we get

$$(\nabla^2 + k^2) \bar{G} = \delta(r - r_0) e^{i\omega t_0}. \quad (46)$$

The solution of this equation, which decays to zero as $r \rightarrow \infty$, is

$$\bar{G} = -\frac{e^{-ik|r-r_0|}}{4\pi|r-r_0|} e^{i\omega t_0}. \quad (47)$$

Taking inverse Fourier transform, we get

$$G = \frac{-\delta(t - t_0 - \frac{r-r_0}{c})}{4\pi|r-r_0|}. \quad (48)$$

Knowing Green's function solution, the solution of Eq. 44 is obtained by integrating Green's function over volume, i.e.,

$$\phi(r, t) = \int G(r, r_0, t, t_0) f(r_0, t_0) dr_0 dt_0. \quad (49)$$

Thus, the displacement field can be constructed for any given source.

Method of Characteristics for Wave Equations

The method of characteristics also reduces a PDE into a set of ODEs. This is demonstrated by the following equation describing erosion due to stream power in geomorphology (► [Earth-Surface Processes](#)) in the presence of tectonic uplift:

$$a \frac{\partial h}{\partial x} + b \frac{\partial h}{\partial t} = c, h(0) = f(x). \quad (50)$$

This PDE is equivalent to the following three ODEs, called characteristic and compatibility equations:

$$\frac{dx}{ds} = a, \frac{dt}{ds} = b, \frac{dh}{ds} = c. \quad (51)$$

The initial conditions for these equations at $s = 0$ are, η being parameter along x -axis,

$$x = x_0(\eta), t = t_0(\eta), h = h_0(\eta). \quad (52)$$

We also have $t = 0, x = \eta, h = f(\eta)$. Thus, the solutions of characteristic curves are, a, b , and c being constants,

$$x = as + \eta, t = bs, h = cs + f(\eta). \quad (53)$$

Transforming in the original domain, we get solution as

$$h(x, t) = ct/b + f(x - at/b). \quad (54)$$

This can be extended to initial and boundary value problems.

Ray Method in Wave Propagation

Seismological knowledge is built on using the solution of wave equation in interpreting earthquake data. The most commonly used solution is ray solution of wave equation (Aki and Richards 2002). For time harmonic waves ($\phi = \phi' e^{-i\omega t}$), the wave equation reduces to

$$\nabla^2 \phi' = \frac{(-i\omega)^2}{\alpha^2} \phi'. \quad (55)$$

High-frequency waves propagate along rays. In this, the solution is sought in the following form:

$$\phi' = A(x) e^{-i\omega B(x)}. \quad (56)$$

Substituting in the wave equation, we get

$$\nabla^2 A - w^2 A |\nabla B|^2 - i(2w \nabla A \cdot \nabla B + w A \nabla^2 B) = -\frac{w^2}{\alpha^2} A. \quad (57)$$

By separating real and imaginary parts, we get the following two equations:

$$\nabla^2 A - w^2 A |\nabla B|^2 + \frac{w^2}{\alpha^2} A = 0, \quad (58a)$$

$$2w \nabla A \cdot \nabla B + w A \nabla^2 B = 0. \quad (59b)$$

The real part is rewritten as

$$\frac{1}{Aw^2} \nabla^2 A - |\nabla B|^2 + \frac{1}{\alpha^2} = 0. \quad (60)$$

In high-frequency situation, the first term can be ignored to get the eikonal or ray equation:

$$|\nabla B|^2 = \frac{1}{\alpha^2} = n^2. \quad (61)$$

The reciprocal of velocity, denoted by n , is called the slowness parameter. Wavefront is given by $B = \text{constant}$. Ray path is given by ∇B . This is a first-order hyperbolic equation and can be solved by using the method of characteristics. A simple solution is

$$A = 1, B = nx. \quad (62)$$

Thus, we get plane wave solution as

$$\phi = e^{-i\omega(nx+t)}. \quad (63)$$

The imaginary part gives transport equation as

$$2\nabla A \cdot \nabla B + A \nabla^2 B = 0. \quad (64)$$

This equation is used to find the amplitude of the propagating waves.

Similarity Solutions

It is possible to transform PDE into ODE by symmetry transformations. Following dilation, symmetry transformation gives the so-called similarity solution:

$$x' = \frac{x}{L}, t' = t/L^\beta. \quad (65)$$

In case of heat diffusion problem

$$\frac{\partial T}{\partial t} - \kappa \frac{\partial^2 T}{\partial z^2} = 0, 0 < z < \infty, 0 < t < \infty, \quad (66a)$$

$$T(z, 0) = T_m, T(0, t) = T_0, \quad (66b)$$

a similarity parameter $\eta = z/2\sqrt{\kappa t}$ is defined such that the diffusion equation reduces to

$$\frac{d^2 T}{d\eta^2} + 2\eta \frac{dT}{d\eta} = 0. \quad (67)$$

To set up boundary condition, temperature is also made nondimensional by using $T' = (T - T_m)/(T_m - T_0)$ such that

$$T'(0) = 1, T'(\infty) = 0. \quad (68)$$

The solution is obtained in terms of the error function as

$$T' = 1 - \text{erf}(\eta), \text{ where } \text{erf}(\eta) = \frac{2}{\sqrt{\pi}} \int_0^\eta e^{-y^2} dy. \quad (69)$$

This solution is used in quantifying the heat flow, bathymetry, seismicity, and many other aspects of structure and evolution of the oceanic lithosphere (Turcotte and Schubert 2002).

Numerical Methods

Very few PDEs representing the real Earth situations have analytical solutions for the given geometry and distribution of properties. In these cases, recourse to numerical methods is taken (► [Computational Geosciences](#)). Numerical methods aim at approximating PDEs to a system of linear algebraic equations which can be solved iteratively. There are two main methods for solving PDEs numerically, finite difference and finite element. In finite difference method, the derivatives occurring in the equations are approximated by difference forms, with values of the variables at discrete nodes. This results in a set of algebraic equations. In the finite element method, whole geometry is divided into finite elements; such elements could be finite interval in 1D or triangle in 2D or tetrahedron in 3D. Within the elements, continuous variation of the variables is assumed in terms of their values on the nodes of the elements. PDE is written in variational form over the domain, and the equations in terms of nodal values are derived by integrating the variation

equation over the elements. This again results in a set of algebraic equations.

Finite Difference Method

Semi-Discrete Method

We illustrate this method on the following heat diffusion problem:

$$\frac{\partial T}{\partial t} - \kappa \frac{\partial^2 T}{\partial z^2} = g(z), 0 < z < d, 0 < t < \infty, \quad (70a)$$

$$T(z, 0) = f(z), T(0, t) = 0 = T(d, t). \quad (70b)$$

We discretize z -axis as $z_i = i\Delta z$, $i = 0, 1, 2, \dots, n + 1$; $\Delta z = d/(n + 1)$. On this grid, we have the following discrete form of the second space derived in the equation

$$\frac{\partial^2 T}{\partial z^2} \approx \frac{T(z_{i+1}, t) - 2T(z_i, t) + T(z_{i-1}, t))}{(\Delta z)^2}. \quad (71)$$

Thus, the governing equation is approximated as

$$\frac{dT_i}{dt} \approx \kappa \frac{T_{i+1} - 2T_i + T_{i-1}}{(\Delta z)^2} + g(z_i), i = 1, 2, 3 \dots n. \quad (72a)$$

$$T_i = T(z_i, t), T_0 = 0 = T_{n+1}, T_i(0) = f(z_i). \quad (72b)$$

These equations can be solved as a system of first-order ODEs.

Semi-Discrete Collocation Method

In this method, the space part of solution is written in terms of some basis functions (► [Radial Basis Functions](#)), $\phi_j(z) : j = 1, \dots, n$ as

$$T(z, t) \approx \sum_{j=1}^n a_j(t) \phi_j(z). \quad (73)$$

We substitute this expression in the governing equation and get

$$\begin{aligned} \sum_{j=1}^n a'_j(t) \phi_j(z_i) &= \kappa \sum_{i=1}^n a_j(t) \phi_j''(z_i) \\ &+ \sum_{i=1}^n g(z_i) \phi_j(z_i). \end{aligned} \quad (74)$$

This can be written in matrix form as

$$a' = \kappa M^{-1} N a + M^{-1} b. \quad (75)$$

Here $a = [a_1, a_2, \dots, a_n]^T$; $m_{ij} = \phi_j(z_i)$; $n_{ij} = \phi_j''(z_i)$; $b_{ij} = g(z_i)\phi_j(z_i)$. This is a system of ODEs which can be solved by following ODE solution methods.

Fully Discrete Methods

Here, both space and time domain are discretized. The time is discretized as $t_k = \Delta t$, $k = 1, 2, \dots$. Fully discrete form of the governing equation is

$$\frac{T(t_{k+1}, z_k) - T(t_k, z_k)}{\Delta t} \approx \kappa \frac{T(z_{l+1}, t_k) - 2T(z_i, t_k) + T(z_{l-1}, t_k)}{(\Delta z)^2} + g(z_i), \quad (76a)$$

$$\begin{aligned} T(t_{k+1}, z_k) = & T(t_k, z_k) \\ & + \frac{\kappa \Delta t}{(\Delta z)^2} (T(z_{l+1}, t_k) - 2T(z_i, t_k) + T(z_{l-1}, t_k)) \\ & + g(z_i) \Delta t. \end{aligned} \quad (76b)$$

The initial and boundary conditions are

$$T(z_i, 0) = f(z_i), T(0, t_k) = 0 = T(z_{n+1}, t_k). \quad (77)$$

This problem can be solved iteratively. This is called explicit method. The space and time steps are chosen so that the solution is stable. This requires meeting the Courant-Friedrichs-Lewy (CFL) criteria:

$$\frac{\kappa \Delta t}{(\Delta z)^2} \leq \frac{1}{2}. \quad (78)$$

Finite Element Method

The most useful method in FEM uses weak formulation of the problem. Weak solutions are applicable to geophysical situations for which classical solutions having smoothness and other regulatory condition are not applicable. In this case, the PDE is reduced to set of ODEs. Here, governing equation is multiplied by a test function and integrated over the whole domain. Thus, for one-dimensional heat diffusion equation, we get

$$\int_0^d v(z) \left(\frac{\partial T}{\partial t} - \kappa \frac{\partial^2 T}{\partial z^2} \right) dz = \int_0^d v(z) g(z) dz. \quad (79)$$

The second term on the right-hand side is integrated by parts, and using the boundary condition on the test function, we get

$$\int_0^d v(z) \frac{\partial T}{\partial t} dz + \kappa \int_0^d \frac{\partial v}{\partial z} \frac{\partial T}{\partial z} dz = \int_0^d v(z) g(z) dz. \quad (80)$$

To proceed further to get a set of ODEs, the solution is written in terms of orthogonal basis functions as

$$T(z, t) \approx \sum_{j=1}^n a_j(t) \phi_j(z). \quad (81)$$

The test function is also taken as

$$v(z) = \phi_i(z). \quad (82)$$

Substituting both in the integral expressions, we get

$$\begin{aligned} & \int_0^d \phi_i(z) \frac{\partial}{\partial t} \left(\sum_{j=1}^n a_j(t) \phi_j(z) \right) dz \\ & + \kappa \int_0^d \frac{\partial \phi_i(z)}{\partial z} \frac{\partial}{\partial z} \left(\sum_{j=1}^n a_j(t) \phi_j(z) \right) dz \\ & = \int_0^d \phi_i(z) g(z) dz. \end{aligned} \quad (83)$$

We denote

$$M_{ij} = \int_0^d \phi_i(z) \phi_j(z) dz, \quad (84a)$$

$$N_{ij} = \int_0^d \frac{\partial \phi_i(z)}{\partial z} \frac{\partial \phi_j(z)}{\partial z} dz, \quad (84b)$$

$$g_i = \int_0^d \phi_i(z) g(z) dz. \quad (84c)$$

We then get

$$M \frac{da}{dt} + Na = g. \quad (85)$$

The initial condition is obtained by using the orthogonality properties of ϕ_j

$$T(z, 0) = f(z) = \sum_{j=1}^n a_j(0) \phi_j(z). \quad (86)$$

These ODEs are solved to get the solution of the PDE.

Deep Learning Method

PDEs also are used in deep learning algorithms. Space-time observations are fitted to PDEs with unknown coefficients which are then determined by using an optimization process based on neural network ([► Neural Networks](#)). This then helps to know the nature of physical processes involved in

generating the observations. Deep learning methods (► [Machine Learning and Geosciences](#)) are also used to find the solution of PDEs by approximating the functions by neural networks. In one such method, called deep Galerkin method (Raissi 2018), the solution which in the FEM is approximated by certain basis functions is instead approximated by neural network. This method helps to solve faster when grids and basis functions become large in multi-dimensional geometries.

Summary

Most processes in geosciences vary with space and time. These processes are governed by conservation laws and constitutive relationships for various materials occurring in Earth. Combining these, we get laws of continuum physics written as partial differential equations. These equations are in general second-order partial differential equations. General strategy in solving these equations is to convert them into a set of ordinary differential equations by using the method of separation of variable, transformation methods, or numerical methods. We have presented these developments as applied to the study of processes related to the Earth.

Cross-References

- [Computational Geoscience](#)
- [Earth Surface Processes](#)
- [Eigenvalues and Eigenvectors](#)
- [Flow in Porous Media](#)
- [Fast Fourier Transform](#)
- [Laplace Transform](#)
- [Machine Learning](#)
- [Neural Networks](#)
- [Ordinary Differential Equations](#)
- [Radial Basis Functions](#)

Acknowledgments AM carried out this work under the project MLP6404-28(AM) with CSIR-NGRI contribution number NGRI/Lib/2021/Pub-02.

Bibliography

- Aki K, Richards PG (2002) Quantitative seismology, 2nd edn. University Science Books
- Anderson RS, Anderson SP (2010) Geomorphology: the mechanics and chemistry of landscapes. Cambridge University Press
- Carslaw HS, Jaeger JA (1959) Conduction of heat in solids, 2nd edn. Oxford University Press
- Hadamard J (1902) Sur les problèmes aux dérivées partielles et leur signification physique. Princeton University Bulletin, pp 49–52

- Kennett BLN, Bunge H-P (2008) Geophysical continua – deformation in the Earth's interior. Cambridge University Press
- Morse P, Feshbach F (1953) Methods of theoretical physics, vol 1. McGraw-Hill
- Officer CB (1974) Introduction to theoretical geophysics. Springer
- Raissi M (2018) Deep hidden physics models: deep learning of nonlinear partial differential equations. J Mach Learn Res 19:1–24
- Turcotte DL, Schubert B (2002) Geodynamics, 2nd edn. Cambridge University Press

Particle Swarm Optimization in Geosciences

Joseph Awange¹, Béla Paláncz² and Lajos Völgyesi²

¹School of Earth and Planetary Sciences, Discipline of Spatial Sciences, Curtin University, Perth, WA, Australia

²Department of Geodesy and Surveying, Budapest University of Technology and Economics, Budapest, Hungary

Definition

One of the nature-inspired meta-heuristic global optimization methods, the particle swarm optimization (PSO), is here introduced and its application in geosciences presented. The basic algorithm with an illustrative example is discussed and compared with other global methods like simulated annealing, differential evolution, and Nelder-Mead method.

Global Optimization

In global optimization problems, where an objective function $f(x)$ having many local minima is considered, finding the global minimizer (also known as global solution or global optimum) x^* and the corresponding f^* value is of concern.

Global optimization problems arise frequently in engineering, decision-making, optimal control, etc. (Awange et al. 2018). There exist two huge but almost completely disjoint communities (i.e., they have different journals, different conferences, different test functions, etc.) solving these problems: (i) a broad community of practitioners using stochastic nature-inspired meta-heuristics and (ii) academicians studying deterministic mathematical programming methods.

To solve global optimization problems, where evaluation of the objective function is an expensive operation, one needs to construct an algorithm that is able to stop after a fixed number of evaluations M of $f(x)$, from which the lowest obtained value of $f(x)$ is used. For global optimization, it is not the dimension of the problem that is important in local optimization but rather the number of allowed function evaluations (often called budget). In other words, when one has the possibility to evaluate $f(x)$

M times (these evaluations are hereinafter called trials), in the global optimization problem of the dimension 5, 10, or 100, the quality of the found solution after M evaluations is crucial and not the dimensionality of $f(x)$. This happens because it is not possible to adequately explore the multidimensional search region D at this limited budget of expensive evaluations of $f(x)$. For instance, if $D \subset R_{20}$ is a hypercube, then it has 220 vertices. This means that one million of trials are not sufficient not only to explore well the whole region D but even to evaluate $f(x)$ at all vertices of D . Thus, the global optimization problem is frequently *NP-hard*.

Meta-heuristic algorithms widely used to solve real-life global optimization problems have a number of attractive properties that have ensured their success among engineers and practitioners. First, they have limpid nature-inspired interpretations explaining how these algorithms simulate behavior of populations of individuals. Other reasons that have led to a wide spread of meta-heuristics are the following: they do not require high level of mathematical preparation to understand them, their implementations are usually simple, and many codes are freely available. Finally, they do not need a lot of memory as they work at each moment with only a limited population of points in the search domain. On the flip side, meta-heuristics have some drawbacks, which include usually a high number of parameters to tune and the absence of rigorously proven global convergence conditions ensuring that sequences of trial points generated by these methods always converge to the global solution x^* . In fact, populations used by these methods can degenerate prematurely, returning only a locally optimal solution instead of a global one or even nonlocally optimal point if it has been obtained at one of the last evaluations of $f(x)$ and the budget of M evaluations has not allowed it to proceed with an improvement of the obtained solution.

Nature-Inspired Global Optimization

Meta-heuristic algorithms are based on natural phenomenon such as particle swarm optimization simulating fish schools or bird groups; firefly algorithm simulating the flashing behavior of the fireflies; artificial bee or ant colony representing a colony of bees or ants in searching the food sources; differential evolution and genetic algorithms simulating the evolution on a phenotype and genotype level, respectively; harmony search method that is inspired by the underlying principles of the musicians' improvisation of the harmony; black hole algorithm that is motivated by the black hole phenomenon, namely, if a star gets too close to the black hole, it will be swallowed by the black hole and is gone forever; immunized evolutionary programming where adaptive mutation and selection operations are based on adjustment of artificial immune system; and cuckoo search based on

the brood parasitism of some cuckoo species, along with Levy flights random walks, just to mention some of them.

Particle Swarm Optimization

The idea of the particle swarm optimization (PSO) was inspired by the social behavior of big groups of animals, like flocking and schooling patterns of birds and fish, and suggested in 1995 by Russell Eberhart and James Kennedy, (see in Yang 2010).

Let us imagine a flock of birds circling over an area where they can smell a hidden source of food. The one who is closest to the food chirps the loudest, and the other birds swing around in his/her direction. If any of the other circling birds comes closer to the target than the first, it chirps louder and the others veer over toward him/her. This tightening pattern continues until one of the birds happens to land upon the food. So the motion of the individuals of the group is influenced by the motion of their neighborhood (Fig. 1).

In the searching space, we define discrete points as individuals (particles) which are characterized by their position vector determining their data (fitness) values to be maximized (similar to fitness function) and their velocity vectors indicating how much the data value can change and a personal best value indicating the closest the particle's data has ever come to the optimal value (target value).

Let us consider a population of N individuals, where their location vectors are $\vec{p}_i$, the objective function to be maximized is $\mathcal{F}$, and the fitness of an individual is $\mathcal{F}(\vec{p}_i)$. These values measure the attraction of the individual, namely, the higher its fitness value, the more the followers in its neighborhood, which will follow its motion. The local velocity of the individual $\vec{v}_i$ will determine its new location after Δt time step.

The velocity value is calculated according to how far an individual's data is from the target. The further it is, the larger is the velocity value. In the bird example, the individuals furthest from the food would make an effort to keep up with the others by flying faster toward the best bird (the individual having the highest fitness value in the population).



Particle Swarm Optimization in Geosciences, Fig. 1 Birds are swarming

Each individual's personal best value only indicates the closest the particle's data has ever come to the target since the algorithm started.

The best bird value only changes when any particle's personal best value comes closer to the target than best bird value. Through each iteration of the algorithm, best bird value gradually moves closer and closer to the target until one of the particles reaches the target.

Definitions and Basic Algorithm

We can give the following somewhat simplified basic algorithm of the PSO:

1. Generate the position and velocity vectors of each particle randomly:

$$P = (\vec{p}_0, \dots, \vec{p}_i, \dots, \vec{p}_N) \text{ and } V = (\vec{v}_0, \dots, \vec{v}_i, \dots, \vec{v}_N).$$

2. Compute the fitness of the particle:

$$\mathcal{F}(P) = (\mathcal{F}(\vec{p}_0), \dots, \mathcal{F}(\vec{p}_i), \dots, \mathcal{F}(\vec{p}_N)).$$

3. Store and continuously update the position of each particle where its fitness has the highest value:

$$P = (\vec{p}_0^{\text{best}}, \dots, \vec{p}_i^{\text{best}}, \dots, \vec{p}_N^{\text{best}}).$$

4. Store and continuously update the position of the best particle where its fitness has the highest value (global best):

$$\vec{p}_g^{\text{best}} = \max_{\vec{p} \in P} (\mathcal{F}(\vec{p})).$$

5. Modify the velocity vectors of each particle considering its personal best and the global best:

$$\vec{v}_i^{\text{new}} = \vec{v}_i + \varphi_1 (\vec{p}_i^{\text{best}} - \vec{p}_i) + \varphi_2 (\vec{p}_g^{\text{best}} - \vec{p}_i) \quad \text{for } 1 \leq i \leq N$$

Remark: Coefficient φ_1 ensures escaping from local optimum, since the motion of a particle does not follow the boss blindly!

6. Update the positions ($\Delta t = 1$):

$$\vec{p}_i^{\text{new}} = \vec{p}_i + \vec{v}_i^{\text{new}} \quad \text{for } 1 \leq i \leq N.$$

7. Stop the algorithm if there is no improvement in the objective function during a certain number of iterations or the number of iterations exceeds the limit.

8. Go back to (2) and compute new fitness value with the new positions.

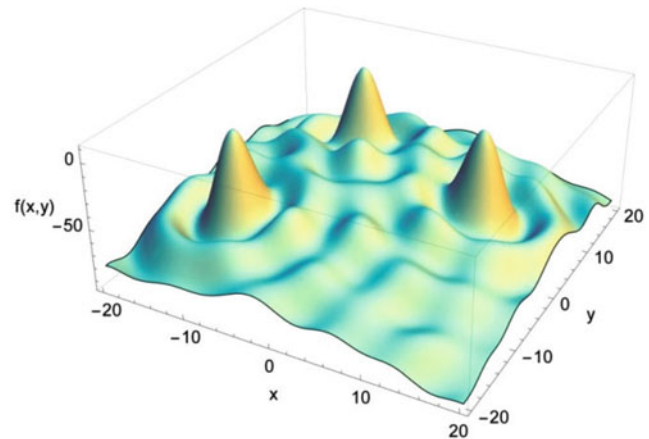
Illustrative Example

Let us demonstrate the algorithm with a 2D problem. Let us suppose that the local geoid can be described by the following function (see Fig. 2):

$$\begin{aligned} f(x, y) = & -\sqrt{(6+x)^2 + (-9+y)^2} - \\ & \sqrt{(-11+x)^2 + (-3+y)^2} - \sqrt{(11+x)^2 + (9+y)^2} + \\ & \frac{50 \sin \left[\frac{\sqrt{(6+x)^2 + (-9+y)^2}}{\sqrt{(6+x)^2 + (-9+y)^2}} \right]}{\sqrt{(6+x)^2 + (-9+y)^2}} + \\ & \frac{50 \sin \left[\frac{\sqrt{(-11+x)^2 + (-3+y)^2}}{\sqrt{(-11+x)^2 + (-3+y)^2}} \right]}{\sqrt{(-11+x)^2 + (-3+y)^2}} + \\ & \frac{50 \sin \left[\frac{\sqrt{(11+x)^2 + (9+y)^2}}{\sqrt{(11+x)^2 + (9+y)^2}} \right]}{\sqrt{(11+x)^2 + (9+y)^2}} \end{aligned}$$

The parameters to be initialized are

$N = 100$; (number of individuals),
 $\lambda = \{\{-20, 20\}, \{-20, 20\}\}$; (location ranges),
 $\mu = \{\{-0.1, 0.1\}, \{-0.1, 0.1\}\}$; (velocity ranges),
 $M = 300$; (number of iterations),
 $\varphi_1 = 0.2$; (local exploitation rate),
 $\varphi_2 = 2.0$; (global exploitation rate).



Particle Swarm Optimization in Geosciences, Fig. 2 The 2D objective function

Figure 3 shows the distribution of the individuals after 300 iterations, while Table 1 shows the results of different global methods.

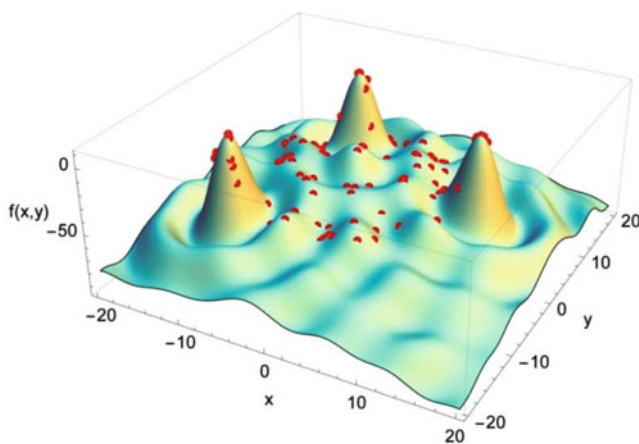
Variants of PSO

Although particle swarm optimization (PSO) has demonstrated competitive performance in solving global optimization problems, it exhibits some limitations when dealing with optimization problems with high dimensionality and complex landscape.

Numerous variants of even a basic PSO algorithm are possible in an attempt to improve optimization performance. There are certain research trends; one is to make a hybrid optimization method using PSO combined with other optimizers such as the genetic algorithm.

Another research trend is to try and alleviate premature convergence (i.e., optimization stagnation), e.g., by reversing or perturbing the movement of the PSO particles. Another approach of dealing with premature convergence is the use of multiple swarms (multi-swarm optimization).

Another school of thought is that PSO should be simplified as much as possible without impairing its performance, leading to the parameters being easier to fine-tune such that they perform more consistently across different optimization problems.



Particle Swarm Optimization in Geosciences, Fig. 3 The population distribution after 300 iterations

Particle Swarm Optimization in Geosciences, Table 1 The result of different global methods

Method	x_{opt}	y_{opt}	f_{opt}
Particle swarm	-6.01717	9.06022	10.8406
Simulated annealing	-6.01717	9.06022	10.8406
Differential evolution	-11.0469	-9.07418	5.3015
Nelder-Mead	-11.0469	-9.07418	5.3015

Initialization of velocities may require extra inputs; however, PSO variant has been proposed that does not use velocity at all.

As the PSO equations given above work on real numbers, a commonly used method to solve discrete problems is to map the discrete search space to a continuous domain, to apply a classical PSO, and then to remap the result, for example, by just using rounded values.

In general, an important variant and strategy of global optimization is to employ global method to find the close neighborhood of the global optimum and then apply local method to improve the result (Paláncz 2021).

PSO Applications in Geosciences

In this section, some case studies are introduced to illustrate the applicability of PSO algorithm in geosciences.

Particle Swarm Optimization for GNSS Network Design

The global navigation satellite systems (GNSS) are increasingly becoming the official tool for establishing geodetic networks. In order to meet the established aims of a geodetic network, it has to be optimized, depending on some design criteria (Grafarend and Sansò 1985). Optimization of a GNSS network can be carried out by selecting baseline vectors from all of the probable baseline vectors that can be measured in a GNSS network. Classically, a GNSS network can be optimized using the trial and error method or analytical methods such as linear or nonlinear programming or in some cases by generalized or iterative generalized inverses. Optimization problems may also be solved by intelligent optimization techniques such as genetic algorithms (GAs), simulated annealing (SA), and particle swarm optimization (PSO) algorithms. The efficiency and the applicability of PSO were demonstrated using a GNSS network, which has been solved previously using a classical method. The result shows that the PSO is effective, improving efficiency by 19.2% over the classical method (Doma 2013).

GNSS Positioning Using PSO-RBF Estimation Model

Positioning solutions need to be more accurate, precise, and obtainable at minimal effort. The most frequently used method nowadays employs a GNSS receiver, sometimes supported by other sensors. Generally, GNSS suffer from signal perturbations (e.g., from the atmosphere, nearby structures) that produce biases on the measured pseudo-ranges. With a view to optimize the use of the satellite signals received, a positioning algorithm with pseudo-range error modeling with the contribution of an appropriate filtering process includes, e.g., extended Kalman filter (EKF) and Rao-Blackwellized filtering (RBWF), which are among the most widely used algorithms to predict errors and to filter

the high frequency noise. A new method of estimating the pseudo-range errors based on the PSO-RBF model is suggested by Jgouta and Nsiri (2017), which achieves an optimal training criterion. The PSO is used to optimize the parameters of neural networks with radial basis function (RBF) in their work. This model offers appropriate method to predict the GNSS corrections for accurate positioning, since it reduces the positioning errors at high velocities by more than 50% compared to the RBWF or EKF methods (Jgouta and Nsiri 2017).

Using PSO to Establish a Local Geometric Geoid Model

There exist a number of methods for approximating local geoid surfaces (i.e., equipotential surfaces approximating mean sea level) and studies carried out to determine local geoids. In Kao et al. (2017), the PSO method as a tool for modeling local geoid is presented and analyzed. The ellipsoidal heights (h), derived from GNSS observations, and known orthometric heights (H) from first-order leveling from benchmarks were first used to create local geometric geoid model. The PSO method was then used to convert ellipsoidal heights (h) into orthometric heights (H). The resulting values were compared to those obtained from spirit leveling and GNSS methods. The adopted PSO method improves the fitting of local geometric geoid by quadratic surface fitting method, which agrees with the known orthometric heights within ± 1.02 cm (Kao et al. 2017).

Application of PSO for Inversion of Residual Gravity Anomalies over Geological Bodies with Idealized Geometries

A global particle swarm optimization (GPSO) technique was developed and applied by Singh and Biswas (2016) to the inversion of residual gravity anomalies caused by buried bodies with simple geometry (spheres, horizontal, and vertical cylinders). Inversion parameters, such as density contrast of geometries, radius of body, depth of body, location of anomaly, and shape factor, were optimized. The GPSO algorithm was tested on noise-free synthetic data, synthetic data with 10% Gaussian noise, and five field examples from different parts of the world. This study shows that the GPSO method is able to determine all the model parameters accurately even when shape factor is allowed to change in the optimization problem. However, the shape was fixed a priori in order to obtain the most consistent appraisal of various model parameters. For synthetic data without noise or with 10% Gaussian noise, estimates of different parameters were very close to the actual model parameters. For the field examples, the inversion results showed excellent agreement with results from previous studies that used other inverse techniques. The computation time for the GPSO procedure was very short (less than 1 s) for a swarm size of less than 50. The advantage of the GPSO method is that it is extremely fast and

does not require assumptions about the shape of the source of the residual gravity anomaly (Singh and Biswas 2016).

Application of a PSO Algorithm for Determining Optimum Well Location and Type

Determining the optimum type and location of new wells is an essential component in the efficient development of oil and gas fields. The optimization problem is, however, demanding due to the potentially high dimension of the search space and the computational requirements associated with function evaluations, which, in this case, entail full reservoir simulations. In Onwunalu and Durlofsky (2010), the particle swarm optimization (PSO) algorithm is applied to determine optimal well type and location. Four example cases are considered that involve vertical, deviated, and dual-lateral wells and optimization over single and multiple reservoir realizations. For each case, both the PSO algorithm and the widely used genetic algorithm (GA) are applied to maximize net present value. Multiple runs of both algorithms are performed, and the results are averaged in order to achieve meaningful comparisons. It is shown that, on average, PSO outperforms GA in all cases considered, though the relative advantages of PSO vary from case to case (Onwunalu and Durlofsky 2010).

Introducing PSO to Invert Refraction Seismic Data

Seismic refraction method is a powerful geophysical technique in near surface study. In order to achieve reliable results, processing of refraction seismic data in particular inversion stage should be done accurately. In Poormirzaee et al. (2014), refraction travel times' inversion is considered using PSO algorithm. This algorithm, being a meta-heuristic optimization method, is used in many fields for geophysical data inversion, showing that the algorithm is powerful, fast, and easy. For efficiency evaluation, different synthetic models are inverted. Finally, PSO inversion code is investigated in a case study at a part of Tabriz City in northwest of Iran for hazard assessment, where the field dataset are inverted using the PSO code. The obtained model was compared to the geological information of the study area. The results emphasize the reliability of the PSO code to invert refraction seismic data with an acceptable misfit and convergence speed (Poormirzaee et al. 2014).

PSO Algorithm to Solve Geophysical Inverse Problems: Application to a 1D-DC Resistivity Case

The performance of the algorithms was first checked by Fernández-Martínez et al. (2010) using synthetic functions showing a degree of ill-posedness similar to that found in many geophysical inverse problems having their global minimum located on a very narrow flat valley or surrounded by multiple local minima. Finally, they present the application of these PSO algorithms to the analysis and solution of an inverse problem associated with seawater intrusion in a

coastal aquifer in southern Spain. PSO family members were successfully compared to other well-known global optimization algorithms (binary genetic algorithms and simulated annealing) in terms of their respective convergence curves and the seawater intrusion depth posterior histograms (Fernández-Martínez et al. 2010).

Anomaly Shape Inversion via Model Reduction and PSO

Most of the inverse problems in geophysical exploration consist of detecting, locating, and outlining the shape of geophysical anomalous bodies imbedded into a quasi-homogeneous background by analyzing their effects on the geophysical signature. The inversion algorithm that is currently in use creates very fine mesh in the model space to approximate the shapes and the values of the anomalous bodies and the geophysical structure of the geological background. This approach results in discrete inverse problems with a huge uncertainty space, and the common way of stabilizing the inversion consists of introducing a reference model (through prior information) to define the set of correctness of geophysical models. A different way of dealing with the high underdetermined character of these kinds of problems consists of solving the inverse problem using a low dimensional parameterization that provides an approximate solution of the anomaly via particle swarm optimization (PSO), e.g., Fernández-Muñiz et al. (2020). This method has been designed for anomaly detection in geological setups that correspond with this kind of problem. These authors show its application to synthetic and real cases in gravimetric inversion performing at the same time uncertainty analysis of the solution. The two different parameterizations for the geophysical anomalies (polygons and ellipses) show that similar results are obtained. This method outperforms the common least-squares method with regularization (Fernández-Muñiz et al. 2020).

One-Dimensional Forward Modeling in Direct Current (DC) Resistivity

One-dimensional forward modeling in direct current (DC) resistivity is actually computationally inexpensive, allowing the use of global optimization methods (GOMs) to solve 1.5 D inverse problems with flexibility in constraint incorporation. GOMs can support computational environments for quantitative interpretation in which the comparison of solutions incorporating different constraints is a way to infer characteristics of the actual subsurface resistivity distribution. To this end, the chosen GOM must be robust to changes in the cost function and also be computationally efficient. The performance of the classic versions of the simulated annealing (SA), genetic algorithm (GA), and particle swarm optimization (PSO) methods for solving the 1.5 D DC resistivity inverse problem was compared using synthetic and field

data by Barboza et al. (2018). The main results are as follows: (1) All methods reproduce synthetic models quite well; (2) PSO and GA are comparatively more robust to changes in the cost function than SA; (3) PSO first and GA second present the best computational performances, requiring less forwarding modeling than SA; and (4) GA gives higher performance than PSO and SA with respect to the final attained value of the cost function and its standard deviation. To put them into effective operation, the methods can be classified from easy to difficult in the order PSO, GA, and SA as a consequence of robustness to changes in the cost function and of the underlying simplicity of the associated equations. To exemplify a quantitative interpretation using GOMs, these solutions were compared with least-absolute and least-squares norms of the discrepancies derived from the lateral continuity constraints of the log-resistivity and layer depth as a manner of detecting faults. GOMs additionally provided the important benefit of furnishing not only the best solution but also a set of suboptimal quasi-solutions from which uncertainty analyses can be performed (Barboza et al. 2018).

Summary

This study investigated the particle swarm method, its variants, and their applications in geosciences, i.e., optimization of GNSS network, approximation of local geoid surface, inversion of residual gravity anomalies, determination of optimal well locations, inversion of refraction seismic data, and solution of geophysical inverse as well as modeling direct current resistivity.

Bibliography

- Awange JL, Paláncz B, Lewis RH, Völgyesi L (2018) Mathematical geosciences: hybrid symbolic-numeric methods. Springer International Publishing, Cham, p. 596, ISBN: 798-3-319-67370-7
- Barboza FM, Medeiros WE, Santana JM (2018) Customizing constraint incorporation in direct current resistivity inverse problems: a comparison among three global optimization methods. *Geophysics* 83: E409–E422
- Doma MI (2013) Particle swarm optimization in comparison with classical optimization for GPS network design. *J Geodetic Sci* 3: 250–257. <https://doi.org/10.2478/jogs-2013-0030>
- Fernández-Martínez JL, Luis J, Gonzalo E, Fernandez P, Kuzma H, Omar C (2010) PSO: a powerful algorithm to solve geophysical inverse problems: application to a 1D-DC resistivity case. *J Appl Geophys* 71:13–25. <https://doi.org/10.1016/j.jappgeo.2010.02.001>
- Fernández-Muñiz Z, Pallero JL, Fernández-Martínez JL (2020) Anomaly shape inversion via model reduction and PSO. *Comput Geosci* 140. <https://doi.org/10.1016/j.cageo.2020.104492>
- Grafarend EW, Sansò F (1985) Optimization and design of geodetic networks. Springer, Berlin/Heidelberg. <https://doi.org/10.1007/978-3-642-70659-2>

- Jgouta M, Nsiri B (2017) GNSS positioning performance analysis using PSO-RBF estimation model. *Transp Telecommun* 18(2):146–154. <https://doi.org/10.1515/tjt-2017-0014>
- Kao S, Ning F, Chen CN, Chen CL (2017) Using particle swarm optimization to establish a local geometric geoid model. *Boletim de Ciências Geodésicas* 23. <https://doi.org/10.1590/s1982-21702017000200021>
- Onwunaliu JE, Durlinsky LJ (2010) Application of a particle swarm optimization algorithm for determining optimum well location and type. *Comput Geosci* 14:183–198. <https://doi.org/10.1007/s10596-009-9142-1>
- Paláncz B (2021) A variant of the black hole optimization and its application to nonlinear regression. <https://doi.org/10.13140/RG.2.2.28735.43680>. <https://www.researchgate.net/project/Mathematical-Geosciences-A-hybrid-algebraic-numerical-solution>
- Poormirzaee R, Moghadam H, Zarean A (2014) Introducing Particle Swarm Optimization (PSO) to invert refraction seismic data. In: *Conference Proceedings, near surface geoscience 2014 – 20th European meeting of environmental and engineering geophysics*, Sep 2014, vol 2014, pp 1–5. <https://doi.org/10.3997/2214-4609.20141978>
- Singh A, Biswas A (2016) Application of global particle swarm optimization for inversion of residual gravity anomalies over geological bodies with idealized geometries. *Nat Resour Res* 25:297–314. <https://doi.org/10.1007/s11053-015-9285-9>
- Yang X (2010) *Nature-inspired metaheuristic algorithms*, 2nd edn. Luniver Press, Frome

Pattern

Uttam Kumar

Spatial Computing Laboratory, Center for Data Sciences,
International Institute of Information Technology Bangalore
(IIITB), Bangalore, Karnataka, India

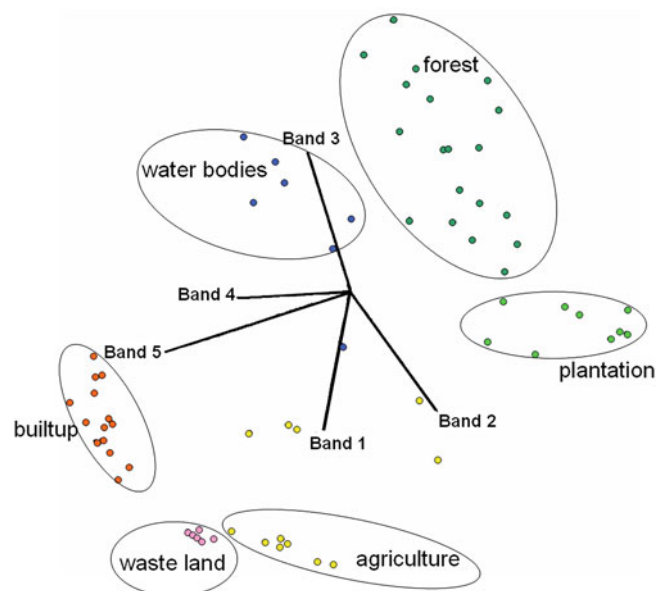
Definition

Spatial data analysis and spatial data mining encompass techniques to find pattern, detect anomalies and test hypotheses from spatial data (Shekhar and Xiong 2008). Pattern denotes the way in which we see and interpret a behavior and present them, where the term behavior refers to a particular, objectively existing configuration of characteristics (Andrienko and Andrienko 2006). A pattern is a construct reflecting essential features of a behavior in a parsimonious manner, i.e., in a substantially shorter and simpler way than specifying every reference and the equivalent characteristics. In the context of data mining, a pattern is an expression E in a language L describing facts in a subset F_E of a set of facts F (dataset) so that E is simpler than the enumeration of all facts in F_E (Andrienko and Andrienko 2006). So, a pattern may reflect a combination of qualities, acts, tendencies, etc., forming a consistent or characteristic arrangement. A pattern results from observation or analysis, an image of a behavior showing how an observer might understand it, which is often subjective. Therefore, different observers understand the

same behavior differently and represent it by different patterns. It is also apparent that one observer may use different patterns to describe the same behavior.

Introduction

Geographic knowledge discovery finds novel and useful geospatial knowledge hidden in Big Data. This is a human-centered process involving many activities such as data cleaning, data reduction, and data mining through scalable techniques to extract patterns to form knowledge (Shekhar and Xiong 2008). In fact, the basis of data mining is to discover patterns occurring in the database, such as associations, classification models, sequential patterns, etc. (Han and Kamber 2006). Areas of machine learning such as recognition, classification, clustering, prediction, association mining, sequential pattern mining, etc. constitute pattern discovery. Recognition and classification are complementary functions because classification is concerned with establishing criteria that can be used to identify or distinguish patterns appearing in the data. These criteria can depend on representatives of each class to numeric parameters for measurement, to syntactical descriptions of key features. Recognition is the process by which tools are subsequently used to find a particular pattern within the data (Russ 2002). Clustering of data has wider applications in many domains, for example, spatial data clustering of remotely sensed images from space-borne satellites to identify various land use classes (see Fig. 1).



Pattern, Fig. 1 Clustering of pixels for different land use classes in multispectral data having 5 spectral bands

For efficient data mining, the evaluation of pattern interestingness is made as deep as possible into the mining process so as to confine the search to only the interesting patterns. Often, domain knowledge are used to guide the discovery process to facilitate concise pattern discovery at various levels of abstraction. Integrity constraints and business rules help in evaluating the interestingness of the discovered patterns (Han and Kamber 2006).

Data pattern can be mined from many different kinds of database such as data warehouse, relational database, transactional database, object-relational database, spatial database, time-series database, sequence database, text database, multimedia database, data streams, World Wide Web, etc. Patterns such as frequent itemsets, subsequences and substructures occur frequently in the data. For example, in a transactional dataset of a supermarket, frequent itemset refers to a set of items that frequently appear together (example, milk and bread), whereas a frequent occurring subsequence that customers may tend to purchase from an electronic retail shop are computer and printer followed by ink cartridges, which represents a sequential pattern. A substructure can refer to different structural forms such as graph, trees, lattice, etc. Mining these frequent patterns leads to the discovery of interesting associations and correlations within data (Han and Kamber 2006).

In the context of spatial data, a tessellation (or tiling) is often used as a way of partition of space into mutually exclusive cells that together make up the complete study space. These are used for vector-based representation to characterize geographic fields and objects through a regular pattern. Examples of two regular tessellations are shown in Fig. 2 that represent square and hexagonal cells. It is fascinating to see that in all regular tessellations, cells are of the same shape and size, and that the field attribute value assigned to a cell is associated with the entire area occupied by the cell. Square cell tessellation pattern is common because of their ease in geo-referencing. It is interesting to notice how tessellation represents different patterns for building topology (By 2004).

Pattern can be described by terms like radial, checkboard, etc. In the context of terrain data, pattern refers to the spatial arrangement of objects and implies the characteristic

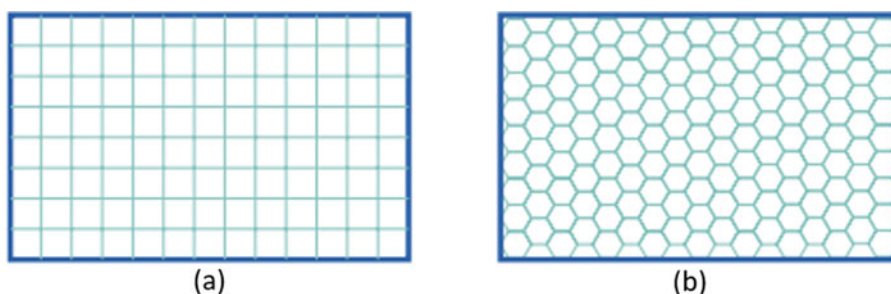
repetition of certain forms or relationships. For example, land use types have specific characteristic patterns such as irrigation types, different types of housing in urban areas, rivers with tributaries, etc. as observed from satellite images. Therefore, pattern is normally used as an interpretation element in image analysis (Kerle et al. 2004).

Types of Patterns

Pattern can be expressed mathematically, through attributes or numbers such as pattern observed in the recorded summer temperature across a city. A pattern may be broadly classified into the following types:

- Association pattern – it means description of a set of references as a unified whole on the basis of similarity of their features (Andrienko and Andrienko 2006). This unification is based on identical or similar characteristics where the attribute values correspond to the references. For example, cities can be grouped together as a cluster based on their population, where population is a close characteristic. This idea can be extended to account for clustering based on multiple attributes such as a cluster of cities with similar male and similar female proportions or low male and high female proportions or high male and low female proportions, etc. Association patterns are supplemented with summary characteristics of a combination of references from the individual characteristics. Numeric characteristics are summarized by aggregation of individual attributes into a single feature such as mean and variance, etc.
- Arrangement pattern – this defines the way characteristics are arranged in certain order for which time can be a reference to signify increasing or decreasing trends (Andrienko and Andrienko 2006). Pattern may also mean stability. The pattern ordering may be random such as cities with increasing population or regular such as a chessboard with black and white boxes. Spatial trends occur in spatial locations that do not exhibit natural

Pattern, Fig. 2 Tessellation: pattern in a square and hexagonal cell representation



ordering, however, some parameters like directions and distances in space can be used for ordering.

- **Differentiation pattern** – here, references are united because they are different from the characteristics of other remaining references/neighborhood that may form another pattern (Andrienko and Andrienko 2006). For example, grouping of regions that have relatively higher traffic than the neighborhood in the northern part of a city. Likewise, similar regions with equal traffic density may also occur in other parts (south, east, west, etc.) of the city. In such examples, references are done on some characteristics along with differentiation operation. Differentiation is possible even in the absence of common references, where the reference may have different characteristics from the remaining reference set. Outlier (local or global) fall in this category, where local outliers differ significantly from their neighborhood and global outliers have references that differ from the entire reference set. A subset of references can also be differentiated from the remaining references because of higher variability of characteristics in one subset than in other references.
- **Compound pattern** – this type of pattern emerges when the summarization of data involves division of a reference set into parts and association of elements on the basis of similar features, resulting in a compound pattern that comprises several subpatterns (Andrienko and Andrienko 2006). In other words, when the distribution does not clearly show a spatial pattern or trend, it can be described as a combination of two or more patterns called compound pattern. For example, “*spatial autocorrelation states that locations that are close are more likely to have similar values than locations that are far apart*” involves distance as a reference set because the characteristics change gradually, i.e., the neighboring objects differ less than the distant objects. In such cases, it makes sense to describe a spatial trend as a compound pattern (combination of two patterns).
- **Distribution summary pattern** – this defines the notion of the distribution (variance) of characteristics over a reference set (Andrienko and Andrienko 2006). Statistical mean, median, quartiles, quantiles, percentiles and variance in different types of distributions are often used to summarize the data. For spatially referenced data, center, entropy, etc. are used to summarize the homogeneity or heterogeneity.
- **Patterns in spatio-temporal data** – spatio-temporal data refers to data that are varying in both space and time domain, i.e., data point moving over space and time, for example, traffic flow at different times on the road.

A spatio-temporal pattern characterizes the spatial relationships among a collection of spatial entities and the behavior of such relationships over time, therefore they are often

multifaceted. Spatial relationship is complex and diverse, i.e. for any pair of spatial entities, a variety of spatial relations exist such as directional, distance based, and topological that are often application specific. Evolving spatial clusters generate pattern which have large number of spatial entities that are similar to one another in the neighborhood. So, spatio-temporal patterns have large number of spatial and temporal relations while spatial clusters may be restricted to relationship based on distances between entities. Spatio-temporal pattern studies are widely used in traffic analysis, security systems, epidemic studies and disease control, etc. (Shekhar and Xiong 2008). Handling spatio-temporal dataset still remains a challenge.

In this context, movement patterns in spatio-temporal data refer to events expressed by a set of entities such as movement of humans, traffic jams, migration of birds, bird flock movement, etc. In these cases, a pattern starts and ends at certain times and may be constrained to a subset of space. In geoscience, pattern is used with various meanings conceptualized as salient movement of events in the geospatial context. Application areas of movement pattern in spatio-temporal data are animal behavior, human movement (for example, tracking humans through mobile phones, cameras, etc.) for traffic management, surveillance and security, in military, sport’s action analysis, etc. The challenge rests in relating movement patterns with the underlying space to answer where, when, and why the entities move the way they do. Therefore, patterns need to be conceptualized with the surroundings (Shekhar and Xiong 2008).

There are a few interesting points that naturally emerge as a result of the pattern analysis of Big Data. It is important to state that not all the patterns are interesting and only a small fraction of them are of interest to a user. So, what makes a pattern interesting? A pattern is interesting if it is easily understood, valid on new dataset with some degree of certainty, useful and novel. A pattern is also interesting if it validates a hypothesis that the user sought to confirm because an interesting pattern also characterizes information that will turn into knowledge. There are several objective measures of pattern interestingness which are based on the structure of discovered patterns. For example, an objective measure for association rules is support and confidence. Support represents the percentage of transactions from a database that the given rule satisfies while confidence assesses the degree of certainty of the detected association (Han and Kamber 2006).

Pattern in Geoscience

There has been several studies that use artificial intelligence based techniques to reduce the multitude of phenomenon in the natural environment into discrete manageable classes. This is because humans are very good at detecting pattern,

but relatively poor in distinguishing subtle brightness difference such as in pixel values (reflectance) in multispectral imageries. In general, the term pattern recognition is the act of taking in raw data and making an action based on the category of the pattern (Duda et al. 2000). In spatial data analysis, pattern recognition or classification is a method used widely by which labels are attached to data (pixels) in view of their spectral characteristics (of the remote-sensing data). This labeling is implemented by any classification algorithm such as maximum likelihood classifier, random forest, support vector machine, etc. that are trained beforehand to recognize pixels with spectral pattern and similarities to derive land use classes. The method of land use detection or recognition has proved to be efficient, objective and reproducible.

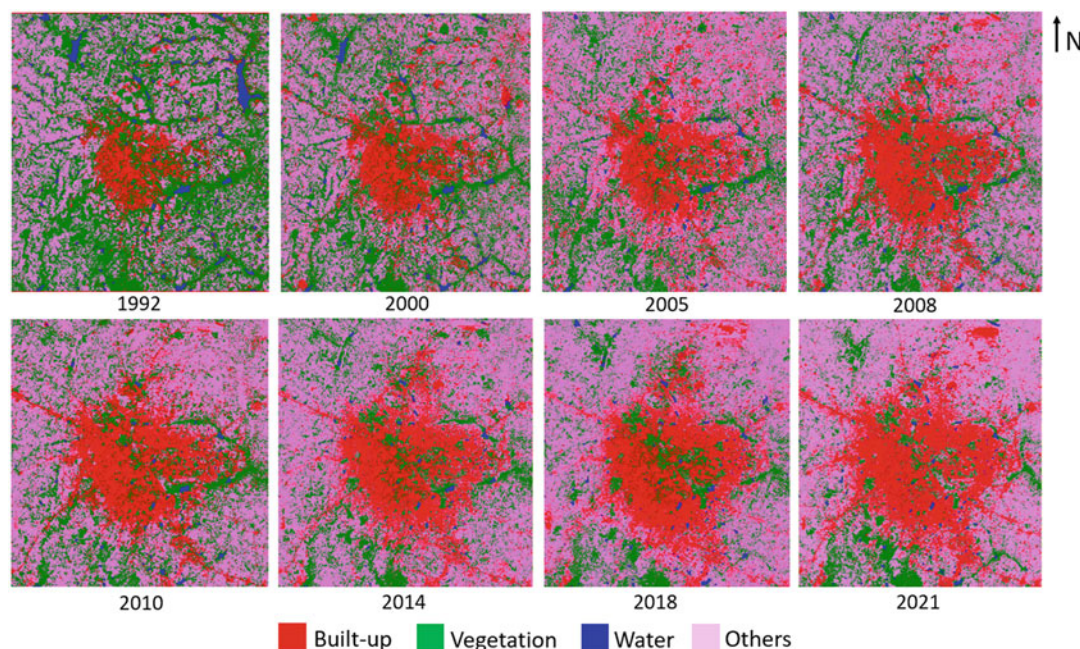
Fig. 3 shows the spatio-temporal pattern of land use, highlighting the phenomenon of urban sprawl of Bangalore City, India derived from time-series Landsat data using a machine-learning approach. The patterns of land use classes such as built-up (including buildings, concrete surfaces, paved roads, parking lots, etc.), vegetation (gardens, lawns, parks, forest, agriculture/horticulture farms, etc.), water bodies (lakes, ponds, etc.), and others class (including fallow land, barren land, rocks and play grounds with soil) are clearly distinguishable. The time-series plot shows the trend/pattern of the growth of urban areas and reduction in the area of vegetation and water bodies as depicted in Fig. 4. Note that the detailed land use statistics and accuracy assessment of the

time-series-classified images have not been discussed here. Such geospatial analysis is very useful in understanding the past and present scenario of a spatial area and helps to model the trend to predict the future urban growth in various directions and distances from the City Center.

Pattern analysis has been used in many studies such as discovering climate change pattern using spatio-temporal dataset, analyzing temporal trends in precipitation and temperature, assessing the spatial distribution of global vegetation cover, for multivariate spatial clustering and geovisualization, discovering pseudopatterns, etc. (Miller and Han 2009). There are other applications including spatio-temporal video data mining to analyze abnormal activity, target analysis, content-based video retrieval under spatiotemporal distribution, urban sprawl pattern analysis, urban temperature analysis, (Li et al. 2015), spatial point pattern analysis (Cressie 2015), etc.

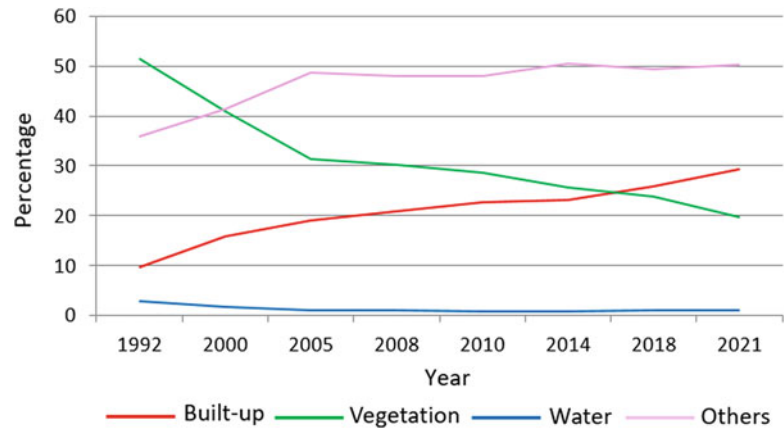
Summary

Pattern refers to the spatial arrangement of objects and implies the characteristic repetition of certain forms or relationships. Here, different types of patterns were broadly discussed. As such, there are no rules for selecting a pattern for a given problem. Discovering an interesting and useful pattern is subjective, so patterns symbolizing important characteristics need to be explored. Common properties for a set of



Pattern, Fig. 3 Spatio-temporal pattern of the urban sprawl of Bangalore City, India

Pattern, Fig. 4 Trend of the spatio-temporal change in land use classes in Bangalore City, India



references are easy to find out, however, distinguishing properties are rather difficult to discover. Patterns need to be distinctive and relevant to the goals of the investigation related to an application. Aggregation at an appropriate level helps in avoiding noninteresting results from summary characteristics. The degree of variation of the characteristics pertaining to the references along with the outliers has to be considered. Finally, pattern depends on the purpose of the analysis, whether a precise specification of a subset is required or an allusion is sufficient.

Cross-References

- [Big Data](#)
- [Cluster Analysis and Classification](#)
- [Machine Learning](#)
- [Pattern Classification](#)
- [Support Vector Machines](#)

Bibliography

- Andrienko N, Andrienko G (2006) Exploratory analysis of spatial and temporal data a systematic approach. Springer, Berlin. <https://link.springer.com/book/10.1007/3-540-31190-4>
- Cressie NAC (2015) Statistics for spatial data. Wiley, Hoboken
- de By RA (2004) Principles of geographic information systems. ITC Educational Textbook Series: The Netherlands
- Duda RO, Hart PE, Stork DG (2000) Pattern classification, 2nd edn. A Wiley-Interscience Publication, New York. ISBN 9814-12-602-0
- Han J, Kamber M (2006) Data mining concepts and techniques. Elsevier, San Francisco
- Kerle N, Janseen LLF, Huurneman GC (2004) Principles of remote sensing. ITC Educational Textbook Series: The Netherlands
- Li D, Wang S, Li D (2015) Spatial data mining theory and application. Springer, Berlin, Heidelberg
- Miller HJ, Han J (2009) Geographic data mining and knowledge discovery, 2nd edn. CRC Press, Boca Raton
- Russ JC (2002) The image processing handbook, 4th edn. CRC Press, Boca Raton
- Shekhar S, Xiong H (2008) Encyclopedia of GIS. Springer, New York

Pattern Analysis

Sanchari Thakur¹, LakshmiKanthan Muralikrishnan², Bijal Chudasama³ and Alok Porwal⁴

¹University of Trento, Trento, Italy

²Cybertech Systems and Software Limited, Thane, Maharashtra, India

³Geological Survey of Finland, Espoo, Finland

⁴Center of Studies in Resources Engineering, IIT Bombay, Mumbai, India

Synonyms

Pattern recognition; Spatial association

Definition

Pattern refers to the rules governing the arrangement of physical events or occurrences in space and time. Depending on the dynamic or static nature of the events, patterns can be spatial, temporal, and spatio-temporal. Patterns can be identified from geoscientific datasets by visualization and mathematical techniques collectively known as pattern analysis. These techniques are chosen depending on the geospatial representation of the events, e.g., as discrete points, lines, and polygons, or as continuous field values (raster or images). Pattern analysis of single type of data (univariate) can indicate whether the events tend to occur close to (clustered) or away from (dispersed) each other, while multivariate pattern analysis can also reveal how the attributes of the events are inter-related.

Pattern Analysis in Geosciences

Geological entities (events) are represented as points, lines, polygonal vectors, and rasters defined by their spatial

coordinates and associated with single or multiple attributes. For instance, at regional scale, mineral deposits are considered as points, lineaments as lines, lithological units as polygons, and surface elevation as raster. The underlying processes and rules governing the occurrence of these geological events result in distinguishable patterns of inter-relationships between them. These patterns can be identified by well-established methods. Statistical pattern analysis techniques test the hypothesis that the observed events are a realization of a random process. Several instances of the process are generated by Monte Carlo simulations. If the observed pattern occurs within the simulation envelope, it represents complete spatial randomness (CSR). Deviations from randomness are indications of positive association (or clustering, i.e., events tend to occur close to each other) or negative association (or dispersion, i.e., events tend to occur away from each other) (Fig. 1). In this chapter, we review some of the pattern analysis approaches with examples of their application in geosciences.

Patterns Analysis in Univariate Data

A univariate point data is defined within a spatio-temporal framework as the presence or absence of an event at a given location and/or time. A spatial point pattern (SPP) X consists of two main components – a set of points with well-defined coordinates $x_i \in \mathcal{R}^2$ (2D cartographic coordinates) and an observation window, that is, a delimiting boundary within which the analysis is restricted. An SPP can be extended to include the time of occurrence of each event to generate a spatio-temporal point pattern (STPP) $\{(x_i, t_i)\}$. In this case, the observation window is bounded in space, $x \in [x_1, x_2]$, and time, $t \in [t_1, t_2]$.

There are several graphical and statistical methods for SPP analysis (Baddeley 2015). Graphical methods visually and quantitatively highlight the regions of heterogeneity in the

PP. The simplest approach divides the observation window into equal and contiguous blocks of any size and shape (quadrats). Clustering is inferred in a quadrat if it has higher number of points in relation to the other quadrats. Alternatively, in kernel smoothed density function (KSDF) techniques, the PP is convolved with a kernel (such as Gaussian) to estimate the probability of occurrence (intensity) of points at every location in the observation window. A constant intensity throughout the window implies a homogeneous point process. Spatially varying intensity could indicate either inhomogeneity or clustering.

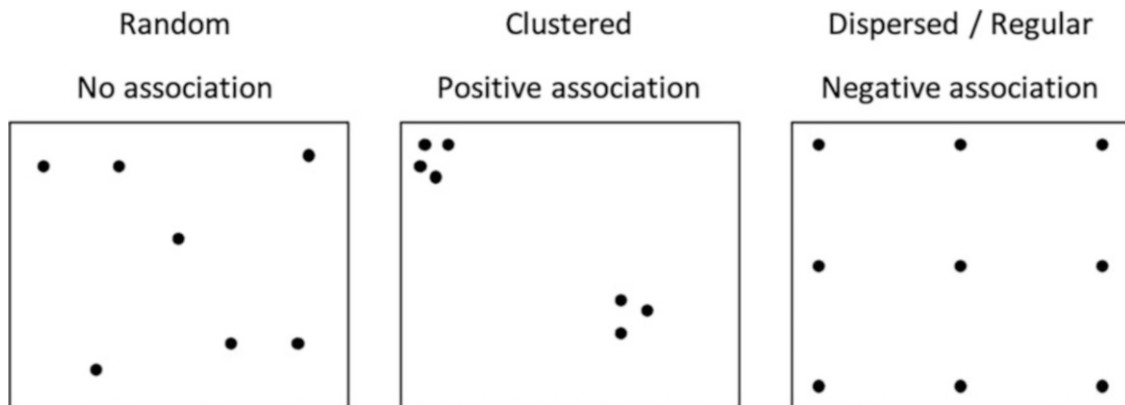
Such inferences can be associated with a level of statistical significance by hypothesis testing methods. For example, a chi-square goodness-of-fit test compares the observed quadrat counts to that expected from a homogeneous point Poisson process to infer CSR. However, distance-based techniques are more widely used for statistical SPP analysis. These are based on estimating the frequency distribution of the distance r between the closest pairs of points (Baddeley 2015, Diggle 2013, Lisitsin 2015). The inter-event distance calculator, G-function, is given by:

$$G(r) = \frac{\sum_{i \neq j}^N \text{count}(\min \|x_i - x_j\| < r)}{N} \quad (1)$$

Here, the numerator calculates the number of points x_i whose nearest neighbor x_j is closer than the distance r . N is the total number of points. The point-to-event distance calculator, F-function, is given by:

$$F(r) = \frac{\sum_{i=1}^M \sum_{j=1}^N \text{count}(\min \|u_i - x_j\| < r)}{N} \quad (2)$$

It calculates the distances to the nearest event x_j from selected M arbitrary points u_i in the empty space. The F-function provides a measure of the shapes of the empty spaces between the point patterns and is particularly useful when the



Pattern Analysis, Fig. 1 Different categories of spatial patterns, where the events are represented as points

observation window is a concave polygon. The inter-event and point-to-event functions are combined into another summary statistic called the J-function given by:

$$J(r) = \frac{1 - G(r)}{1 - F(r)} \quad (3)$$

$J(r) = 1$ represents CSR. If the empty spaces are more (i.e., $F(r)$ is small) and the points are closer to each other (i.e., $G(r)$ is large), $J(r) < 1$ and the PP is clustered. Conversely, $J(r) > 1$ indicates dispersion.

G , F , and J are first-order distance functions, in that only the nearest points are considered. There are well-established multi-distance functions that consider multiple points at different distances from each other, such as the K-function (Baddeley 2015). These methods can be used for hypothesis testing by appropriately selecting the point processes such as homogeneous Poisson point process (for constant point intensity), inhomogeneous Poisson point process (for variable but predictable intensity), doubly stochastic Poisson process, or Cox process (if intensity is also random) (see ► “Point Pattern Statistics”). In general, border correction is necessary to account for censoring errors caused by restricting the PP analysis to within the boundary of an observation window.

Similar to SPP, STPP analysis aims at understanding the interaction among events (Diggle 2013) in terms of the degree of clustering, but in both space and time. Thus, SPP analysis methods can be extended to STPP analysis by adding the temporal coordinate. Statistical STPP analysis methods include the use of Knox’s statistic, scan statistics, and tests related to the nearest-neighbor distribution, and second-order properties of the point process expressed through the Ripley’s K-function for homogeneous and inhomogeneous spatio-temporal point processes.

An application of STPP analysis is the modeling of earthquake events, which are known to trigger aftershocks close to the epicenter and occurrence time of a high-magnitude earthquake. Shi et al. (2009) used STPP analysis to study the sequences of aftershocks of the Wenchuan earthquake of 2008. They found the aftershock events to be clustered within 60 km and 260 hours from the high-magnitude earthquake. Besides points, univariate analysis applied to single-band rasters identifies relationships between neighboring pixels (e.g., texture analysis, hydrological and terrain modeling) using image processing techniques.

Patterns in Multivariate Data

Different geological processes are intricately linked to one another. Therefore, it seldom suffices to independently analyze only the coordinates of events that are also associated with other geological information. If each point x_i in an

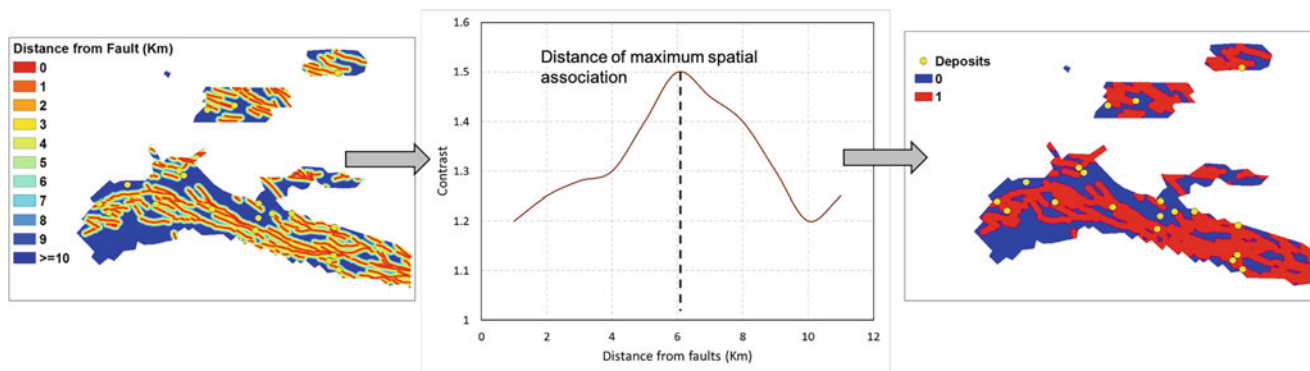
SPP/STPP is associated with a mark m , the set of the point-mark combinations $\{(x, m)\}$ describes a marked point pattern (MPP) (Ho and Stoyan 2008). In geoscience applications, the marks are typically real-valued attributes. In the previous example of earthquake analysis if each earthquake event is also associated with its magnitude, it describes a marked STPP.

The mark can be modeled as a spatially continuous random field Z . However, this model assumes marks and points to be mutually independent, thus inhibiting the pattern analysis of correlations between them. In contrast, density-dependent marked Cox processes allow the modeling of inter-relationships between the points and the marks (Stoyan et al. 1995; Ho and Stoyan 2008). Several MPP distance functions are extensions of SPP analysis, for example, the normalized mean product of marks of a pair of points separated by a distance r . Common MPP processes include log Gaussian Cox process, intensity-marked Cox process, and geostatistical model for preferential sampling (Ho and Stoyan 2008).

A special case of MPPs is when the marks are categorical. Consider n point datasets $X_1, X_2, \dots, X_n$ each containing objects having the same categorical mark. Irrespective of the mark, any two points in the union of the n datasets are defined as neighbors if the distance between them is less than a given distance r_0 . A co-location pattern is defined by an undirected connected graph with points representing nodes and edges representing the neighborhood relationships between them. Patterns are mined from the graph based on spatio-temporal topological constraints and defining proximity rules that associate the objects of X_i with those of X_j , $i \neq j$. These association rules are defined using interestingness measures such as participation ratio, prevalence, and confidence (Mamoulis 2008). An application of co-location analysis is in determining the association of water reservoirs with incidence of diseases.

Patterns between vector geometries can be identified using probabilistic approaches such as the Weights of Evidence (WofE) model based on Bayesian method (Agterberg et al. 1993). It involves quantification of spatial association (in terms of weights) between the targeted geometry (hypothesis) and the evidential geometry (evidence). Figure 2 shows an application of WofE to mineral prospectivity modeling, where the hypothesis to be predicted is the occurrence of mineral deposits (points) and the evidential features are faults (lines). Here, the spatial association between deposits and the presence/absence of faults is assessed by the estimation of posterior probabilities.

Patterns between points and their enclosing polygons can be identified using the SPP analysis techniques by defining the polygons as observation windows. Mamuse et al. (2010) applied distance-based PP analysis to the exploration of nickel sulfide deposits within prospective komatiites in Kalgoorlie terrane. Considering the terrane as observation window, the nickel sulfide deposits and komatiite bodies



Pattern Analysis, Fig. 2 WofE method for identifying distance of spatial association between points and lines by calculating contrast, that is, difference between the probability of deposits within the feature and probability of deposits outside the feature

(represented as points) were found to be clustered. However, considering the komatiites as observation windows, the deposits were found to be randomly distributed or dispersed. This observation was used to estimate the number of deposits in a komatiite body and to identify the spatial controls on nickel sulfide mineralization.

Pattern analysis of multivariate rasters includes evaluation of the correlation and covariance matrices, and cluster analysis. In these methods, the geoscientific information from co-located pixels of each raster band is concatenated to form a feature-vector representing the set of attributes at a given location (corresponding to the pixel). The superset of all feature-vectors extracted from the full raster is then analyzed in the feature-space (where the support axes represent the different variables rather than the spatio-temporal coordinates in the original dataspace) to identify patterns of inter-relationships between the different variables.

A diagnostic step for advanced multivariate analyses involves computation of a correlation matrix, which estimates the pair-wise correlation coefficients between the variables or a covariance matrix, which characterizes the monotonic trend between the variables (positive covariance indicates variations in the same direction, while negative covariance indicates inverse directional variations). These matrices are precursors for advanced multivariate analysis such as principal component analysis (PCA), which is based on eigenvalue decomposition of the covariance matrix. PCA is used for identifying geochemical patterns of anomalous elemental concentrations or alteration patterns indicating mineralization or bedrock lithology. Graphs of the PC loadings indicate that elements (variables) located in proximity in the PC plots are closely related in the geochemical domain (e.g., Filzmoser et al. 2009).

Clustering (e.g., k-means, Isodata) algorithms are also used for finding patterns in the feature-space. Its simplicity lies in the calculation of Euclidean distances between the feature-vectors and grouping them into clusters by their proximity. The identified clusters are mapped back to the corresponding geospatial domains. The actual calculations

are iterative for the identification of cluster centroids and grouping of pixels until the centroids have stabilized or the maximum number of iterations has been reached (see ► “Pattern Recognition”).

Summary and Conclusions

We have seen different types of spatial and spatio-temporal patterns and the techniques for analyzing them as univariate or multivariate data by graphical methods, pattern mining, statistical testing, and probabilistic modeling. Pattern analysis supports the understanding of the underlying physical processes that cause the occurrence of the events at a particular location and/or time. This capability finds interesting applications in geosciences such as (1) interpolation, that is, estimating the values of spatio-temporal variables where observed data is not available; (2) forecasting, that is, predicting occurrence of events; (3) modeling, that is, identifying the underlying processes responsible for the occurrence of the pattern; and (4) simulations, that is, virtual regeneration of physical reality.

Apart from the methods described in this chapter, there are several supervised, semi-supervised, and unsupervised techniques for identifying patterns in images (either single or multiband) and applying them to the automatic classification of geological entities. Recently, pattern analysis has been extended to pattern recognition which incorporates advanced machine learning techniques (such as artificial neural networks, self-organizing maps, convolutional networks), which enable mineralization-related pattern identification from input raster data (see ► “Machine Learning”).

Cross-References

- [Machine Learning](#)
- [Mineral Prospectivity Analysis](#)

- [Multiple Point Statistics](#)
- [Pattern](#)
- [Pattern Classification](#)
- [Pattern Recognition](#)
- [Point Pattern Statistics](#)
- [Predictive Geologic Mapping and Mineral Exploration](#)
- [Spatial Analysis](#)
- [Spatial Autocorrelation](#)
- [Spatial Statistics](#)

Bibliography

- Agterberg FP, Bonham-Carter GF, Cheng QM, Wright DF (1993) Weights of evidence modelling and weighted logistic regression for mineral potential mapping. *Comput Geol* 25:13–32
- Baddeley A (2015) *Spatial point patterns: methodology and applications with R*. CRC Press
- Diggle PJ (2013) *Statistical analysis of spatial and spatio-temporal point patterns*. CRC Press
- Filzmoser P, Hron K, Reimann C (2009) Principal component analysis for compositional data with outliers. *Environmetrics* 20(6):621–632
- Ho LP, Stoyan D (2008) Modelling marked point patterns by intensity-marked Cox processes. *Stat Prob Lett* 78(10):1194–1199
- Lisitsin V (2015) *Spatial data analysis of mineral deposit point patterns: applications to exploration targeting*. *Ore Geol Rev* 71:861–881
- Mamoulis N (2008) Co-location patterns, algorithms. In: Shekhar S, Xiong H (eds) *Encyclopedia of GIS*. Springer, Boston
- Mamuse A, Porwal A, Kreuzer O, Beresford S (2010) Spatial statistical analysis of the distribution of komatiite-hosted nickel sulfide deposits in the Kalgoorlie terrane, western Australia: clustered or not? *Econ Geol* 105(1):229–242
- Shi P, Liu J, Yang Z (2009) Spatio-temporal point pattern analysis on Wenchuan strong earthquake. *Earthq Sci* 22(3):231–237
- Stoyan D, Kendall WS, Mecke J (1995) *Stochastic geometry and its applications*. Wiley, New York

Pattern Classification

Katherine L. Silversides
Australian Centre for Field Robotics, Rio Tinto Centre for Mining Automation, The University of Sydney, Camperdown, NSW, Australia

Definition

Pattern classification is the automated grouping or labeling of samples based on features or patterns in a dataset (Bishop 2006; Duda et al. 2001).

Introduction

Pattern classification involves using the features present in a dataset to automatically label each test point using a machine

learning method. It is a supervised learning method, where a training dataset is used to train the model. The training dataset contains both the input data and the corresponding label for each data point. The trained model is then used to classify the test dataset using the input data to predict the corresponding label (Bishop 2006; Duda et al. 2001).

Pattern classification can be performed on any dataset that can be divided into a set number of classes, each with distinct features in the data. In the geosciences this can include lithofacies classification (Bhattacharya et al. 2016), material classification (Banchs and Rondon 2007), geological mapping (Cracknell et al. 2014), wireline log interpretation (Song et al. 2020), anomaly detection (Gonbadi et al. 2015), and lithology classification (Bressan et al. 2020). Many different data types can be used for the classification including geophysical logs (Bhattacharya et al. 2016; Bressan et al. 2020; Cracknell et al. 2014; Song et al. 2020), geochemical samples (Cracknell et al. 2014; Gonbadi et al. 2015), airborne geophysical data (Cracknell et al. 2014), and seismic surveys (Banchs and Rondon 2007).

Machine learning methods that can be used for classification include support vector machine (SVM) (Bhattacharya et al. 2016; Bressan et al. 2020; Gonbadi et al. 2015), neural networks (NN) (Banchs and Rondon 2007; Bhattacharya et al. 2016; Song et al. 2020), self-organizing maps (SOM) (Bhattacharya et al. 2016; Cracknell et al. 2014), random forests (RF) (Bressan et al. 2020; Cracknell et al. 2014; Gonbadi et al. 2015), boosting (Gonbadi et al. 2015), and decision trees (Bressan et al. 2020). The best method to apply varies between different applications.

NN, SVM, and RF and examples of their application are discussed below. SOMs were designed as an unsupervised learning method; however supervised learning versions have been developed for classification tasks (Bhattacharya et al. 2016; Cracknell et al. 2014; Lau et al. 2006). For more details on SOMs, please refer to the “SOM” entry in this encyclopedia.

Neural Networks

Neural networks (NN) are a popular method for data classification that was designed to mimic the way that the human brain processes information. NN typically contain at least three layers. The first is an input layer that takes the input data and passes it to the hidden layer. The hidden layer learns the relationships between the input variables, then passes weighted data patterns to the output function. This output layer then determines the output, which for classification problems is a label. Multiple hidden layers can be used for complicated problems. In training, a NN is initially assigned random weight coefficients for the hidden layer. Training data is then repeatedly fed through the network and the weight coefficients are modified until the difference between the

outputs and the desired classifications are minimized. Iterative back-propagation is often used for this learning step. When designing a NN there are numerous parameters that can be altered, such as the number of hidden layers, the number of nodes in each hidden layer, learning rate, damping coefficient, and the number of iterations used for learning (Bishop 2006; Bhattacharya et al. 2016; Duda et al. 2001). Many different variations of NN exist, with variations on how the layers are built and the method used for training common changes. More information can be found in the ► [“Artificial Neural Network”](#) entry.

Bhattacharya et al. (2016) apply several different machine learning methods to classify multiple mudstone lithofacies from each other and calcareous siltstone and limestone lithofacies from petrophysical well logs. One of the methods applied is a standard NN using a single hidden layer. For this case, the authors found that the best results were achieved using a SVM.

Banchs and Rondon (2007) apply a Hopfield neural network (HNN) to several example cases, classifying areas of high and low porosity, or areas of sand or shale, using seismic attributes as the input. In a HNN the nodes in the hidden layer are binary instead of weighted, so each node is either on or off.

Song et al. (2020) classified shapes in in wireline logs using recurrent neural networks (RNN). This involved identifying four basic log shapes in spontaneous potential log segments. RNNs use previous outputs as part of the input to recognize trends and larger features. This model used six layers including a layer of Long Short Term Memory (LSTM) units. The RNN model was found to outperform a standard NN as well as other machine learning techniques for this problem.

Support Vector Machine

Support Vector Machine (SVM) is a popular machine learning method that maps the input data to a higher dimensional space to increase the distance between data points from different classes. A hyperplane can then be used to divide the points into classes. Support vectors are the points that are on the edges of the classes, closest to the hyperplane. The distances between the support vectors represent the margin. A model with larger margins is considered to be more robust, and less likely to misclassify edge points (Bhattacharya et al. 2016; Bressan et al. 2020; Schölkopf et al. 2000).

A kernel function is used to describe the mapping between the input space and the new higher dimensional space. Possible kernels include linear, polynomial, radial basis function (RBF), and multilayer perceptron. This kernel is optimized using the training data to minimize the margin. While ideally there should be no misclassifications in the training data, perfect separation may be impossible due to overlapping

classes. In this case a small number of errors may be tolerated. While SVM is a binary classifier, it can be used for multi-class problems using either a one-vs-all or pairwise one-vs-one method (Bhattacharya et al. 2016; Bressan et al. 2020). More information can be found in the ► [“Support Vector Machines”](#) entry.

Bhattacharya et al. (2016) used SVM with RBF to classify multiple mudstone lithofacies from each other and calcareous siltstone and limestone lithofacies from petrophysical well logs. The authors found that in this case SVM performed better than both NN and SOM.

Bressan et al. (2020) used SVM to classify sedimentary well logs into four lithology groups based on seven geophysical logs. The authors found that the SVM method used produced poor results, and suggested that this was due to the log data not containing a highly defined pattern, and insufficient data for proper training. In comparison, RF performed well on this data.

Gonbadi et al. (2015) applied SVM with a RBF kernel to identify geochemical anomalies in a porphyry copper to target new exploration drilling. Despite testing using two different feature selection methods, the authors found that a boosting method (AdaBoost) was better suited to this problem.

Random Forest

Random forest (RF) is a pattern classification method that is based on an ensemble of decision trees (Breiman 2001). Each individual decision tree is created by independently splitting the data using sequential queries at each node. This splitting divides the tree into branches, with leaves (each resembling a class) at the end of each sub-branch. Each decision tree has a different set of rules. A classification prediction can then be calculated by averaging the results from all of these trees, combining many weak classifiers into a single strong classifier. As more trees are included in the RF the generalization error converges to a limit without overfitting. The generalization error depends on both the strength of the trees and the correlation between the trees (Breiman 2001; Gonbadi et al. 2015).

The generalization error and variance of the RF can be reduced using bagging and random selection of features. Bagging involves generating a random individual subset, with replacement, of the dataset to train each decision tree. The samples not used in the training are then used to estimate the generalization error. For each of these trees a set number of features are randomly chosen to use in the decision nodes. As this method does not overfit, it has been shown to perform well in classifications involving multisource and high-dimensional data (Cracknell et al. 2014; Gonbadi et al. 2015). More information can be found in the ► [“Random Forest”](#) and ► [“Decision Trees”](#) entries.

Cracknell et al. (2014) used RF to classify geophysical and geochemical data into 21 discrete lithological units. The data was from an area with known volcanic-hosted massive sulfide (VHMS) deposits where field mapping is difficult due to poor outcrops and thick vegetation. Comparing the RF map to the geological interpretation demonstrated that the RF map not only performed well, but also included additional details about the spatial distributions of key lithologies.

Bressan et al. (2020) applied RF to seven geophysical logs to classify sedimentary well logs into four lithology groups. Compared to SVM, decision tree and multilayer perceptron (MLP) methods, RF performed well on this data. In two of the three scenarios considered RF obtained the best results. RF also performed well in all scenarios for cross-validation.

Gonbadi et al. (2015) applied a RF with 350 trees to identify geochemical anomalies to identify the best targets for new exploration drilling in a porphyry copper region. RF was chosen because it does not assume that the variables have a normal distribution. However, the authors found that a boosting method (AdaBoost) was better suited to this problem.

Summary

Pattern classification involves using a machine learning method to automatically group or label samples based on features or patterns in a dataset. It is a supervised learning method, requiring a training dataset that contains both the input data and the corresponding label for each data point. The trained model can then be used to predict the classification of new samples. Many different classification methods can be used. Popular methods include NN, SVM, and RF. Pattern classification has been applied to problems such as lithology or material classification, geological mapping, wireline log interpretation, and anomaly detection.

Cross-References

- [Artificial Neural Network](#)
- [Decision Tree](#)
- [Random Forest](#)
- [Self-Organizing Maps](#)
- [Support Vector Machines](#)

Bibliography

- Banchs R, Rondon O (2007) Seismic pattern recognition based upon Hopfield neural networks. *Explor Geophys* 38:220–224
- Bhattacharya S, Carr TR, Pal M (2016) Comparison of supervised and unsupervised approaches for mudstone lithofacies classification: case

- studies from the Bakken and Mahantango-Marcellus Shale, USA. *J Nat Gas Sci Eng* 33:1119–1133
- Bishop CM (2006) *Pattern recognition and machine learning*. Springer, Berlin, 710p
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32
- Bressan TS, de Souza MK, Girelli TJ, Chemale F Jr (2020) Evaluation of machine learning methods for lithology classification using geophysical data. *Comput Geosci* 139:104475
- Cracknell MJ, Reading AM, McNeill AW (2014) Mapping geology and volcanic-hosted massive sulfide alteration in the Hellyer–Mt Charter region, Tasmania, using Random Forests™ and self-organising maps. *Aust J Earth Sci* 61(2):287–304
- Duda R, Hart P, Stork D (2001) *Pattern classification*, 2nd edn. Wiley, New York, 654p
- Gonbadi AM, Tabatabaei SH, Carranza EJM (2015) Supervised geochemical anomaly detection by pattern recognition. *J Geochem Explor* 157:81–91
- Lau KW, Yin H, Hubbard S (2006) Kernel self-organising maps for classification. *Neurocomputing* 69(16–18):2033–2040
- Schölkopf B, Smola AJ, Williamson RC, Bartlett PL (2000) New support vector algorithms. *Neural Comput* 12(5):1207–1245
- Song S, Hou J, Dou L, Song Z, Sun S (2020) Geologist-level wireline log shape identification with recurrent neural networks. *Comput Geosci* 134:104313

Pattern Recognition

Rogério G. Negri

São Paulo State University (UNESP), Institute of Science and Technology (ICT), São José dos Campos, São Paulo, Brazil

Definition

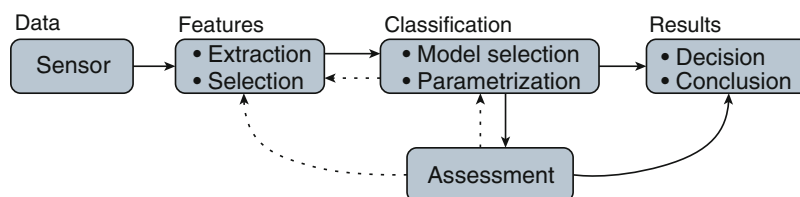
Pattern recognition comprises a field of study regarding how machines and algorithms may recognize patterns and structures on data sets and making decisions about them.

Pattern recognition embraces a research area that focuses on problems of classification and object description. With multidisciplinary aspects, the pattern recognition studies interface with statistic, engineering, artificial intelligence, computer science, data mining, image and signal processing, etc. Typical examples of applications are automatic character recognition, medical diagnosis, bank customer profiles monitoring, and even the most recent issues like recommendations systems and face recognition.

Process Overview

A pattern recognition process is not limited to apply methods in data sets but also includes distinct stages from the problem formulation, data collection, and representation until classification, assessment, and interpretation. Figure 1 depicts a common pattern recognition workflow.

Pattern Recognition,
Fig. 1 Pattern recognition
 workflow overview



In a broad view, a “pattern” is usually represented by a vector whose components values stands for a specific attribute previously measured by a sensor. Waves captured from an acoustic system, the number of financial transactions in a period, the spectral responses registered by a digital camera, and the climatic variables at a specific instant are examples of patterns in practical situations.

Based on the observed data and aiming to simplify or identify non-obvious information about them, associations are made between each of these data to a specific class or group/cluster. This pattern-to-class association process is called *classification*, which demands the modeling (i.e., choosing a model and respective parameterization) of a rule able to perform such a task.

As an intermediate part of this process, it needs to highlight the data preprocessing and assessment of the classification results. Computing additional values from the objects to provide a better description leads to a different stage known as “feature extraction.” Regarding all the available information, it is essential to use only the meaningful attributes in the pattern recognition process that favors error reduction and accuracy improvement. To do so, a careful “feature selection” is usually performed. Also, computing the error/accuracy levels should be systematically and persistently carried in an “assessment” stage.

Learning Paradigms

According to the previous discussions, a pattern recognition process lies in modeling a decision rule that can classify patterns regarding classes or clusters. The learning paradigm is directly related to how the modeling is carried.

Among different paradigms in the literature, the supervised, unsupervised, and semi-supervised are the most common. However, it is valid to mention other paradigms, for example, reinforcement, evolutionary, and instance-based learning.

The main characteristic that distinguishes supervised and unsupervised learning refers to how the classification rules are modeled. While supervised learning builds a model using information from a set of ground-truth examples in which the expected response is known (i.e., the class), unsupervised learning is based on similarities identified over the set of patterns without information about presumed classes. Examples whose response/class is known are called “labeled

patterns.” The set of labeled patterns required by supervised methods is defined as “training set.”

When the classes comprised by the classification are previously defined, supervised learning is preferable. However, insufficiently labeled data availability may turn unfeasible a supervised training. In this case, unsupervised learning rises as an alternative. Nonetheless, it should stress that results based on such a learning paradigm are restricted to groups/clusters of similar patterns without assigning them to a class.

These limitations became a starting point to the development of the semi-supervised paradigm, usually defined as a midterm between supervised and unsupervised, since it simultaneously uses labeled and unlabeled patterns to model the classification rules (Chapelle et al. 2006).

Approaches

Beyond the learning paradigm, the pattern recognition methods are also characterized according to the approach used to define the decision rules. Examples of approaches comprise template matching, statistical models (parametric or non-parametric), construction of decision surfaces based on geometric/structural criteria, and neural networks.

Template matching involves comparing and identifying similarities between patterns taken as referential (i.e., the templates) and other patterns whose class is unknown (Pavlidis 2013). This process should be independent of scale, rotation, and translation changes.

The statistical approach establishes decision regions using probability distributions representing each class included in the classification problem (Webb and Copsey 2011). The Bayes decision theory is an essential mathematical referential to define the rules and make decisions. Naive Bayes, maximum likelihood classification, and *K*-nearest neighbors are examples of statistical approach methods.

Methods with a geometric/structural approach are characterized by decision rules expressed by separation hyperplane or partitions in the attribute space (Bishop 2006). Support vector machines, decision trees, graph-based models, and the *K*-means algorithm follows the geometric/structural approach.

Lastly, methods based on the neural network approach simulate an arrangement of neurons that adapt to solve complex tasks with nonlinear input and outputs (Haykin 2009). Among many neural networks proposed in the literature, the multilayer perceptron, self-organizing maps, and convolutive neural networks are popular.

In addition, it is worth highlighting the possibility of combining the cited approaches into an ensemble scheme (Theodoridis and Koutroumbas 2008). Methods like AdaBoost and random forest are examples of ensembles widely adopted.

Applications in the Geosciences

Several studies and applications in geosciences employ pattern recognition concepts. Examples include seismic analysis, meteorology, land use and land cover mapping, and natural disaster prevention.

Curilem et al. (2016) use pattern recognition methods to extract features from seismic signals and build a system for volcanic events classification. Chattopadhyay et al. (2020) demonstrate that the integration of clustering and neural network concepts provides a potential method for weather pattern forecasting. Extreme rainfall forecasting using pattern recognition is another topic of relevance discussed in Nayak and Ghosh (2013). Distinctly, as proposed in Wang et al. (2015), an approach based on the support vector machine method allows reconstructing the historical data of weather type, which comprises an essential component for photovoltaic power forecasting. Mapping the land use and land cover is another topic of extreme relevance in several applications related to the geosciences (Giri 2012).

Summary

Pattern recognition comprises a research area focused on developing and applying mathematical and computational concepts to solve data classification problems. The amount and variety of pattern recognition applications in geosciences are vast.

Cross-References

- Pattern
- Pattern Analysis
- Pattern Classification

Bibliography

- Bishop CM (2006) Pattern recognition and machine learning (information science and statistics). Springer, Berlin/Heidelberg
- Chapelle O, Scholkopf B, Zien A (2006) Semi-supervised learning. MIT Press, Cambridge, MA
- Chattopadhyay A, Hassanzadeh P, Pasha S (2020) Predicting clustered weather patterns: a test case for applications of convolutional neural networks to spatio-temporal climate data. *Sci Rep* 10(1317). <https://doi.org/10.1038/s41598-020-57897-9>

- Curilem M, Huenupan F, Beltrn D, San-Martin C, Fuentealba G, Franco L, Cardona C, Acua G, Chacn M, Khan MS, Becerra Yoma N (2016) Pattern recognition applied to seismic signals of Llama Volcano (Chile): an evaluation of station-dependent classifiers. *J Volcanol Geotherm Res* 315:15–27. <https://doi.org/10.1016/j.jvolgeores.2016.02.006>
- Giri CP (2012) Remote sensing of land use and land cover: principles and applications. Remote sensing applications series. CRC Press
- Haykin S (2009) Neural networks and learning machines, vol 10. Prentice Hall
- Nayak MA, Ghosh S (2013) Prediction of extreme rainfall event using weather pattern recognition and support vector machine classifier. *Theor Appl Climatol* 114:583603. <https://doi.org/10.1007/s00704-013-0867-3>
- Pavlidis T (2013) Structural pattern recognition. Springer series in electronics and photonics. Springer, Berlin/Heidelberg
- Theodoridis S, Koutroumbas K (2008) Pattern recognition, 4th edn. Academic, San Diego
- Wang F, Zhen Z, Mi Z, Sun H, Su S, Yang G (2015) Solar irradiance feature extraction and support vector machines based weather status pattern recognition model for short-term photovoltaic power forecasting. *Energ Buildings* 86:427–438. <https://doi.org/10.1016/j.enbuild.2014.10.002>
- Webb AR, Copsey KD (2011) Statistical pattern recognition, 3rd edn. Wiley. <https://doi.org/10.1002/9781119952954>

Pawlowsky-Glahn, Vera

Ricardo A. Olea
Geology, Energy and Minerals Science Center,
U.S. Geological Survey, Reston, VA, USA



Fig.1 Vera Pawlowsky-Glahn, courtesy of Prof. Juan José Egozcue

Biography

Vera met her professional destiny the day she had a meeting with her adviser Wolfdietrich Skala to choose the topic of her dissertation. He offered her a job as collaborator on a compositional data analysis project, but she wanted to work on geostatistics. Vera first proved that spurious correlation was a

problem in both disciplines. The rest is history. Vera started a long spatiotemporal trek that took her to the top specialist on each discipline: Georges Matheron in France and Aitchison in Hong Kong. Progress toward the solution to the mapping of compositional data came in several installments. First was her dissertation (Pawlowsky 1986), then a translation with a few new developments (Pawlowsky-Glahn and Olea 2004) and finally the complete solution to the problem via the formulation of the Aitchison geometry (Pawlowsky-Glahn and Egozcue 2001). Over the years, in cooperation with multiple scientists around the world, Vera expanded her interest to all aspects of compositional modeling, writing more than 100 papers. The main accomplishments of her career are summarized in Pawlowsky-Glahn et al. (2015). Without Vera, Aitchison's ideas may have been lost or, at best, it would have taken much longer to receive the acceptance that they have today. The compositional data specialists around the world prepared a Festschrift (Filzmoser et al, 2021) in appreciation for Vera's contribution to the field.

Upon finishing graduate school, Vera spent her career with the Polytechnical University in Barcelona (1986–2000), serving as Associate Professor, and with the University of Girona (2000–2018), where she reached the position of Chair of the Computer Sciences, Applied Mathematics and Statistics Department. Vera has been quite active in the International Association for Mathematical Geosciences (IAMG), organizing the third IAMG conference in 1997, serving as chair of several committees, convener of numerous sessions, the 2007 Distinguished Lecturer, and the 11th President (2008–2012). IAMG presented her with the 2006 Krumbein Medal for her valuable contributions to science, the profession and IAMG, and also the 2008 Cedric Griffiths Teaching Award. She was instrumental in starting the Compositional Data Association in 2015, serving as the founding president (2015–2017).

Vera was born in Barcelona on 25 September 1951, the seventh of nine children. As a Latvian-Russian-German descendant, she graduated from high school from the *Deutsche Schule* of Barcelona in 1970. She attended the University of Barcelona, enrolling in the Mathematics Department, not in the Biology Department as she had originally planned, because there was no waiting queue to enroll in Mathematics. She earned the equivalent of a master's degree in mathematics in 1982. After that, she decided to pursue graduate studies in Germany, where she received a PhD in natural sciences from the Free University of Berlin in 1986. Vera's partner is Juan José Egozcue and she is the mother of zoo, aquarium, and wildlife veterinarian Tania Monreal-Pawlowsky.

Bibliography

- Filzmoser P, Hron K, Martín-Fernández JA, Palarea-Albaladejo J, 2021. Advances in compositional data analysis: Festschrift in honour of Vera Pawlowsky-Glahn. Springer.
- Pawlowsky V (1986) Räumliche Strukturanalyse und Schätzung Ortsabhängiger Kompositionen mit Anwendungsbeispielen aus der Geologie. Dissertation, Free University of Berlin, 170 pp
- Pawlowsky-Glahn V, Egozcue JJ (2001) Geometric approach to statistical analysis on the simplex. *Stoch Env Res Risk A* 15(5):384–398
- Pawlowsky-Glahn V, Olea RA (2004) Geostatistical analysis of compositional data. Oxford University Press, Oxford, 181 pp
- Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015) Modeling and analysis of compositional data. Wiley, Chichester, 247 pp

Pengda, Zhao

Qiuming Cheng¹ and Frits Agterberg²

¹State Key Lab of Geological Processes and Mineral resources, China University of Geosciences, Beijing, China

²Central Canada Division, Geomathematics Section, Geological Survey of Canada, Ottawa, ON, Canada



Fig. 1 photographed at the cultural party celebrating the 60th anniversary of the founding of China University of Geosciences. https://en.wikipedia.org/wiki/Zhao_Pengda (CC BY-SA 3.0)

Biography

Professor Zhao Pengda was born in Qingyuan, Liao Ning Province in 1931. He graduated from the Geological Department of Beijing University in 1952, whereupon he was appointed as assistant to teach in the newly founded Beijing Geological Institute. In 1954, he was sent to Moscow, USSR as a postgraduate student, and studied methods of prospecting and exploration for mineral deposits under Professor A. A. Yakren at the Moscow Geological Surveys College obtaining his Ph.D. in 1958. The title of his dissertation was: “The geological characteristics and methods of exploration for stockwork type tin and tungsten ore deposits.” In 1960, he became Associate Professor at the Beijing Geological Institute, where he attained the rank of full professor in 1980. Meanwhile, the Institute was moved to Wuhan and named Wuhan College of Geology in 1975. Appointed its President in 1983, Professor Zhao retained this post following reorganization of the college in 1987 to create the China University of Geosciences (CUG). Later during his reign, CUG opened its second campus in Beijing.

Professor Zhao has been researching the use of mathematical and statistical methods in mineral exploration since 1956. In 1976, he began applying mathematical models to predict and assess mineral resources, specifically iron and copper deposits in a Mesozoic volcanic basin in southern China. This resulted in the book “Statistical Prediction for Mineral Deposits” (Zhao et al. 1983). In total, he has published three books and over 120 papers. His later research includes work on expert systems, statistical theory and methods of prediction, and integrated modeling of geochemical anomalies and ranking by favorability. He has taught courses in mathematical geology, statistical analysis of geoexploration data, and statistical prediction of mineral deposits.

In 1990, Professor Zhao organized and chaired the international workshop: “Statistical Prediction and Assessment for Mineral Deposits,” held in Wuhan and attended by 14 experts from outside China including Dr. Richard McCammon, then President of the International Association of Mathematical Geosciences (IAMG). As Father of Mathematical Geology in China, Professor Zhao has been a member of the IAMG for many years, furthering the field of mathematical geology through teaching and publications (e.g., Zhao 1992). In 1993 he was awarded the William Christian Krumbein Medal, which is the highest award the IAMG can bestow upon one of its members. In 1993, Professor Zhao also became Academician of the Chinese Academy where he chaired its geoscience section for many years. After his retirement as President of the China University of Geosciences he remains actively involved in research and supervision of Ph.D. students. In May 2020, CUG organized Professor Zhao’s 90th Birthday Celebration.

Bibliography

- Zhao P (1992) Theories, principles and methods for the statistical of mineral deposits. *Math Geol* 24:589–592
 Zhao P, Hu W, Li Z (1983) Statistical prediction of mineral deposits. Geological Publishing House, Beijing. (in Chinese)

Plurigaussian Simulations

Nasser Madani

School of Mining and Geosciences, Nazarbayev University, Nur-Sultan city, Kazakhstan

Definition

Plurigaussian simulation is a stochastic technique for spatial modeling of categorical variables. An important characteristic of this technique is that the relationship and contact between categories can be identified by domain expert and imposed into the process of modeling.

Introduction

Categorical variables in geoscience-related fields usually are divided into two parts. The first part, nominal variables are the categories without any ordering. Some examples of this family are lithology, facies, rock types, alternations, permeable/impermeable, etc. The second part, ordinal variables are the categories with an intrinsic ordering or ranking. Several examples of this family are lithofacies sequence in sedimentary environment, class of mineral grades and geochemical anomalies (e.g., poor, medium, rich), intensity of minerals in a rock sample quantified on a quantitative scale (e.g., none, little, moderate, much, very much), clarity and color grades of particular minerals and stones, etc. Let’s denote the categorical variables as geo-domains throughout this text. Three-dimensional spatial modeling of these geo-domains can be implemented either by deterministic or stochastic approaches. In the former, the layout of the boundaries between two adjacent geo-domains can be specified based on the geological knowledge and one may obtain only one single scenario for positioning of this boundary. In the latter, however, a stochastic process determines the probable area of the boundaries among the geo-domains. In this context, several geostatistical simulation techniques are developed in the last decades. Among others, plurigaussian simulations are of paramount importance for modeling the geo-domains where some particular restrictions exist in contact boundaries. For instance, in modeling of a mineral deposit, from the

geological perspectives, it is possible that some of geo-domains never touch each other, meaning that there is no contact between them. This characteristic, known as forbidden contact can be modeled in resulting maps/scenarios by plurigaussian simulations.

History

Theoretical background of plurigaussian simulations was first introduced by Galli et al. (1994). In the past almost 26 years since the initiation of the first idea, many scholars used this technique for modeling the categorical variables in reservoir characterization, ore body modeling, rock fractures, and hydrogeology (Armstrong et al. 2011).

Essential Concepts

Plurigaussian simulations are natural extension of truncated Gaussian simulation. Matheron et al. (1987) first introduced the truncated Gaussian simulation for the purpose of simulating first the lithofacies in an oil reservoir and then to simulate the porosity and permeability within the built lithofacies. This methodology is developed for the cases when the geo-domains follow a sequential ordering. Plurigaussian simulation, however, is developed to model more complicated types of contacts that exist among the underlying geo-domains. For instance, Fig. 1 shows an example resulting map of truncated and plurigaussian simulations for modeling four geo-domains. Figure 1a shows that there is a sequence from geo-domain A to geo-domain D, which is obtained from the former method while Fig. 1b illustrates that the geo-domain A only touches geo-domain B and never is in contact with geo-domains C and D.

The rationale behind the truncation Gaussian and plurigaussian simulation is that the former considers one Gaussian random field (Fig. 1c), and the latter employs two or more Gaussian random fields (Fig. 1c and d) that are simulated entire area under consideration, and then a truncation rule (Fig. 1e and f) is used to turn these Gaussian values into geo-domains. Hence, the actual application of the plurigaussian simulations requires defining five main steps:

Step 1: Choosing the Truncation Rule

As shown in Fig. 1, the Gaussian random fields (Fig. 1c and d) need to be truncated in order to obtain the categorical maps (Fig. 1a and b). This needs definition of a truncation rule known as “flag,” where one can identify how the geo-domains are in content together. For instance, in Fig. 1e, the flag can represent the sequential contact between the geo-domains A to D and Fig. 1f depicts the flag that geo-domain A is

only in contact with geo-domain B and the latter is only in contact with geo-domains C and D. As a result, the configuration of the underlying flag that should be identified based on the geological interpretation of the domain identifies the permissible and forbidden contacts between geo-domains whether or not to comply with chronological ordering.

Step 2: Estimating the Truncation Thresholds

Splitting the Gaussian maps, given the preidentified flag, to obtain the categorical maps requires determining the corresponding thresholds. These numerical values can be obtained based on the proportion of each geo-domain. For N Gaussian random fields belonging to a vector random field $\mathbf{Z} = \{\mathbf{Z}(\mathbf{x}) : \mathbf{x} \in R^d\}$ with N components, and M geo-domains, $M - 1$ thresholds can be identified. Let's denote $g(\cdot)$ as the joint probability density function of the Gaussian random fields. For the purpose of computing the probability of presenting the i -th geo-domain at a particular sample location $\mathbf{x}$, one requires resolving the following formula:

$$P_i(\mathbf{x}) = \int_{D_i} g(z_1, \dots, z_N) dz_1 \dots dz_N \quad (1)$$

Where P_i is the proportion and D_i is the segment of the flag linked to the i th geo-domain (a subset of R^N defined by the thresholds). This formula may be resolved numerically or by trial and error.

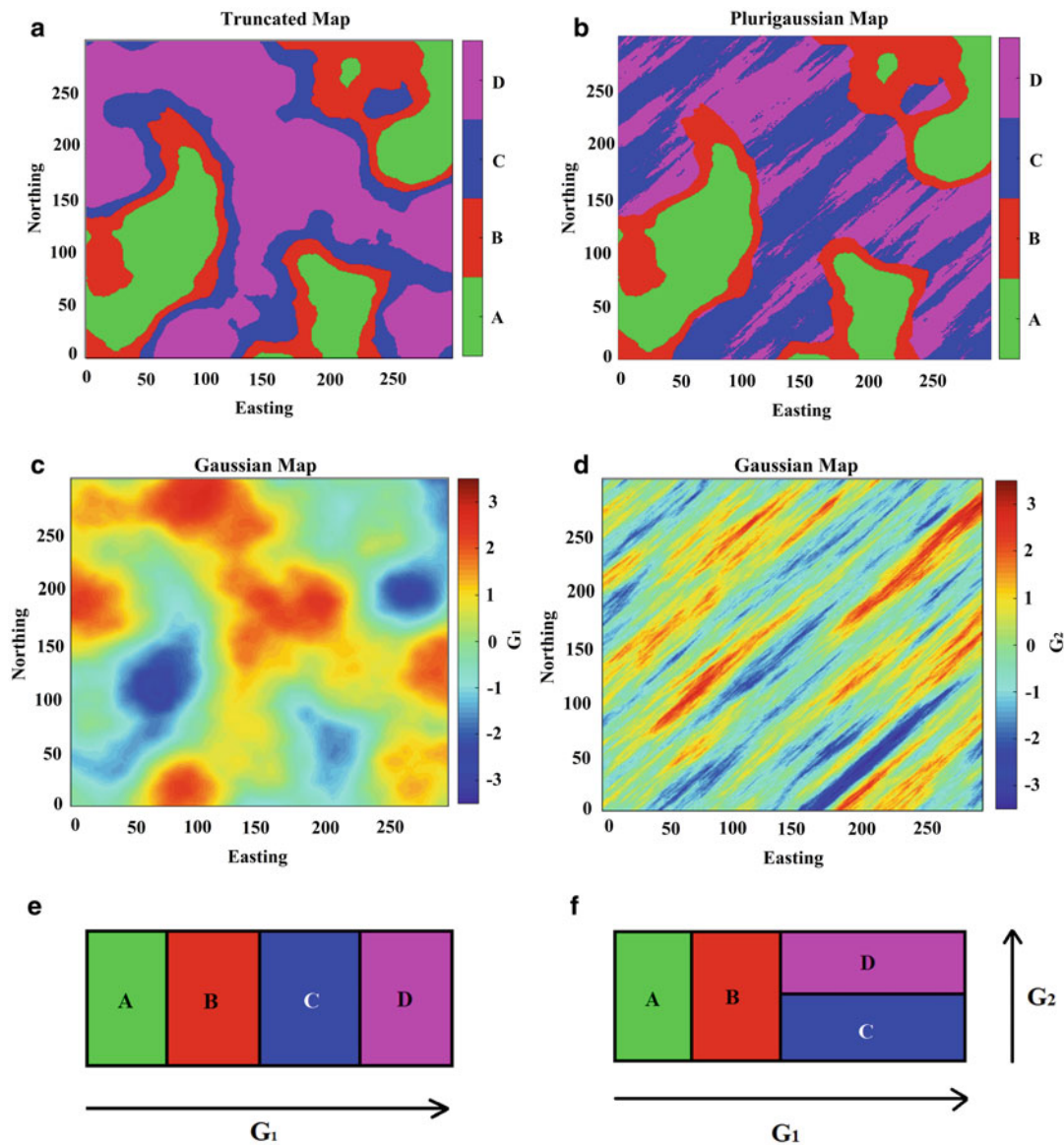
Step 3: Inferring the Variogram of the Gaussian Random Fields

Variogram inference of the underlying Gaussian random field (s) is a key factor in plurigaussian simulation. Based on the mathematical relationship between the indicator variograms and the variograms of the Gaussian random field(s), one may infer a proper theoretical model for the variogram of interest by deriving its properties. The experimental indicator variogram can be computed experimentally through the recorded geo-domains at sample locations. For any separation vector h , the indicator cross variogram between two geo-domains (with indices i and j) is deduced by related non-centered covariance (Armstrong et al. 2011):

$$\gamma_{ij}(\mathbf{h}) = C_{ij}(\mathbf{0}) - \frac{1}{2} [C_{ij}(\mathbf{h}) + C_{ij}(-\mathbf{h})] \quad (2)$$

$$C_{ij}(h) = \text{prob}\{G(x) \in D_i, G(x+h) \in D_j\} \quad (3)$$

If D_i and D_j are rectangular parallelepipeds of R^N and the components of the vector random field G are independent, then the second member of Eq. (3) is a function of the direct covariances or variograms of the components of G and can be computed by utilizing Hermite polynomials (Emery 2007).



Plurigaussian Simulations, Fig. 1 Schematic representation of truncated and plurigaussian simulation techniques

Step 4: Generating Gaussian Random Fields at Sampling Points

Once the variogram of Gaussian random fields and the corresponding threshold(s) are obtained, the categorical variable at sampling points should be converted to the Gaussian values. This step generates the Gaussian values that first respects the variogram fitted in step 3 and also falls in the proper intervals imposed by flag and the truncation thresholds. For this purpose, a method called Gibbs sampler (Casella and George 1992) is used to generate such Gaussian values. The common procedure is as follows:

Initialization

In each sample point u_α , simulate a vector with N components u_α in $D_{i(\alpha)}$, where $i(\alpha)$ is the index of the geo-domain given at u_α .

Iteration

- (I) Select a data location u_α , regularly or randomly.
- (II) Compute the distribution of $G(u_\alpha)$ conditional to the other data $\{G(u_\alpha) : \beta \neq \alpha\}$. In the stationary case, this is a Gaussian distribution, with mean equal to the simple kriging estimation of $G(u_\alpha)$ and covariance matrix equals to the covariance matrix of the simple kriging errors.
- (III) Simulate a vector Z_α according to the previous conditional distribution.
- (IV) If Z_α is in compliant to the geo-domain dominating at u_α ($Z_\alpha \in D_{i(\alpha)}$), substitute the existing value of $G(u_\alpha)$ by Z_α .
- (V) Go back to a) and loop many times.

The Gibbs sampler presented is reversible, aperiodic, and irreducible. In a nutshell, the distribution of simulated values through the sample points converges to the conditional distribution of the target Gaussian random fields if the number of iterations increases significantly.

Step 5: Simulating Values at Target Grid Nodes

In this step, the simulation should be implemented to simulate the Gaussian random fields at the grid nodes, conditionally to the values acquired in the Step 4 and the same variogram model inferred in Step 3. For this, any algorithm can be applied for simulation of the Gaussian random fields: Cholesky decomposition, sequential Gaussian simulation, discrete spectral simulation, continuous spectral simulation, and turning bands (Chilès and Delfiner 2012, and references therein). Then, the simulated Gaussian values at target grid nodes should be converted back into the geo-domains using the preidentified flag.

Applications in More Detail

To illustrate the applicability of the plurigaussian simulations in a practical example, a small case study is presented in this section. This exemplar is inspired from a siliciclastic platform, for which it is composed of Sand, Shale Sand (ShSand), Shale (Sh), Limestone (Lst), and argillite limestone (argLst). For this case study, 100 sample points are irregularly available in an area and each facies is quantified at each sample point (Fig. 2). Based on the geological interpretation, Sand is in touch with ShSand and ShSand itself is in contact with Sh, whereas Sh is in contact with both Lst and argLst. This gradual change of facies from Sand to Lst and argLst is compatible with the sedimentology of the geological setting for this platform. All these knowledge contribute to identifying a flag as identified in Fig. 2. The variogram model of Gaussian random fields is obtained through an iteration over experimental variogram of indicators. Therefore, one cubic variogram with range of 100 m is inferred for both Gaussian random fields. The simulation results (two scenarios) are also shown in this figure. In total, 100 scenarios are generated and the probability map of each facies is calculated. These maps are applicable for quantification of uncertainty. In the end, one can obtain the most probable map as illustrated.

Current Investigations

There are several investigations of plurigaussian simulations, however, in this section; we only highlight some important developments of this simulation algorithm. Standard method of plurigaussian simulation (Emery 2007) generally is restricted to model few geo-domains and the Gaussian

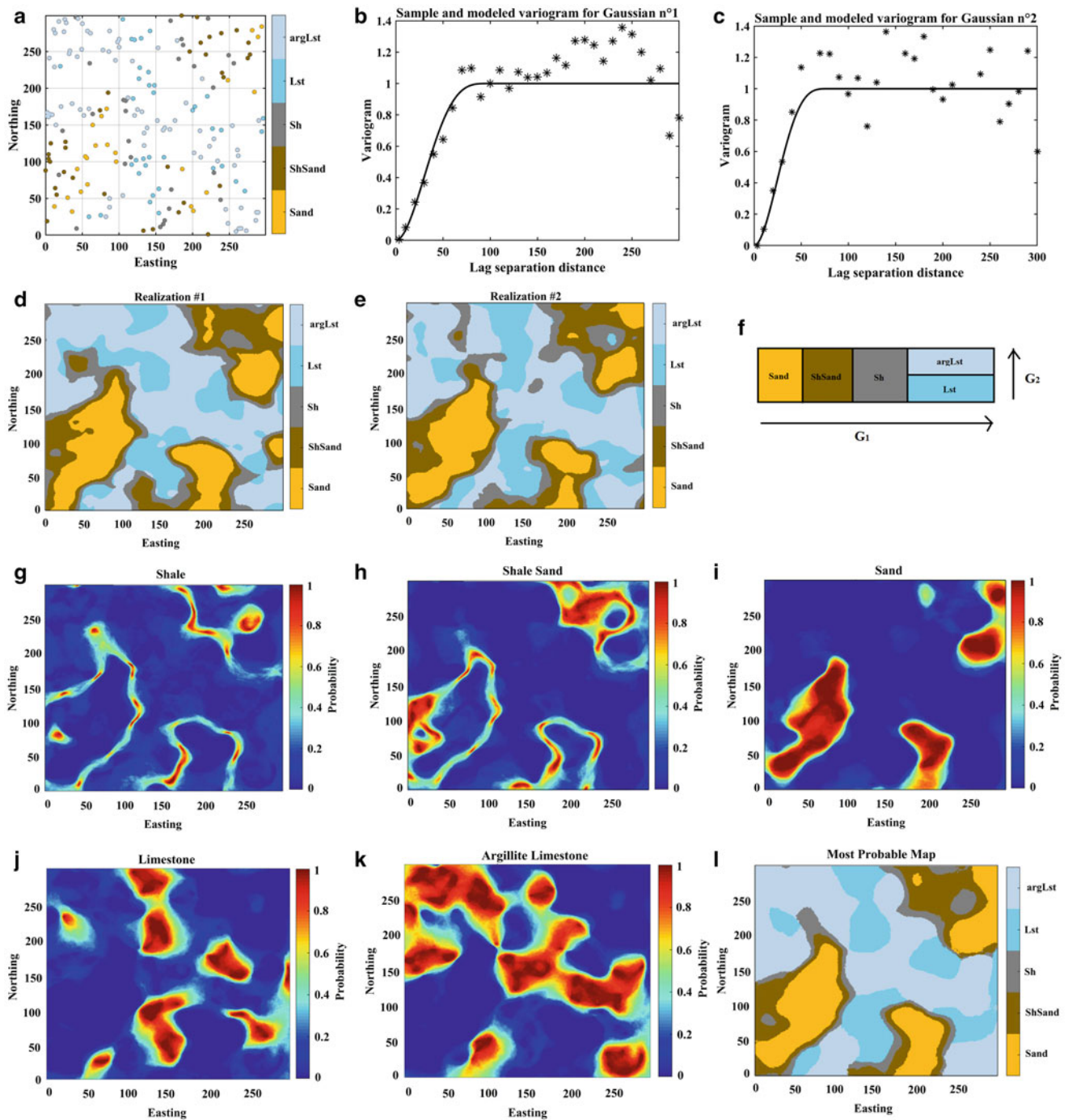
random fields. Madani and Emery (2017) provided a generalization of the conventional plurigaussian simulation that allows the number of Gaussian random fields to be increased, appropriate for the chronological ordering of geological settings in such a way that younger geo-domains cross-cut the older ones. In such a case, all the geo-domains must be in contact altogether where modeling forbidden contact is not feasible. Maleki et al. (2016) introduced another exercise of this method, for which the geo-domains are, in the same manner, ordered hierarchically, but the forbidden contacts can also be dictated into the process of modeling. All these techniques are on the basis of stationary phenomena or homogeneous variability. Nevertheless, in the case of heterogeneous characteristics of geo-domains, one needs to renovate the stationary assumptions (Madani and Emery 2017) or utilize secondary information in replicating the possible trend for the corresponding geo-domains entire the region. For instance, an accessible information may be the seismic data. Another generalization of plurigaussian simulation is related to modeling the cyclic and rhythmic facies architectures (Le Blévec et al. 2018). The developed plurigaussian model takes into account the dampened hole-effect variogram model to delineate the rhythmicity in the vertical direction and a different stationary variogram model in horizontal plane. Plurigaussian simulations can also be integrated with multi-Gaussian simulation for establishing a modeling technique between geo-domains and continuous variables. Emery and Silva (2009) developed a joint simulation algorithm, where one categorical variable and one continuous variable can be modeled altogether given that the variability of the latter across the boundary of former is soft and there is a good association between these two variables. In this case, it is possible to consider several permissible and forbidden contacts among the geo-domains.

Controversies and Gaps in Current Knowledge

Despite the rapid development of plurigaussian simulations in the last decades, there remain some avenues for further research for modeling the geo-domains by this technique.

Deterministic Evaluation of Truncation Threshold

One of the main concern in plurigaussian simulations is identification of truncation thresholds based on the available data at sampling points. There are some approaches to define the declustering proportions to make it representative; however, fitting variogram step, Gibbs sampler and final truncation maps highly depend on the predetermined truncation threshold. Therefore, misspecification of these values leads to producing a huge bias in the results. Incorporation of uncertainty for truncation threshold in the process of modeling needs investigation.



Plurigaussian Simulations, Fig. 2 Application of plurigaussian simulation for modeling the facies in a siliciclastic platform

Variogram Analysis

In the case of having complicated contact relationship between geo-domains, manual variogram fitting is cumbersome. This is more difficult when the aim is to use joint simulation for incorporation of continuous variable as proposed in Emery and Silva (2009). An automatic technique or

semiautomatic technique may facilitate this issue. Another issue is related to lack of data. Whenever some geo-domains are scarce, modeling the variogram and inference its corresponding Gaussian variogram is not trivial. More work is needed to solve the issue of variogram analysis in the presence of scarce data.

Heterogeneity Characteristics of Geo-Domains

Nonstationary representation of geo-domains in a geological complex motivates one using the proportion curves (Armstrong et al., 2011). A common exemplar comes up when the proportions of the geo-domain of interest change vertically and laterally. Spatially varying proportions may be connected with spatially varying thresholds. Taking the advantage of spatially varying geo-domain proportions is disputing for the variogram inference and the computation of spatially varying truncation thresholds itself. More research is required to solve the issue of nonstationary phenomena of geo-domains without considering the spatially varying proportion curves.

Convergence of Gibbs Sampler

In the case of simulating plenty of sample points, a moving neighborhood often is used for solving the kriging system, based on the screen-effect approximation. However, this means that the Markov Chain may no longer converge to the required distribution or does not converge at all. More investigation is required to solve the issue of computational resources for using unique neighborhood in Gibbs sampler instead of moving neighborhood in the case when the data are numerous.

Complex Geological Features

Plurigaussian simulations still are suboptimal for modeling some complex geological features such as connectivity constraints, long-range geological structures, and coniform shapes (e.g., intrusive bodies, metamorphic geological process, infiltration by gravity, or deposition).

Cosimulation

Plurigaussian simulations still are restricted to modeling of one categorical variable. However, this variable may contain only a few geo-domains. More investigation is required for establishing a cosimulation algorithm for modeling two or more categorical variables. For instance, sometimes, rock types and alteration show good association in a deposit. The idea of further research can use of a cosimulation algorithm to reproduce such an association. Concerning this, one possible solution can be multivariate categorical modeling with hierarchical truncated simulation (Silva and Deutsch 2019).

Joint Simulation

Using plurigaussian simulations for joint simulation as proposed by Emery and Silva (2009) is restricted to only one categorical variable and one continuous variable. More investigation is needed for developing an approach for joint simulation of more than one continuous variable, while the categories remain to either one or several variables. Another avenue of research is related to the hierarchical simulation

algorithm for identifying truncation rule while there exist one or more geo-domains in this joint technique.

Summary and Conclusion

Plurigaussian simulations are advanced techniques for modeling the geo-domains with complexity in contact relationship. Flexibility of identifying the truncation rule (flag) and incorporation of geological perspectives lead to producing more realistic results compared to other simulation techniques. In this regard, particularly, forbidden contacts can be imposed into the process of modeling in the resulting maps. For implementation of this algorithm, five main steps were discussed: (1) choosing the truncated rule (flag), (2) estimating the truncation threshold(s), (3) inferring the variogram of the underlying Gaussian random fields, (4) Generating Gaussian random fields at sampling points, (5) simulating values at target grid nodes. Plurigaussian simulations can be implemented in commercial software such as ISATIS and in Petrel through PGS plug-in that is available on Schlumberger's Ocean Store. An open access MATLAB code is also available in IAMG website (iamg.org) attached to the work of Emery (2007). Besides many interesting solutions for modeling the complex geological features in this technique, there still exist several opportunities for further improvement. Few of them were discussed accordingly.

Bibliography

- Armstrong M, Galli A, Beucher H, Le Loc'h G, Renard D, Doligez B, Eschard R, Geffroy F (2011) Plurigaussian simulations in geosciences. Springer, Berlin
- Blévec TL, Dubrule O, John CM, Hampson GJ (2018) Geostatistical modelling of cyclic and rhythmic facies architectures. *Math Geosci* 50(6):609–637. <https://doi.org/10.1007/s11004-018-9737-y>
- Casella G, George EI (1992) Explaining the Gibbs sampler. *Am Stat* 46(3):167–174
- Chilès JP, Delfiner P (2012) Geostatistics: modeling spatial uncertainty. Wiley, New York
- Emery X (2007) Simulation of geological domains using the plurigaussian model: new developments and computer programs. *Comput Geosci* 33(9):1189–1201
- Emery X, Silva DA (2009) Conditional co-simulation of continuous and categorical variables for geostatistical applications. *Comput Geosci* 35(6):1234–1246
- Galli A, Beucher H, Le Loc'h G, Doligez B (1994) HeresimGroup. The pros and cons of the truncated gaussian method. *Geostatistical Simulations*, Kluwer, pp 217–233
- Madani N, Emery X (2017) Plurigaussian modeling of geological domains based on the truncation of non-stationary Gaussian random fields. *Stoch Environ Res Risk Assess* 31:893–913
- Maleki M, Emery X, Cáceres A, Ribeiro D, Cunha E (2016) Quantifying the uncertainty in the spatial layout of rock type domains in an iron ore deposit. *Comput Geosci* 20:1013–1028. <https://doi.org/10.1007/s10596-016-9574-3>

- Matheron G, Beucher H, de Fouquet C, Galli A, Guerillot D, Ravenne C (1987) Conditional simulation of the geometry of fluvio deltaic reservoirs. In: SPE 1987 annual technical conference and exhibition. SPE, Dallas, pp 591–599
- Silva D, Deutsch CV (2019) Multivariate categorical modeling with hierarchical truncated pluri-gaussian simulation. *Math Geosci* 51(1):527–552

Point Pattern Statistics

Dietrich Stoyan
Institut für Stochastik, TU Bergakademie Freiberg, Freiberg,
Germany

Synonyms

Point process statistics

Definition

In the geosciences irregular, random point patterns are studied, in 3D and 2D space. The “points” usually stand for objects of the same type, reducing the information on the objects to their locations; often the points are midpoints or small objects of the same size as the data grid cell. Additional information can be attached by so-called “marks,” which can be real numbers (characterizing perhaps size or type) or geometrical objects (pores or geological bodies). The mathematical models of point patterns are called *point processes*.

Overview

In geoscientific applications, the points may be positions of volcanoes, small deposits of ore, sinkhole centers, trap sites of precious stone placer deposits, centers of pores or crystals, etc. The corresponding patterns occur on all scales, from microscopic to plate-tectonic scale. The goals of point pattern statistics are for a concrete pattern:

- Classification of its type (clustering or regular)
- Understanding interaction between the points
- Determination of range of correlation
- Detection of correlation to covariables
- Fitting a model that helps to understand the process that formed the pattern

Typical point patterns in the geosciences are clustered and cannot be considered as stationary, a standard assumption of point process statistics. “Stationarity” means that a given

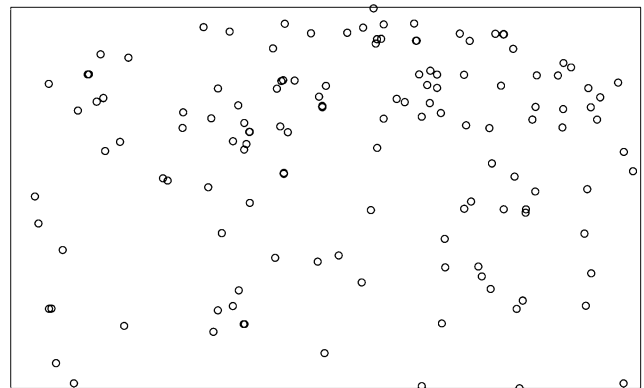
pattern can be considered as a part of a large pattern with similar random fluctuations everywhere in space. However, if one only aims consideration of short-range properties, clever choice of the window of observation gives sense to application of stationary methods, which is justified by Georges Matheron’s idea of “local ergodicity.”

Standard references to point processes and their statistics are Baddeley, Rubak and Turner (2016), and Illian et al. (2008). All formulas mentioned in this article can be found in these books. For statistical analyses of point patterns, the statistical R package *spatstat* is recommended. In the present short text, the theory for spatio-temporal point patterns is not given, as it is applicable to the concrete example considered, where the points are centers of sinkholes (see Fig. 1).

Point Process Theory

General

Many geometrical structures in the geosciences can be described by point patterns, where randomly distributed objects are described by points. This is done in the two- and three-dimensional case, while the present text considers mainly the 2D case. Sometimes the objects are additionally characterized by marks (e.g., numbers or vectors), which leads to marked point patterns. If the marks are bodies, one speaks of germ-grain processes or object-models. Clearly, a point pattern is not a regionalized variable or random field as studied in classical geostatistics, since it is given only in the isolated points that constitute it; also with real-valued marks,



Point Pattern Statistics, Fig. 1 Point pattern of 134 sinkhole centers in the region of the city Tampa. This sinkhole pattern from January 2020 in a rectangle of 19.0 km × 11.5 km in the region of the city of Tampa is a small subpattern of the large and spatially inhomogeneous pattern of all sinkholes in Florida, located at the South-West edge of a large cluster of sinkholes. It was selected as a nearly homogeneous pattern, with the aim to study possible short-range interaction of sinkholes in a relatively homogeneous geological situation with reduced influence of covariables such as population density (and thus of observation intensity). (Source: Florida Geological Survey publications, http://fddeploc.dep.state.fl.us/geodb_query/PubIndexQuery.asp)

it is not a regionalized variable. However, it is possible to transform a point pattern to such a variable by smoothing, see below.

The corresponding stochastic models are called *point processes* and *marked point processes*. It is always assumed that the structures are locally finite, that the number of points in bounded sets is finite. The corresponding mathematical theory is presented in Chiu et al. (2013) and Illian et al. (2008). It studies the probability laws of point processes, the interactions of points, and the description by statistical summary characteristics and many point process models. The present text concentrates on the facts necessary for first statistical steps.

Statistical methods for stationary point processes can be applied for the investigation of short-range interaction of the points, for the study of relationships at distances in the order of first neighbor distances. For this purpose, the window of observation W of observation is adapted to the point pattern so that it looks homogeneously distributed within W . For example, it may make sense to consider points in the interior of a single cluster, while the analysis of a total single cluster with much empty space around is questionable, see Illian et al. (2008, p. 264 and 268).

Intensity Function and Intensity

As the first step of a statistical analysis of a point pattern its *intensity function*, the spatially variable point density, $\lambda(x)$, should be determined, where x is a deterministic point that varies in the window of observation W . Mathematically $\lambda(x)$ yields the mean of the number of points $N(A)$ in a set A by integration,

$$E(N(A)) = \int_A \lambda(x) dx. \quad (1)$$

The intensity function can be estimated by smoothing (see below). Figure 2 shows the estimated intensity for the sinkhole pattern, with a simple parametric form.

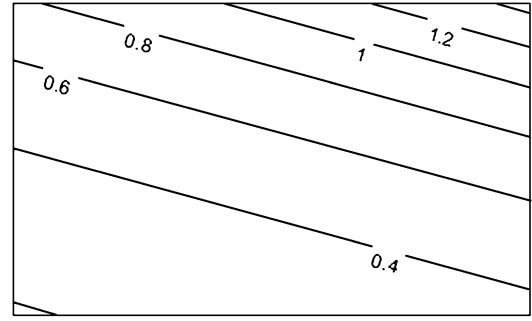
In geological applications also relations to covariables help in the description of point density, for example, the distance to faults or relations to curvature or strain, see Eq. (7).

In the stationary case, $\lambda(x)$ is constant, called *intensity* and denoted by λ . Equation (1) simplifies to

$$E(N(A)) = \lambda \cdot v(A), \quad (2)$$

where $v(A)$ denotes the area of A . Clearly, λ is estimated statistically by

$$\hat{\lambda} = N(A)/v(A). \quad (3)$$



Point Pattern Statistics, Fig. 2 Empirical intensity function of the sinkhole pattern. According to the positions of the sinkholes at the South-West edge of a large cluster of sinkholes, the isolines of $\lambda(x)$ were chosen as lines in SE-NW direction, that is, $\lambda(x)$ is estimated in a parametrized way, not by smoothing. The intensity is measured in number per km^2 .

Pair Correlation Function and Further Summary Characteristics

The variability of point patterns is characterized by various summary characteristics. In general, the most powerful is the *pair correlation function* $g(r)$. It plays the role of Ripley's *K-function* $K(r)$ in the older literature and may be seen as an analogue of the variogram of geostatistics.

A short explanation is as follows. The term $\lambda K(r)$ can be interpreted as the mean number of points in a disc of radius r centered at a typical point of the pattern, not counting the disc center. Then the pair correlation function is given by

$$g(r) = \frac{dK(r)}{dr} / (2\pi r) \text{ for } r \geq 0. \quad (4)$$

Note that the intensity λ is eliminated. In the context of $g(r)$, the variable r should be interpreted as inter-point distance. Always $g(r)$ is nonnegative. If the point distribution is completely random, then

$$g(r) \equiv 1. \quad (5)$$

For large r the function $g(r)$ always takes on the value 1. If there is a finite distance r_{corr} with

$$g(r) = 1 \text{ for } r \geq r_{\text{corr}}$$

then r_{corr} is called *range of correlation*. This means that there are no correlations between point positions at distances larger than r_{corr} . And if there is a minimum inter-point distance r_0 , sometimes called *hard-core distance*, then

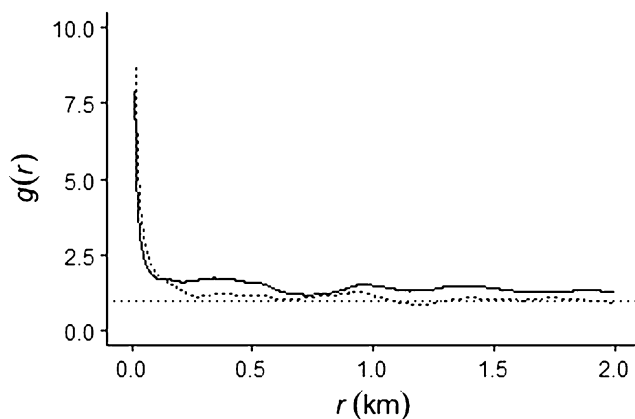
$$g(r) = 0 \text{ for } r \leq r_0.$$

The estimation of the pair correlation function is described in Baddeley et al. (2016) and Illian et al. (2008). Corresponding to the nature of $g(r)$ estimators of λ are used

for normalization. It should be determined only for large point patterns of (say) more than 100 points. The interpretation of empirical pair correlation functions is an art, which needs experience (Illian et al. 2008).

If the pattern analyzed deviates from the behavior expected for a sample of a stationary point process, the estimated $g(r)$ can for large r significantly deviate from the theoretical value 1. In such cases, the estimation may be improved by using ideas of the statistics for *second-order intensity-reweighted stationary* point processes. Then the estimators of intensity λ are replaced by estimators of the intensity function $\lambda(x)$ (Baddeley et al. 2016).

Figure 3 shows the empirical pair correlation function $g(r)$ for the sinkhole pattern of Fig. 1 obtained by the stationary and nonstationary approach. The curves have the typical form of pair correlation functions for cluster point processes. They suggest a cluster diameter between 0.2 and 0.5 km. There are



Point Pattern Statistics, Fig. 3 Empirical pair correlation functions. Empirical pair correlation functions for the pattern of sinkhole centers in Fig. 1, estimate ignoring nonstationarity (—) and estimate with intensity-reweighting (---) in comparison to the value of 1 for the CSR case (...). It is typical that the curve for the stationary approach is above that for the nonstationary approach. See the text for more information

nine point pairs of an inter-point distance smaller than 40 m. (This is very well related to a localized intense rain event.) The closest has a distance of 13 m, and the corresponding events happened in 6 days distance. Another pair of distance 40 m appeared in 2 days. Thus, there is strong clustering in space and time. The large values of $g(r)$ for small r indicate clustering also in this selected subpattern. The pattern may consist of many micro clusters, and these microclusters seem to be randomly distributed; they often are simply close point pairs. Any prediction based only on the pattern is difficult.

Point process statistics has many other summary characteristics, which may be considered as belonging to the class of multiple-point statistics. The most popular ones are:

- Distribution function $D(r)$ of the *distance to the nearest neighbor* from a typical point of the process
- *Spherical contact distribution function* $H_s(r)$, that is, distribution function of the distance from a random test point to the nearest point of the pattern

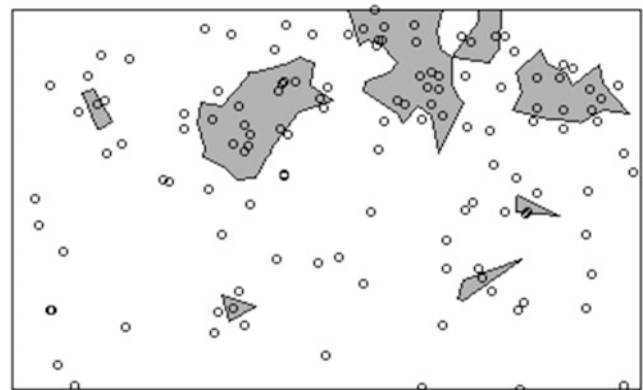
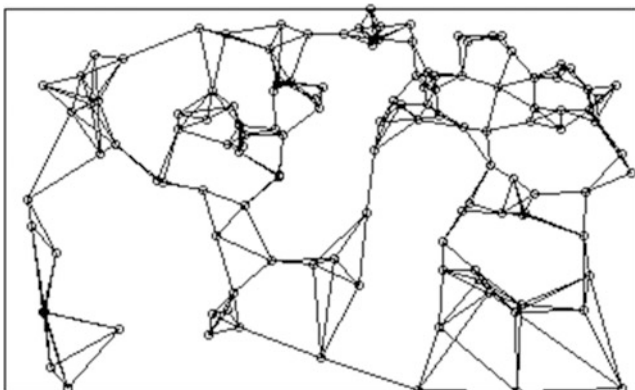
Many others can be found in Baddeley et al. (2016) and Illian et al. (2008). For the determination of directionalities in point patterns, for example, in the context of strain analysis, the Fry method is used (Rajala et al. 2018).

For the sinkhole pattern, Fig. 4 shows two summary characteristics, which aim to show gaps or regions of high point density:

- The 4-neighborhood graph, see Illian et al. (2008, p. 258)
- The Allard-Fraley estimate of regions of high point density, see Baddeley et al. (2016, p. 192).

Marked Point Processes

To each point x_n of a point process, a further datum m_n , called *mark*, can be added, such as a real number (perhaps tonnage



Point Pattern Statistics, Fig. 4 4-neighborhood graph and high-intensity regions. (left) Each sinkhole center is connected with its four nearest neighbors, which shows gaps in the point pattern. (right) High-intensity regions

of the ore deposit at x_n), a surface (perhaps a fault centered at x_n), or a three-dimensional object centered at x_n , see the article *Stochastic Geometry in the Geosciences*.

Stationarity of marked point processes means invariance of the distribution under translations which shift the points but retain the marks, that is, $[x, m] \rightarrow [x + \delta, m]$.

The intensity of a stationary marked point process is the intensity of the point process with the marks stripped off. The marks are characterized by the mark distribution, which in the case of real-valued marks is simply given by a distribution function. The corresponding mean is called the mean mark.

For the case of real-valued marks, stochastic models are available in the literature, for which formulas exist for model characteristics. Two simple cases are (a) independent marks and (b) marks that come from an independent random field $\{Z(x)\}$, that is, the mark of the point x_n is $Z(x_n)$. In the statistics of marked point processes summary characteristics called *mark-correlation functions* are of value, that is, functions that describe the spatial variability of the marks relative to the points.

An example for the application of these characteristics is Ketcham et al. (2005), which studies three-dimensional porphyroblastic textures. Figure 5 shows a volume rendering of a slice of the material obtained using computed tomography. For statistical analysis, such samples were converted to marked point patterns, where crystal centers = points and crystal volumes = marks.

Point process statistics for these data suggests distributional order or regularity at length scales less than the mean nearest-neighbor distance and clustering at great length scales. Crystals close together tend to be small, even after

accounting for the physical limitation that one crystal cannot nucleate within another. In another microstructure application, the points could be pore centers and the marks pore volumes.

Point Process Models

The Homogeneous Poisson Process, CSR

The theory of point processes has a standard “null model”: the *homogeneous Poisson process*. In its context statisticians often speak about *complete spatial randomness* (CSR). Between the points of a homogeneous Poisson process, there is no interaction, and it can be assumed that the points take their locations independently. The point numbers in disjoint sets are independent; this holds true not only for two sets but for any number of sets.

The homogeneous Poisson process is a stationary and isotropic point process. Its only parameter is the intensity λ . The name “Poisson process” comes from the fact that the number $N(A)$ of points in any set A has a Poisson distribution with mean $\mu = \lambda \cdot v(A)$, which means

$$P(N(A) = i) = \frac{\mu^i}{i!} \exp(-\mu) \text{ for } i = 0, 1, \dots \quad (6)$$

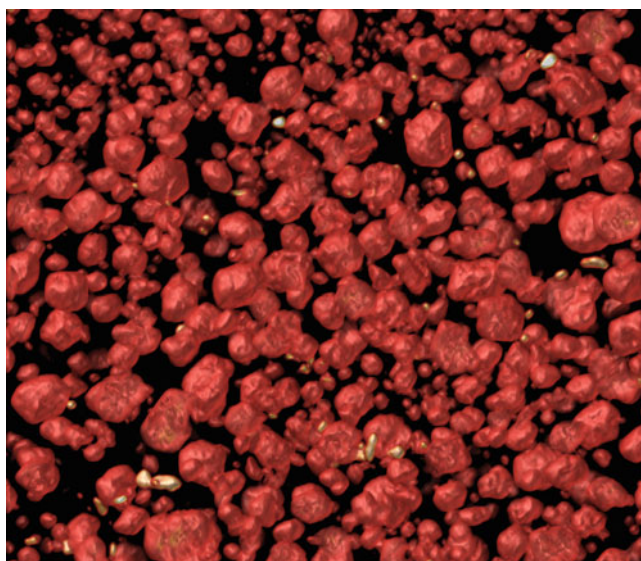
For the Poisson process, there exist many formulas for its distributional characteristics (Baddeley et al. 2016; Illian et al. 2008). In particular, the pair correlation function $g(r)$ has the simple form of Eq. (5).

When a statistician is able to prove by a test that a given point pattern can be seen as CSR, that is, as a sample of a homogeneous Poisson process, then any search for interaction between the points can stop, and any prediction is senseless. (That’s why the term “null model” is used.) Therefore, there are many tests of the CSR hypothesis, see Baddeley et al. (2016), Illian et al. (2008), and Myllymäki et al. (2017). Many of these consider an empirical summary characteristic for the pattern and analogous characteristics obtained by simulating CSR patterns, rejecting the CSR hypothesis if the empirical characteristic differs significantly from the simulated ones. For the pattern of sinkholes, the test using the pair correlation function leads to rejection of the CSR hypothesis. The values of $g(r)$ for r between 0.2 and 0.4 km are too large.

For the homogeneous Poisson process, the nearest neighbor distance distribution function $D(r)$ and the spherical contact distribution function $H_s(r)$ coincide and are equal to $1 - \exp(-\lambda r^2)$.

The Inhomogeneous Poisson Process and the Cox Process

The inhomogeneous Poisson process shares the most properties with the homogeneous one, only there is a spatially variable intensity function $\lambda(x)$. Care is necessary to



Point Pattern Statistics, Fig. 5 A porphyroblastic structure. Garnet crystals (red) and higher-density trace phases (such as zircon and iron oxides) (yellow) in a sample of garnet amphibolite schist. The width of the object shown is 25.5 mm and the thickness 3.1 mm. (Data courtesy of Richard Ketcham)

understand the main point: Of course, every point pattern has a randomly fluctuating point density, even a pattern for a stationary point process. However, in the case of the inhomogeneous Poisson process, the point density fluctuates in a predictable sense: If one could observe more than one sample, one would observe similar point densities in the same regions, for example, high point density along a geological fault; nevertheless the particular point locations may be fully random.

The point numbers in disjoint sets are independent and the number $N(A)$ of points in any set A has a Poisson distribution, the mean of which depends both on size and location of A . It is given by Eq. (1).

An example of application of an inhomogeneous Poisson process is the study of spatial correlation between gold deposits and geological faults in Baddeley (2018), where an intensity of the form

$$\lambda(x) = \exp(\alpha + \beta d(x)) \quad (7)$$

was used, where $d(x)$ is the nearest distance from x to a fault.

The inhomogeneous Poisson process can be made still more random: also the intensity function may be random. The resulting point processes are called *Cox processes* (after the English statistician David R. Cox) or “doubly stochastic Poisson processes” because of the two stochastic elements in the model. In this case, the intensity function is a random field: if this field is stationary then the Cox process is stationary too.

Of course, this field has to be positive, as an intensity function. Therefore, a Gaussian random field is not suitable in this context. A natural solution is to start from a Gaussian field $\{Z(x)\}$ and then to use the exponential $\lambda(x) = \exp(Z(x))$. This leads to so-called *log-Gaussian Cox processes*, which have been widely used in the modeling of environmental heterogeneity. Because of their construction, they can be seen as a link between point process statistics and geostatistics. They are simple enough to derive formulas for their statistical characteristics. For example, intensity and pair correlation function of a stationary and isotropic log-Gaussian Cox process are given by

$$\lambda = \exp\left(\mu_Z + \frac{1}{2}\sigma_Z^2\right) \quad (8)$$

and

$$g(r) = \exp(k_Z(r)) \text{ for } r \geq 0, \quad (9)$$

where μ_Z is the mean, σ_Z^2 the variance, and $k_Z(r)$ the covariance function of the Gaussian field $Z(x)$.

Cluster and Regular Point Processes

Some Cox processes are cluster point processes. Typical examples are Poisson cluster processes. These can be thought to be constructed starting from a homogeneous Poisson process. The points of this process, sometimes called “mother points,” are cluster centers (not points of the final process), and around each of these points “daughter points” are randomly scattered. The union of all daughter points is the cluster process. Formulas for its distributional characteristics can be found in Baddeley et al. (2016), and Illian et al. (2008).

Cluster point patterns appear frequently in the geosciences. A famous case are spatio-temporal patterns of epicenters of earthquakes, which are clustered in space and time (Ogata 2017). Of course, also the sinkhole pattern can be considered as a sample of a cluster process.

The pair correlation function of a cluster point process may have a pole at $r = 0$. The existence of a pole may indicate a fractal behavior of the structure related to the points (Agterberg 2014). However, poles can appear also in situations without any fractal structure, for example, when daughter points are scattered on segments, perhaps close to faults.

While the simulation of cluster processes is easy, their statistics present a challenge (Baddeley et al. 2016; Illian et al. 2008).

Another class of point processes are *regular point processes*, for example, “hard-core processes” with a positive minimum inter-point distance. Such a model would be used for the porphyroblastic structure in Fig. 5. “Gibbs processes” as mentioned below in the context of Markov Chain Monte Carlo are a popular class of such models.

Tools of Point Pattern Statistics

Smoothing

A regionalized variable is a quantity $Z(x)$ defined for every location x of space. Examples are porosity at x (either 0 or 1) or ore content at x (a nonnegative number, being the ore content of a sample (ball or cube) centered at x). The statistics of regionalized variables are statistics of random fields and sometimes this just goes under the name *geostatistics*. Point processes are not regionalized variables, but sometimes it makes sense to regionalize such structures, in order then to apply the strong methods of geostatistics. A natural approach is smoothing. For example, consider the number of points $Z(x)$ of a planar point process in the disc $b(x, R)$ divided by πR^2 for every point x of the space for some chosen radius R . For inhomogeneous point processes, this leads to estimates of the intensity function $\lambda(x)$. A problem is then the choice of the smoothing parameter R .

Simulation Techniques

Point processes are simulated in order to visualize the structure of some models, to generate samples for the investigation of statistical procedures, so-called test-patterns, or to demonstrate uncertainty. Here some important techniques are described, limited to the stationary case.

Simple Simulations

The simulation of many point processes models is a simple task. Here are two pseudocodes for the first steps, where *random* is a random number uniform on $[0, 1]$:

Simulation of a random point $x = (\xi, \eta)$ in a rectangle of side lengths a and b .

1. $\xi = \text{random} \cdot a$
2. $\eta = \text{random} \cdot b$
3. print $x = (\xi, \eta)$

Simulation of a sample of a homogeneous Poisson process of intensity λ in a rectangle.

1. generate a Poisson random number n of mean $\lambda \cdot a \cdot b$
2. for $i = 1$ to n generate random points x_i in the rectangle
3. print $(x_1, \dots, x_n)$

An elegant form of simulation of a Poisson random number is described in Illian et al. (2008, p. 71). In spatstat, a Poisson process pattern can be generated by the function `rpoispp`.

For more general models, care is necessary with the boundaries, for example, when simulating cluster processes: then mother points outside the window must not be ignored. Here the technique of so-called “exact simulation” can be used.

Markov Chain Monte Carlo

In particular, point processes with inhibition between the points, that is, with some degree of regularity, are simulated by the Markov chain Monte Carlo technique. A simple example is the simulation of the hard-core Gibbs process. It goes for a sample of n points in a window of simulation W as follows, if the hard-core or minimum distance between the points is R :

Place n points within W so that the distance between all point pairs is not smaller than R . Take randomly one of the points and give it a new position chosen randomly in the part of W where the hard-core condition is not violated. Continue this procedure many times, until the simulation process comes into a stationary state; in the beginning, there is some burn-in period. This method can be generalized both to the cases of a random number of points and forms of

interaction between the points more subtle than the simple hard interaction.

The sequence of point patterns in W after each step is a Markov chain in the state space of point patterns.

Reconstruction

Sometimes no model is available and one only needs patterns with similar statistical properties as a given empirical pattern, for example as training images. Then the method of reconstruction is recommended, a method in the spirit of multipoint statistics (Illian et al. 2008). Its first step is the determination of some statistical summary characteristics such as pair correlation function $g(r)$ plus(!) distribution function $D(r)$ or $H_s(r)$ for the original pattern. Then a new point pattern is constructed that has similar statistical summary characteristics. The construction starts with a CSR pattern of the same point number as the empirical pattern. Then one of its points is chosen randomly and replaced by a random point in W . If the new pattern has summary characteristics closer to the characteristics of the empirical pattern, the new point is accepted and a new trial is made. If not, a new point is chosen, etc. This iteration is repeated for many steps, where the determination of the summary characteristics of the whole pattern must be well organized. The minimization of the discrepancy may be improved by so-called simulated annealing. Also conditional simulation is possible, which means here simulation of a point pattern in some region given a subpoint pattern.

Summary

Point patterns make sense everywhere in geostatistics as representatives of locations of groups of objects irregularly distributed in space; marked point patterns stand for object systems. In statistical studies of microstructures, there are good chances to apply the powerful methods for stationary point processes. However, in larger scales big problems appear with instationarity and anisotropy. Here further research should change the situation, with adapted finite models and statistics.

Cross-References

- [Geostatistics](#)
- [Multiple Point Statistics](#)
- [Simulated Annealing](#)
- [Simulation](#)
- [Spatiotemporal](#)
- [Stochastic Geometry in the Geosciences](#)

Bibliography

- Agterberg F (2014) Geomathematics: theoretical foundations, applications and future developments. Springer, Cham/Heidelberg/New York/Dordrecht/London
- Baddeley A (2018) A statistical commentary on mineral prospectivity analysis. In: Sagar BSD, Cheng Q, Agterberg F (eds) Handbook of mathematical geosciences. Springer, Cham, pp 25–65
- Baddeley A, Rubak E, Turner R (2016) Spatial point patterns. Methodology and applications with R. CRC Press, Boca Raton
- Chiu SN, Stoyan D, Kendall WS, Mecke J (2013) Stochastic geometry and its applications. Wiley, Chichester
- Illian J, Penttinen A, Stoyan H, Stoyan D (2008) Statistical analysis and modelling of spatial point patterns. Wiley, Chichester
- Ketcham RA, Meth C, Hirsch DM, Carlson WD (2005) Improved methods for quantitative analysis of three-dimensional porphyroblastic textures. *Geosphere* 1:42–59
- Myllymäki M, Mrkvicka T, Grabarnik P, Seijo H, Hahn U (2017) Global envelope tests for spatial processes. *J R Stat Soc Ser B* 79:381–404
- Ogata Y (2017) Statistics of earthquake activity: models and methods for earthquake predictability studies. *Annu Rev Earth Planet Sci* 45:497–527
- Rajala T, Redenbach C, Särkkä A, Sormani M (2018) A review on anisotropy analysis of spatial point patterns. *Spat Stat* 28:141–168

Polarimetric SAR Data Adaptive Filtering

Mohamed Yahia^{1,2}, Tarig Ali^{1,3}, Md Maruf Mortula³ and Riadh Abdelfattah^{4,5}

¹GIS and Mapping Laboratory, American University of Sharjah United Arab Emirates, Sharjah, UAE

²Laboratoire de recherche Modélisation analyse et commande de systèmes- MACS, Ecole Nationale d'Ingénieurs de Gabes – ENIG, Université de Gabes Tunisia, Zrig Eddakhlania, Tunisia

³Department of Civil Engineering, American University of Sharjah, Sharjah, UAE

⁴COSIM Lab, Higher School of Communications of Tunis, Université of Carthage, Carthage, Tunisia

⁵Département of ITI, Télécom Bretagne, Institut de Télécom, Brest, France

Definition

Remote sensing systems represent nowadays an important source of information for the study of the Earth's surface. Due to their operation in all weather and all-time conditions, synthetic aperture radar (SAR) and polarimetric SAR (PoSAR) sensors are broadly applied in geoscience and remote sensing applications. Nevertheless, due to the coherent nature of the scattering mechanisms, SAR images are contaminated by the speckle noise. The speckle noise reduces the accuracy of post-processing as well as human

interpretation of the images. Hence, speckle filtering forms an important preprocessing task before the extraction of the useful information from SAR images.

Introduction

Adaptive filters and particularly the minimum mean square error (MMSE) filters have been extensively applied in PoLSAR to reduce speckle. Most MMSE-based PoLSAR filters have mainly concentrated on the collection of homogeneous pixels. In the refined Lee filter (Lee et al. 1999), the homogeneous pixels are chosen from eight edge aligned windows. In the Lee sigma filter (Lee et al. 2015), similar pixels are chosen from the sigma range. In Lee et al. (2006), the scattering model-based technique chose pixels of the same scattering media. In Wang et al. (2016), similar pixels are chosen using both intensity information and scattering mechanism. Based on a linear regression of means and variances of the filtered pixels, an infinite number of looks prediction (INLP) technique has been applied to improve the filtering performance of MMSE-based SAR image filtering (Yahia et al. 2020a). In Yahia et al. (2017, 2020b), an iterative MMMSE (IMMSE)-based filter is introduced. The IMMSE is initialized by a polarimetric filter ensuring a high speckle reduction level. Then, to improve the spatial detail preservation, the IMMSE filter is implemented for few iterations.

SAR Polarimetry

PoSAR data can be expressed using the scattering matrix S (Lee et al. 1999)

$$S = \begin{bmatrix} S_{hh} & S_{hv} \\ S_{vh} & S_{vv} \end{bmatrix} \quad (1)$$

The indices v and h represent the vertical and horizontal polarization, respectively. $S_{vh} = S_{hv}$ for reciprocal backscattering. PoSAR data can be expressed also in a vector form

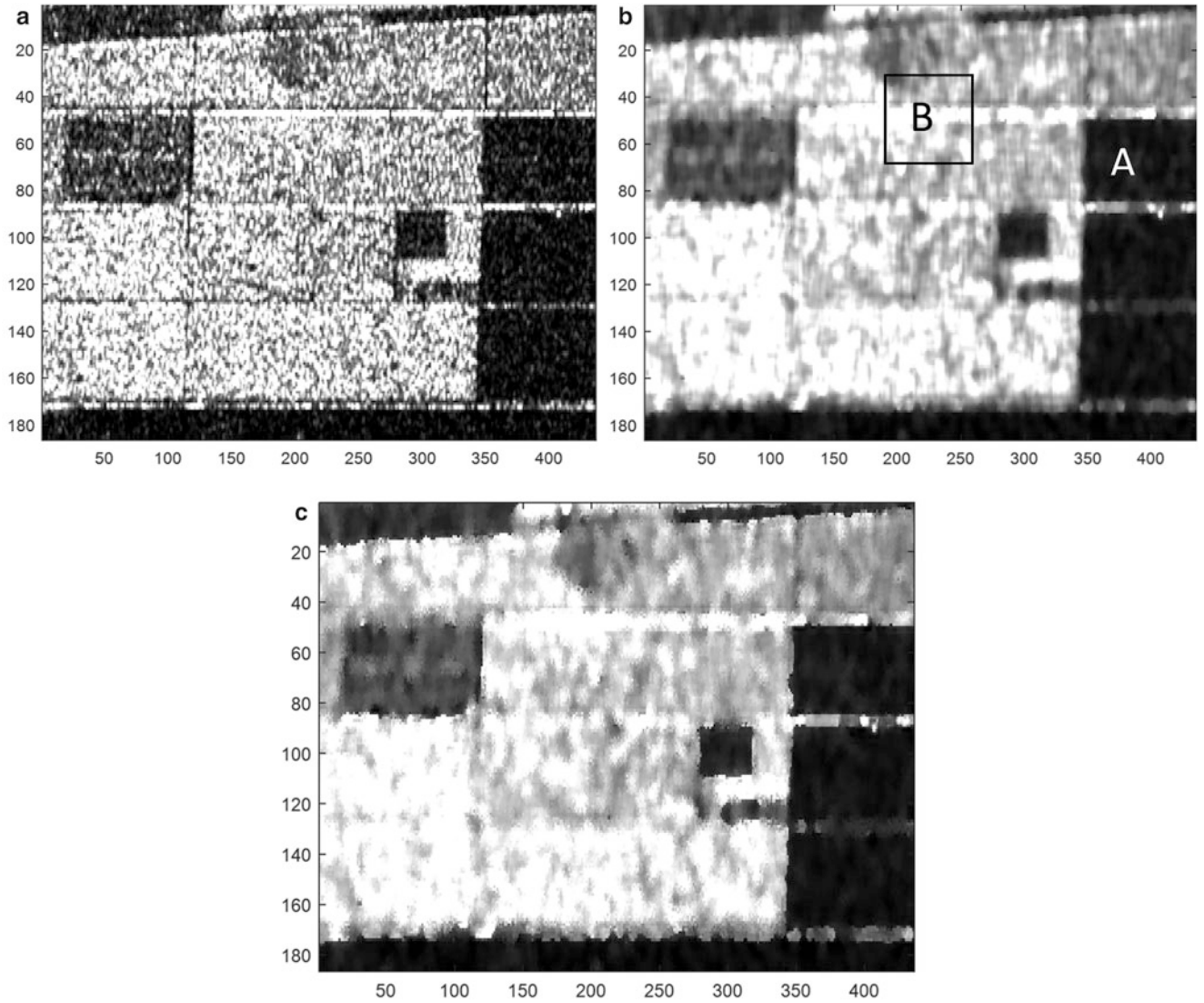
$$\Omega = \begin{bmatrix} S_{hh} & \sqrt{2}S_{hv} & S_{vv} \end{bmatrix}^t \quad (2)$$

where “ t ” represents the transpose operator. The total power of a pixel i.e., $span$ is expressed as (Lee et al. 1999)

$$span = |S_{hh}|^2 + 2|S_{hv}|^2 + |S_{vv}|^2 \quad (3)$$

The covariance matrix C is expressed as (Lee et al. 1999)

$$C = E(\Omega \Omega^{*t}) \quad (4)$$



Polarimetric SAR Data Adaptive Filtering, Fig. 1 Span images: (a) original image, (b) Sigma filter, (c) INLP filter. A and B zones are used to compute ENL (22) and EPD (23), respectively

where “*” is the complex conjugation and $E(.)$ is the mathematical expectation.

Adaptive MMSE PolSAR Filtering

The PolSAR speckle filtering constraints are (Lee et al. 2015):

1. Conserve the polarimetric information.
2. Minimize the speckle noise in homogeneous zones (i.e., average all pixels $\hat{X}(i) = \overline{C}(i)$). Where $\hat{X}(i)$ represents the filtered covariance matrix.
3. Preserve the spatial details. The value of the original

covariance matrix should be maintained (i.e., $\hat{X}(i) = C(i)$).

SAR images are affected by the multiplicative (Lee et al. 1999)

$$y(i, j) = x(i, j) v(i, j), \quad (5)$$

v is the speckle noise with standard deviation σ_v and unit mean. x is the noise free pixel. For convenience, the (i, j) index is eliminated. The MMSE-based filtered pixel $\hat{x}$ is (Lee et al. 1999)

$$\hat{x} = \bar{y} + d(\text{var}(y))(y - \bar{y}) \quad (6)$$

where

$$b = \text{var}(x)/\text{var}(y), \quad (7) \quad b = x \quad (16)$$

Then

$$b = (\text{var}(y) - \bar{y}^2 \sigma_v^2) / (1 + \sigma_v^2) \text{var}(y) \quad (8)$$

$\bar{y}$ and $\text{var}(y)$ are computed adaptively using a local square window W .

In the refined (Lee et al. 1999) and sigma (Lee et al. 2015) filters, the MMSE estimate of the covariance matrix (i.e., $\hat{C}$) is

$$\hat{C} = \bar{C} + b(C - \bar{C}) \quad (9)$$

where b given in (8) is estimated from the span image (3).

Figure 1 displays the original and the filtered span images of Les-Landes PolSAR data, respectively. It can be observed that the sigma filter performed high spatial detail preservation level. However; the speckle reduction is low. The speckle reduction level could be increased by increasing the size of the window size but the spatial detail preservation will be decreased. To increase the performance of the MMSE-based filters, the MMSE formulation has been improved as follows

Infinite Number of Looks Prediction INLP Filter

By combining Eqs. (5) and (6) (Yahia et al. 2017).

$$\bar{\hat{x}} = \bar{y} + \frac{\text{var}(x)}{\text{var}(y)}(y - \bar{y}) = \bar{y} = x, \quad (10)$$

$$\begin{aligned} \text{var}(\hat{x}) &= E\left(\left(\bar{y} + \frac{\text{var}(x)}{\text{var}(y)}(y - \bar{y})\right) - \bar{y}\right)^2 \\ &= \left(\frac{\text{var}(x)}{\text{var}(y)}\right)^2 E(y - \bar{y})^2, \end{aligned} \quad (11)$$

$$\text{var}(\hat{x}) = \left(\frac{\text{var}(x)}{\text{var}(y)}\right) \text{var}(x) = d\text{var}(x), \quad (12)$$

then,

$$\hat{x} = x + \frac{(y - x)}{\text{var}(x)} \text{var}(\hat{x}). \quad (13)$$

Finally,

$$\hat{x} = a \text{var}(\hat{x}) + b, \quad (14)$$

where

$$a = (y - x)/\text{var}(x), \quad (15)$$

and

Equation (14) shows that the filtered pixel $\hat{x}$ is linearly related to its variance $\text{var}(\hat{x})$. This rule is applied to estimate the INLP filtered pixel (i.e., the parameter b) (Yahia et al. 2017).

The linear regression in (x, y) coordinates gives (Vaseghi 2000):

$$b = \bar{y} - a \bar{x}, \quad (17)$$

and

$$a = \sum_{i=1}^M (x_i y_i - M \bar{x} \bar{y}) / MV_x, \quad (18)$$

where V_x and $\bar{x}$ are the variance and the mean of the vector x , respectively. In (18), the parameters x_i and y_i represent the samples $\hat{x}_i$ and $\text{var}(\hat{x}_i)$, respectively. To conserve PolSAR information, it has been suggested that all the elements of the covariance matrices must be filtered equally as the multi-look process (Lee et al. 2015). However, it is observed in (18) that the filtered pixel b is not linearly dependent on y_i due to the square values of y_i . Hence, (17) can be written as:

$$b = \bar{y} \left(1 + \frac{\bar{x}^2}{V_x}\right) - \frac{\bar{x}}{MV_x} \sum_{i=1}^M x_i y_i = P \bar{y} + Q y', \quad (19)$$

where

$$P = \left(1 + (\bar{x}^2/V_x)\right) \quad (20)$$

and Q is a vector enclosing the samples

$$Q_i = (\bar{x}/MV_x) x_i. \quad (21)$$

To make b linearly dependent on y_i , the parameters P and Q must have the same values for all covariance matrix elements. Hence, the parameters P and Q were estimated from the *span* image and introduced in (19) to filter all the elements of the covariance matrix similarly. The real and imaginary coefficients of the off-diagonal elements of the covariance matrix were processed independently.

To assess the performance of the studied speckle filtering techniques quantitatively in terms of speckle reduction level, the equivalent number of looks is selected:

$$\text{ENL} = \hat{x}/\text{var}(\hat{x}) \quad (22)$$

To assess the ability of the filters in terms of spatial detail preservation, the Edge Preservation Degree (EPD) is

employed (Feng et al. 2011). The EPD in horizontal direction is:

$$EPD_H = \frac{\sum_{m,n} |\hat{x}(m,n)/\hat{x}(m,n+1)|}{\sum_{m,n} |y(m,n)/y(m,n+1)|}. \quad (23)$$

Table 1 shows the quantitative filtering results. The INLP filter ensured higher speckle reduction level (i.e., $ENL = 31 > 25$) and better spatial detail preservation ($EPD_{H-INLP} = 0.84126 > EPD_{H-sigma} = 0.8398$) than the 7×7 sigma. Figures 1b and c display the filtered sigma and INLP span images, respectively. Quantitative outcomes are verified visually where the INLP filter ensured better spatial detail preservation than sigma filter. The spatial resolution is significantly increased by the improved INLP filter. The spatial details in Figs. 1c are more contrasted than the one in Figs. 1b.

In addition to the ability of the adaptive MMSE filtering technique to reduce speckle, the conventional adaptive filters such as refined Lee, Lee sigma, are the most computationally efficient methods among other better speckle filtering techniques for huge datasets. In fact, the computational complexity of the PolSAR filter constitutes currently a great concern for practical implementations since for many of the new SAR systems such as TerraSAR-X, Cosmo-SkyMed, and Gaofen-3, the image sizes are huge (i.e., bigger than $10,000 \times 10,000$ pixels).

Unlike the scalar MMSE, in PolSAR MMSE-based filters, the span is used to estimate the weight b . However, it has been

demonstrated in Yahia et al. (2020c) that the span is a non-linear filter and the span statistics are function of the entropy H . Figure 2 displays the ENL of the span image. It can be observed that the statistics of the span change within different extended homogenous areas.

As result, for optimal application of the adaptive MMSE PolSAR filters, σ_v in (8) must be adapted to the statistics of the span image.

Summary and Conclusions

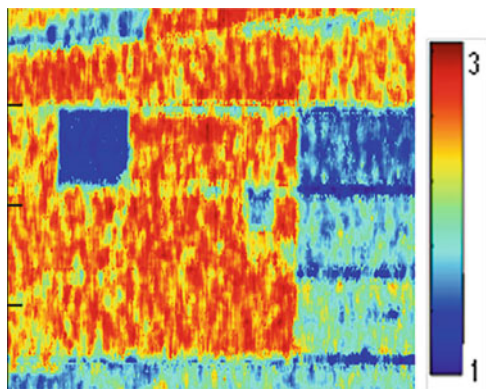
In this study, the adaptive MMSE PolSAR filtering is investigated. The sigma Lee PolSAR filter is particularly implemented. Results show that this filter was able to preserve spatial details with low ability to reduce speckle. To improve the adaptive MMSE filtering performance, the INLP filter is introduced. The filtered pixel was derived using a linear regression of means and variances of the filtered pixels. The linear regression rule has been adapted to PolSAR data in order to preserve the polarimetric information. Results show that the INLP outperformed sigma filter in terms of spatial detail preservation and speckle reduction.

Bibliography

- Feng H, Hou B, Gong M (2011) SAR image despeckling based on local homogeneous-region segmentation by using pixel-relativity measurement. *IEEE Trans Geosci Remote Sens* 49(7):2724–2737
- Lee JS, Grunes MR, De Grandi G (1999) Polarimetric SAR speckle filtering and its implication for classification. *IEEE Trans Geosci Remote Sens* 37(5):2363–2373
- Lee JS, Grunes MR, Schuler DL, Pottier E, Ferro-Famil L (2006) Scattering-model-based speckle filtering of polarimetric SAR data. *IEEE Trans Geosci Remote Sens* 44(1):176–187
- Lee JS, Ainsworth TL, Wang Y, Chen KS (2015) Polarimetric SAR speckle filtering and the extended sigma filter. *IEEE Trans Geosci Remote Sens* 53(3):1150–1160
- Vaseghi SV (2000) Advanced digital signal processing and noise reduction, 4th edn. Wiley, New York, pp 227–254
- Wang Y, Yang J, Li J (2016) Similarity-intensity joint PolSAR speckle filtering. In: International Conference on Radar
- Yahia M, Hamrouni T, Abdelfattah R (2017) Infinite number of looks prediction in SAR filtering by linear regression. *IEEE Geosci Remote Sens Lett* 14(12):2205–2209
- Yahia M, Ali T, Mortula MM, Abdelfattah R, El Mahdi S, Arampola NS (2020a) Enhancement of SAR speckle denoising using the improved IMMSE filter. *IEEE J Select Topics Appl Earth Obs Remote Sens* 13: 859–871
- Yahia M, Ali T, Mortula MM, Abdelfattah R, El Mahdy S (2020b) Polarimetric SAR speckle reduction by hybrid iterative filtering. *IEEE Access* 8(1)
- Yahia M, Ali T, Mortula MM (2020c) Span Statistics And Their Impacts On Polsar applications. *IEEE Geosci Remote Sens Lett*. <https://doi.org/10.1109/LGRS.2020.3039109>

Polarimetric SAR Data Adaptive Filtering, Table 1 Quantitative results

	ENL A zone	EPD _H	EPD _V
Sigma 7×7	25	0.8398	0.8183
INLP ($L = 40$)	31	0.84126	0.8214



Polarimetric SAR Data Adaptive Filtering, Fig. 2 ENL of the span

Pore Structure

Weiren Lin and Nana Kamiya

Earth and Resource System Laboratory, Graduate School of Engineering, Kyoto University, Kyoto, Japan

Synonyms

Interstice; Void

Definition

A pore is a small open space or passageway existing between solid materials in a rock, also called a void or an interstice. Generally, pores primarily form at the same time as the formation of the host rock and/or are secondarily introduced by mechanical failures and/or chemical actions after rock formation. From their morphological features, pores can be classified into isolated pores and connected pores. Usually, the pore structure is described in terms of the geometrical features, including their shape, size, volume, connectivity, and tortuosity, of pore populations rather than an individual pore.

Introduction

Invariably, pores exist in the earth materials from soils in the shallow subsurface to rocks in the crust and at least in the upper mantle. In this chapter, pores and pore structures in rocks will be the focus and will hopefully benefit areas related to solid earth in geoscience and geoen지니어ing. Generally, pores in rocks are filled by groundwater at depths below the groundwater level. In rare cases, however, the pores are partially or fully filled by oil (petroleum) and/or natural gas or solid gas hydrate. More importantly, the connected pores play a function in the fluid flow path, which allows material circulations at local and global scales and affects Earth systems. In addition, pore fluid pressure, especially the over pore fluid pressure in seismogenic zones, possibly influences various phenomena in and around plate interfaces, such as great earthquakes and plate tectonics. Therefore, pore structure is one of the most fundamental and important characteristics of rocks in many areas of study, e.g., geology, geophysics, geochemistry, earth resources, and groundwater hydrology.

Pore structures in different rock types are dependent on their origin and their host rocks. Generally, pores in sedimentary rocks exist in the matrix and have a relatively small aspect ratio (the ratio of their length to width). Most pores in volcanic rocks usually exhibit spherical shapes because they

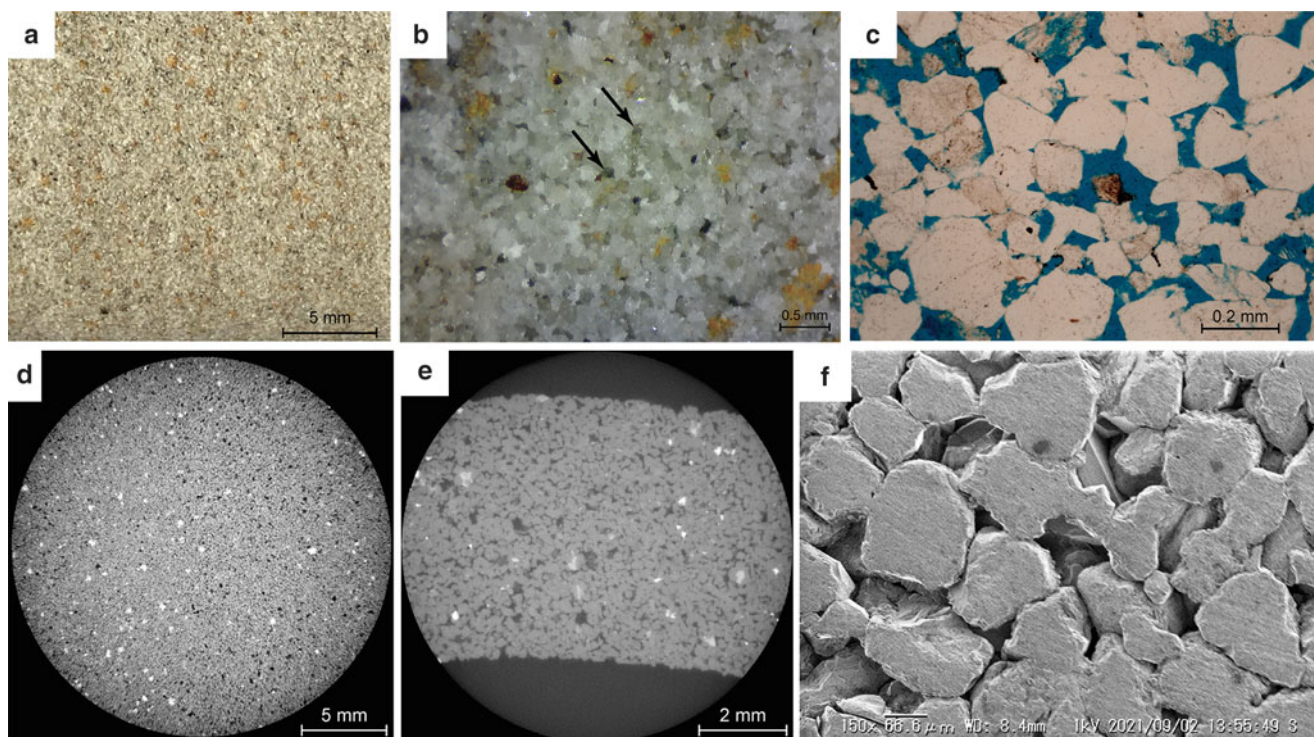
were formed from volcanic gas. On the other hand, pores in crystalline rocks, including plutonic rocks and metamorphic rocks, have larger aspect ratios, e.g., >100 , and are generally called fractures.

Visual Observation of Pore Structure

To reveal the pore structure, visual observation at various scales is the most fundamental approach and has been performed by various methods. In this section, a very popular sedimentary rock called Berea sandstone was used as an example to show several typical images of a pore structure of a rock. This Paleozoic sandstone quarried for more than 200 years in Ohio, USA, has been used widely as a building material and grindstone (Hannibal 2020). Moreover, Berea sandstone with homogeneous fabric and fine-to-medium grain sizes has been used as a rock testing material for laboratory hydrological and mechanical experiments and for the visual observation of pore structures worldwide.

Figure 1a is an image taken by an iPhone (Model: SE) camera, showing that any pore(s) is hard to identify, although mineral grains are visible. Therefore, it may indicate that the pore structure is hard to see with the human naked eye. Several holes (the pores) can be seen from images taken by a stereomicroscope, but the pore structure cannot be clearly identified, probably due to low contrast and the presence of many semitransparent minerals, such as quartz (Fig. 1b). Figure 1c is a polarized optical microscopic photo (open nicol) of a thin section of $\sim 30\ \mu\text{m}$ in thickness. The thin section was made up after fluidic resin dyed blue was impregnated in pores and then fixed thermally (Chen et al. 2018). Therefore, mineral grains and pores with blue color can be easily identified, and the pore structure is clearly visible. From this two-dimensional image, the pores can be considered to connect with each other in the three-dimensional domain, i.e., a well-connected pore network through Berea sandstone exists.

Figure 1d and e shows the images of X-ray CT (computed tomography) at different scales. Since the 1990s, medical X-ray CT with an $\sim 200\text{-}\mu\text{m}$ resolution has been utilized for rock core samples in the geosciences (Cnudde and Boone 2013). The important advantages of X-ray CT are as follows: (i) the internal structure in a rock sample can be imaged without cutting, i.e., nondestructive observation; (ii) the three-dimensional structure can be reconstructed; and (iii) its images are composed of digital data and suitable for numeral simulation and mathematical geosciences. Following the application of medical X-ray CT, micro-CT (high-resolution CT) with a typical resolution of $\sim 10\ \mu\text{m}$ (Fig. 1d and e and recently nano-CT with submicrometer resolutions were developed and applied. From such X-ray CT images, three-dimensional pore network models can be built and then used



Pore Structure, Fig. 1 Images of Berea sandstone with ~20% porosity. (a) A normal optical photo taken by an iPhone (model: SE) camera. (b) An image of stereo microscope, arrows show presence of holes (the pores). (c) An image of thin section taken by a polarized optical microscope under open Nicol condition; the blue color shows pores in where blue resin was impregnated and fixed. (d) A X-ray CT image in a

resolution of ~22 $\mu\text{m}/\text{pixel}$ of a cylindrical sample; the dark gray-black color spots show the pores. (e) A X-ray CT image in a resolution of ~8 $\mu\text{m}/\text{pixel}$ of a rock tip (courtesy to Kazuya Ishitsuka). (f) A SEM image of the sandstone sample surface grinded by fine grit size sandpapers (average particle size <23 μm); WD denotes working distance is an operation parameter of SEM

to simulate permeability and other physical properties (e.g., Takahashi et al. 2016).

Figure 1f is a surface image of a Berea sandstone sample taken via scanning electron microscopy (SEM). This image shows a similar pore structure to the polarized optical microscopic image shown in Fig. 1c. These images (Fig. 1c and f) are clearer than the X-ray CT images (Fig. 1d and e); however, their disadvantage point is two-dimensional. All the images mentioned above were obtained under atmospheric pressure conditions. Generally, pore structure in rocks is dependent on ambient pressure because pressure may cause pores to reduce their volume or size (Takahashi et al. 2016; Lin et al. 2020).

Quantitative Evaluation of Pore Structure

The most important parameter to evaluate pore structure in rocks should be the porosity defined by the following equation:

$$\varphi = \frac{V_{\text{pore}}}{V_{\text{total}}} \quad (1)$$

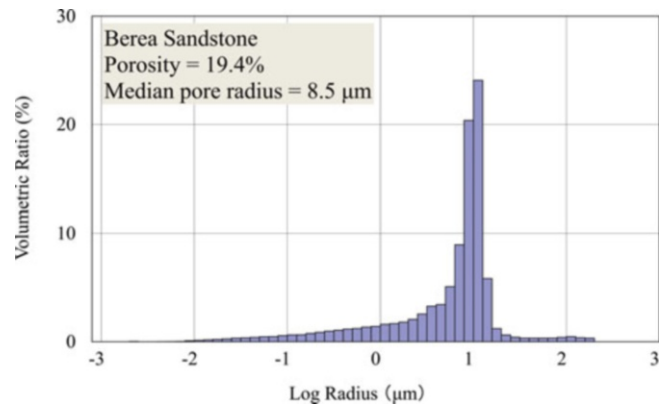
where porosity φ is usually in %, V_{pore} is the cumulative pore volume in a rock sample, and V_{total} is the total volume including both solid materials and pores in the sample. Generally, only the connected pore volume in rock samples can be measured in most laboratory experiments. Thus, the measured porosity is also called effective porosity. Because only connected pores contribute to fluid flow and other transportation properties, “effective” porosity is more important than “true” porosity, including isolated pores. The other similar parameter is the void ratio, defined as the ratio of the cumulative pore volume in a rock sample and the cumulative solid volume of the sample. The void ratio can be determined uniquely from the porosity. Porosity is more popularly used in rock mechanics areas, but the void ratio is more popularly used in soil mechanics areas.

As direct measurement methods to determine the porosity of a rock sample, the cumulative pore volume (V_{pore}) in the sample and the total volume (V_{total}) of the sample must be measured, as shown in Eq. (1). Three main porosity measurement methods are utilized widely. As the first method, called the buoyancy method, derives V_{pore} from the difference between masses (or weights) of a rock sample under completely water-saturated and completely dried states and V_{total} from the difference between masses of the completely

water-saturated sample in air (atmosphere) and in water based on Archimedes' principle (Franklin 1979). The second method uses a helium gas pycnometer to measure the cumulative solid volume (V_{solid}) of a dry sample based on Boyle's law of the volume-pressure relationship (Blum 1997). In addition, the sum of this V_{solid} and the cumulative pore volume V_{pore} obtained by the difference of water-saturated and dried sample masses gives the total sample volume V_{total} . Therefore, porosity φ can be easily determined by $V_{\text{pore}}/V_{\text{total}}$. The third method called mercury intrusion porosimetry uses a porosimeter to determine not only the total volume (both solid and pore volumes) of a dry rock sample but also the cumulative pore volume from the volume of mercury intruded into the pores under high pressures (ASTM D4404–18 2018).

In addition to the laboratory measurement methods of rock porosity mentioned above, several indirect measurement/estimation methods are frequently applied in geoscience and geoenvironmental areas. In particular, geophysical loggings are widely utilized in drilling and/or drilled boreholes because they have an important advantage in which a continuous depth distribution of porosity can be obtained. Nuclear logs, including neutron logs and density logs, are available; the former uses neutrons to measure the quantity of water (H_2O) in the completely water-saturated rocks around the borehole to determine porosity, and the latter uses gamma rays to measure the bulk density of rocks and then to derive porosity. In addition, porosity depth profiles can be estimated based on the correlations (empirical equations) between porosity and elastic wave velocity from sonic (acoustic) logs and electric resistivity from resistivity logs. The porosity of rocks typically changes from <1% for hard rocks to >50% for soft rocks. For example, the porosity of Berea sandstone shown in Fig. 1 is approximately 20% (Lin et al. 2011). Generally, the porosity decreases as the depth increases in sedimentary basins. As an example, porosity changed from ~60% at a depth of 100 m below seafloor (mbsf) to ~20% or less at 3000 mbsf in the Nankai Trough Subduction Zone drilling site C0002 (Kitajima et al. 2017).

Pore size distribution is another important characteristic of pore structure that may affect fluid flow properties through the pore network. Under the assumption of the same effective porosity, for example, the larger the pore size is, the more easily the fluid flows through the pore network. In principle, pore size distribution can be evaluated by analysis of images such as those shown in Fig. 1. A more widely used technique is the mercury intrusion porosimetry method, which can determine not only the cumulative pore volume but also the pore volume corresponding to a certain distribution of pore sizes (ASTM D4404–18 2018). In this method, the non-wetting liquid (mercury in this method) can be forced into the connected pores by applying external pressure. The size of the pores that are intruded is inversely proportional to the applied pressure; thus, the pore size can be estimated from the



Pore Structure, Fig. 2 Pore size distributions of a Berea sandstone sample by mercury intrusion porosimetry. For determination of the pore size distribution, a cylindrical model of pore shape was assumed (ASTM D4404–18 2018)

pressure based on the Washburn equation. Mercury intrusion porosimetry is useful only for measuring the pores connected to the outside of the rock sample. This method gives only the volume of intrudable pores that have an apparent diameter corresponding to a pressure within the pressurizing range of the testing instrument; however, it does not give the true sizes of individual pores. Figure 2 shows the pore size distribution of a Berea sandstone sample obtained by mercury intrusion porosimetry, similar to that presented by Lin et al. (2011).

Relationships Between Porosity and Other Physical Properties

As an internal factor of rocks, pore structure is not only related to the physical properties of the rocks but also affects the properties. In addition, the types of fluids existing in the pores, e.g., water, oil, air, and pore fluid pressure in the pores, may also affect the physical properties of the rocks. The bulk densities of a rock in both the water-saturated and dried states are inversely proportional to porosity. Elastic wave velocities (P- and S-wave velocities), electric resistivity (the inverse of electric conductivity), strengths (e.g., compressive and tensile strengths), and Young's modulus increase as porosity decreases. Permeability shows a positive correlation with porosity but is also dependent on the pore size distribution. The thermal conductivity of rocks shows a negative correlation with porosity because solid minerals have higher thermal conductivity than fluids in the pores (Lin et al. 2020).

Summary

Pores must be available in all rocks occurring in the Earth's crust down to depths of several tens of kilometers

underground. Knowledge of pores and pore structure is important for geosciences and geoengineering. In this chapter, the definition, visual observations, quantitative evaluations, and relationships with the other physical properties are briefly described. Only fundamental knowledge is shown above and may not be sufficient for many readers. Fortunately, abundant previous research articles on the pore structure of rocks have been published in various scientific journals.

Cross-References

- Flow in Porous Media
- Porous Medium

Bibliography

- ASTM D4404-18 (2018) Standard test method for determination of pore volume and pore volume distribution of soil and rock by mercury intrusion porosimetry, ASTM International, West Conshohocken. <https://doi.org/10.1520/D4404-18>
- Blum P (1997) Moisture and density (by mass and volume). In Blum P (ed) Physical properties handbook: a guide to the shipboard measurement of physical properties of deep-sea cores. ODP tech. note, 26. <https://doi.org/10.2973/odp.tn.26.1997>
- Chen Y, Naoi M, Tomonaga Y, Akai T, Tanaka H, Takagi S, Ishida T (2018, 1976) Method for visualizing fractures induced by laboratory-based hydraulic fracturing and its application to shale samples. *Energies* 11. <https://doi.org/10.3390/en11081976>
- Cnudde V, Boone MN (2013) High-resolution X-ray computed tomography in geosciences: a review of the current technology and applications. *Earth Sci Rev* 123:1–17. <https://doi.org/10.1016/j.earscirev.2013.04.003>
- Franklin JA Coordinator (1979) Suggested methods for determining water content, porosity, density, absorption and related properties and swelling and slake-durability index properties : Part 1: suggested methods for determining water content, porosity, density, absorption and related properties. *Int J Rock Mech Min Sci Geomech Abstr* 16: 141–156. [https://doi.org/10.1016/0148-9062\(79\)91452-9](https://doi.org/10.1016/0148-9062(79)91452-9)
- Hannibal JT (2020) Berea sandstone: a heritage stone of international significance from Ohio, USA. *Geol Soc Lond, Spec Publ* 486: 177–204. <https://doi.org/10.1144/SP486-2019-33>
- Kitajima H, Saffer D, Sone H, Tobin H, Hirose T (2017) In situ stress and pore pressure in the deep interior of the Nankai accretionary prism, Integrated Ocean Drilling Program Site C0002. *Geophys Res Lett* 44: 9644–9652. <https://doi.org/10.1002/2017GL075127>
- Lin W, Tadai O, Hirose T, Tanikawa W, Takahashi M, Mukoyoshi H, Kinoshita M (2011) Thermal conductivities under high pressure in core samples from IODP NanTroSEIZE drilling site C0001. *Geochem Geophys Geosyst* 12:Q0AD14. <https://doi.org/10.1029/2010GC003449>
- Lin W, Hirose T, Tadai O, Tanikawa W, Ishitsuka K, Yang X (2020) Thermal conductivity profile in the Nankai accretionary prism at IODP NanTroSEIZE site C0002: estimations from high-pressure experiments using input site sediments. *Geochem Geophys Geosyst* 21:e2020GC009108. <https://doi.org/10.1029/2020GC009108>
- Takahashi M, Kato M, Lin W, Sato M (2016) Three-dimensional pore geometry and permeability anisotropy of Berea sandstone under hydrostatic pressure: connecting path and tortuosity data obtained by microfocus X-ray CT. In Eggers MJ et al (eds) *Developments in engineering geology*. Geological Society London, Engineering Geology Special Publication, 27, pp 207–215. <https://doi.org/10.1144/EGSP27.18>

Porosity

Pham Huy Giao^{1,2} and Pham Huy Nguyen³

¹PetroVietnam University (PVU), Baria-Vung Tau, Vietnam

²Vietnam Petroleum Institute (VPI), Hanoi, Vietnam

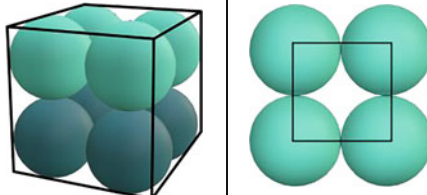
³Imperial College, London, UK

Definition

Porosity is a fundamental property of material in general and geomaterials (soils and rocks) in particular, being used in many fields such as material science, earth science, hydrogeology, rock and soil mechanics, geotechnical and environmental engineering, petroleum geoscience and engineering etc. Porosity ($\emptyset$) is a measure of the void space in soils, rocks and it is defined as the ratio between the pore volume (V_p) and the total volume (V_t) or the bulk volume (V_b). A conceptual model of geomaterial (soils, rocks), also known as a petrophysical model, is shown in Fig. 1a, consisting of three major components, i.e., solid, pore space and pore-filling fluids. Porosity can be in fraction (see Eq. 1a) or in percentage (see Eq. 1b). An example of how to calculate the porosity of an idealized granular medium made of uniform spherical grains arranged in a cubic packing is shown in Fig. 1b.

Fundamental Types of Geomaterial Porosity (Soils, Rocks)

Porosity of idealized spherical grain packings as shown in Fig. 2 has been studied for a long time (e.g., Graton and Fraser 1935), for which one single concept of porosity is enough to characterize the pore space. In nature, however, there are many diverse rock types belonging to three main groups of rocks, i.e., *igneous, sedimentary and metamorphic*. Igneous rocks form due to colling and solidification of a magma rising from the deeper part of Earth's crust or mantle. Sedimentary rocks are formed by deposition and cementation of sediment particles or by precipitation of minerals from water. Metamorphic rocks are formed by changes of preexisting rocks under the extreme conditions of heat, pressure, or reactive fluids. Consequently, the porosity concept has to diversify to characterize better the rocks in the real geological conditions. Before presenting some geological and engineering

<table><tr><td rowspan="2">Vt or Vb</td><td>Vp</td><td>Pores, voids</td><td>Pore-filling fluids (water, air, gas, oil)</td></tr><tr><td>Vm</td><td>Grains</td><td>Matrix (solids)</td></tr></table>	Vt or Vb	Vp	Pores, voids	Pore-filling fluids (water, air, gas, oil)	Vm	Grains	Matrix (solids)	$\phi = \frac{V_p}{V_t} \quad (\text{Eq. 1a})$ $\phi (\%) = \frac{V_p}{V_t} \times 100 \quad (\text{Eq. 1b})$	
Vt or Vb		Vp	Pores, voids	Pore-filling fluids (water, air, gas, oil)					
	Vm	Grains	Matrix (solids)						
(a) Conceptual soil/rock model (petrophysical model): V_t is the total volume (or V_b , the bulk volume), V_p is the pore volume, V_m is the matrix volume (or V_s , the solid volume), ϕ is the porosity.	(b) Porosity of a cubic packing with sphere with radius of r : $V_t = (4r)^3; V_m = 8 \times \frac{4}{3} \pi r^3; V_p = V_t - V_m$ $\phi = \frac{V_p}{V_t} = \frac{64r^3 - \frac{32}{3} \pi r^3}{64r^3} = \frac{5}{6} \pi = 0.476$								

Porosity, Fig. 1 Porosity of a conceptual geomaterial

classifications of porosity in the next section three fundamental types of porosity are explained below:

Total Porosity

As defined in Eq. 1a, b it is the ratio of the total pore space to the total volume of rocks.

Effective Porosity

Is the ratio of interconnected pore volume with respect to the total volume of rocks, being less than the total porosity in most cases. The isolated pores, partially filled pores or micropores that cannot transmit the fluid are not considered in the effective porosity.

Critical Porosity

Raymer et al. (1980) published some empirical velocity–porosity relations based on observations of the velocity change in a transition from solid to suspension, leading to the creation of the critical porosity concept. Initially conceptualized for a relatively simple clean sand, it has expanded to more complex rocks, i.e., rocks have a characteristic porosity above that they behave as a suspension, and this boundary value separates their mechanical and acoustic behavior into two distinct domains (Nur et al. 1998). The critical porosity plays an important role in rock physical modeling and construction of rock physical template for seismo-petrophysical characterization of complex hydrocarbon reservoirs (Giao et al. 2021).

The idealized packings in Fig. 2 represent the most simplified cases of clean sand or sandstone with single-porosity system (i.e., interparticle pores). Fig. 3a shows a well-sorted grain packing whose void space is reduced with time due to precipitation, cementation and infilling of the pore spaces as shown in Fig. 3b. Due fracturing a double porosity systems is developed (Fig. 3c). Some other major rock types like

carbonate rocks that account for approximately 50% of oil and gas production around the world or fractured igneous and metamorphic rocks can have multiple-porosity systems, combining primary and secondary porosities of both matrix and fractures (Fig. 3d). A possible difference between the double porosity systems in Fig. 3c and d is that in the latter case the primary porosity of matrix might be very small, e.g. cases of carbonates, igneous rocks. When the matrix porosity is insignificant the geomaterial is considered a fractured medium. Types and distribution of pores within the rock mass exert strong control on petrophysical characteristics of these rocks.

There are a number of mathematical models of porosity change with time and depth. For example, a one-dimensional compaction model can be described in its simplest form by a nonlinear diffusion equation as follows

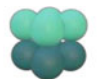
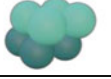
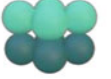
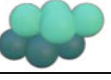
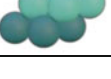
$$\frac{\partial \phi}{\partial t} = \lambda \frac{\partial}{\partial z} \left\{ k(1 - \phi) \left[- \frac{\partial \sigma'}{\partial z} (1 - \phi) \right] \right\} \quad (2)$$

Where: t is time; z is depth; ϕ is the normalized porosity at depth z ; λ is the ratio of the hydraulic conductivity to the sedimentation rate; σ' is the effective overburden/stress; ϕ is the normalized porosity at depth z . Eq. 3 can be used to simulate time-dependent compaction of rocks in a sedimentary basin, which occurs when accumulating sediments compact under their own weight, expelling pore water in the process. A good analytical solution of Eq. 3 can be found in Fowler and Yang (1998).

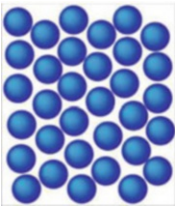
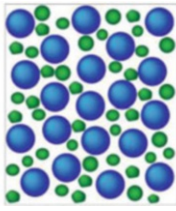
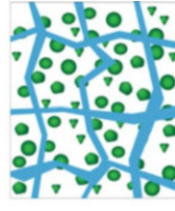
Classification of Porosity

While in geotechnical engineering classification of porosity is not much paid attention to, in petroleum geoscience and

Porosity, Fig. 2 Porosity of various idealized spherical grain packings

Packing	Basal layer	3D view	Top view	The second layer
Cubic: ($\Phi = 47.6\%$)	square packs			same as the underlying layer
Hexagonal ($\Phi = 39.5\%$)				shifted laterally with half a radius in one direction on the underlying layer
Rhombohedral ($\Phi = 26.0\%$)				shifted diagonally, dropped into the square centers of the underlying layer
Orthorhombic ($\Phi = 39.5\%$)	rhombic packs			same as the underlying layer
Tetragonal ($\Phi = 30.2\%$)				shifted laterally with half a radius in one direction on the underlying layer
Triclinic ($\Phi = 26.0\%$)				shifted diagonally, dropped into the square centers of the underlying layer

Porosity, Fig. 3 Types and changes of porosity

			
(a) High (primary) porosity	(b) Low (primary) porosity	(c) Fracture (secondary) porosity	(d) Total (combined) matrix and fracture porosity

engineering the classification of porosity is extensively used for hydrocarbon potential assessment, reservoir characterization, reserve estimation etc. There are two widely used classifications of porosity in these fields, i.e., geological classification and engineering classification (Tiab and Donaldson 2015).

Geological Classification of Porosity

The *primary porosity* was formed during deposition of clastic sedimentary rocks (Fig. 3a and b) or during the cooling and solidification process of igneous rocks, while the *secondary porosity* (Fig. 3c and d) is caused by the changes after formation of rocks.

Primary porosity consists of the following void types: (i) *Intercrystalline voids* between cleavage planes of crystals,

between individual crystals, and in crystal lattices. They are mostly of sub-capillary size (less than $2\ \mu\text{m}$ in diameter), and can be called microporosity; (ii) *Intergranular or interparticle voids* between grains, i.e., interstitial voids of all kinds in all types of rocks. These voids range from sub-capillary to super-capillary size (greater than $0.5\ \text{mm}$ in diameter); (iii) *Bedding voids* along the bedding planes caused by differences of sediments deposited, i.e., particle sizes and arrangements, the depositional environment; (iv) *Bio-induced voids* that are vuggy and irregular voids resulting from the accumulation of detrital fragments of fossils, packing of oolites, voids created by living organisms at the time of deposition.

Secondary porosity is the result of geological processes occurring after original formation or rocks such as post-

Porosity, Table 1 Classification of porosity and permeability for clastic reservoir characterization

Porosity (%)	Reservoir quality	Permeability (mD)	Reservoir quality
0–5	Negligible	<1	Poor/tight
5–10	Poor	1–10	Fair
10–15	Fair	10–50	Moderate
15–20	Good	50–250	Good
>20	Very good	>250	Very good

deposition diagenesis and catagenesis processes for clastic sediments, dolomitization of carbonates, solution, fracturing, faulting of carbonate, igneous and metamorphic rocks. Main types of secondary porosity are: (i) *solution porosity*; (ii) *dolomitization porosity*; (iii) *fracture porosity*;

Engineering Classification of Reservoir Porosity

An example of engineering classification of porosity used in hydrocarbon reservoir characterization is shown in Table 1 below. It is noted that the engineering criteria change with time, depending on reservoir type and technology advancement, say, more than 50 years ago a permeability of 50 mD indicated a tight reservoir (Tiab and Donaldson 2015) or while a 5%-porosity is negligible for a sandstone reservoir it is the maximum one can expect from a fractured granite basement reservoir.

Related Parameters

Void Ratio

In soil mechanics a variant of porosity named void ratio (e) is commonly used instead, defined as the ratio between the volume of voids and volume of matrix as defined below:

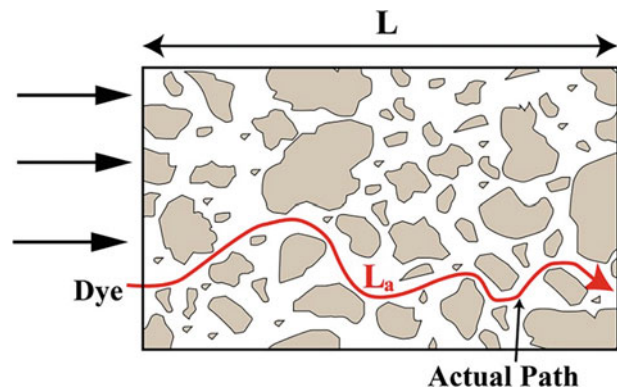
$$e = \frac{V_p}{V_m} = \frac{V_p}{V_t - V_p} = \frac{\emptyset}{1 - \emptyset} = \frac{n}{1 - n} \quad (3)$$

Where: e is void ratio; $\emptyset$ is porosity, similarly n is also porosity as commonly denoted in soil mechanics or geotechnical engineering; V_t , V_p , V_m are the same as in Eq. 1a, b (Fig. 1a).

Tortuosity

Tortuosity (τ) of a porous material is the ratio of actual flow path length (L_a) to the straight distance between the ends of the flow path (L) as defined in Eq. 4 and shown in Fig. 4 with a dye tracer dispersion. Tortuosity can be of different types, i.e., geometric, hydraulic, electrical, and diffusive. It is a useful parameter to study the structure of porous media, their electrical and hydraulic conductivity, the travel time and length for tracer dispersion (Ghanbarian et al. 2013):

$$\tau = \frac{L_a}{L} \quad (4)$$

**Porosity, Fig. 4** Definition of tortuosity

Permeability and Porosity-Permeability Relationship (Poro-perm)

Permeability (k) is defined as the ability of geomaterial (soils, rocks) to let the fluids flow through under a hydraulic/pressure gradient or loading. In 1856 Henry Darcy, a French engineer, based on laboratory experiments of water flow through core sand samples, developed an equation (Eq. 5) that had become a valuable mathematical tool to determine hydraulic conductivity for groundwater flow or permeability for hydrocarbon flow as follows:

$$V = \frac{q}{A_c} = \frac{k}{\mu} \frac{dp}{dl} \quad (5)$$

Where: v is the velocity of fluid flow (cm/s); q is the flow rate (cm³/s); k is the permeability of rocks in Darcy (D), 1D = 0.986923 μm²; A_c is the cross-section of the core sample (cm²); μ is the viscosity of the fluid in centipoises (cP); dp/dl is the pressure gradient in the direction of flow (atm/cm).

Porosity and permeability form a couple of most related petrophysical parameters, with the latter being more difficult to be tested and estimated comparing to the former. That is why one always tries to establish some empirical relationships between these two parameters based on the core measurements. One early and common method is to plot porosity (on a linear scale) against permeability (on a logarithmic scale) and observe a curve fitting trend called poro-perm, which can be

used to predict permeability based on porosity values. For a non-linear porosity-permeability relationship, considering the rock heterogeneity, Kozeny-Carman equation is a good mathematical model for permeability prediction as widely used in petroleum engineering (Tiab and Donaldson 2015), and namely:

$$k = \frac{1}{K_{ps} \tau S_{vgr}^2} \frac{\emptyset^3}{(1 - \emptyset)^2} \quad (6a)$$

$$k = \frac{1}{5S_{vgr}^2} \frac{\emptyset^3}{(1 - \emptyset)^2} \quad (6b)$$

Where: k is permeability (D); K_{ps} is the pore factor; τ is the tortuosity; S_{vgr} is The specific surface area per unit grain volume (cm⁻¹); $\emptyset$ is the effective porosity; The product ($K_{ps} \cdot \tau$) is often taken equal to a numeric value, say, 5 for many porous rocks as shown in Eq. 6b, but it can vary from 2 to 5 and up depending on case by case.

Methods to Determine Porosity

Porosity can be directly measured on the soil and rock samples in the laboratory by various methods such as buoyancy, helium porosimeter, fluid saturation, mercury porosimetry (Glover 2020; Ulusay 2015), thin section analysis, digital rock physics analysis or indirectly determined by analysis of well log curves, e.g., density, sonic, neutron, resistivity, NMR logs (Tiab and Donaldson 2015). In geotechnical engineering, the porosity and void ratio of fine-grained soils like clays are commonly determined using the popular one-dimensional consolidation or oedometer test (ASTM D2345 2020).

Summary or Conclusions

Porosity is a fundamental property of soils and rocks, having many important applications in earth science, hydrogeology, geotechnical engineering, environmental engineering, petroleum geoscience and engineering etc. Extensive studies have been done on porosity since very early time of science and engineering development. Porosity of some clean sands can be calculated based on idealized spherical packings, but with time porosity is reduced due to geological process occurring after the rock formation such as cementation, solution, infilling of the pores, fracturing. Consequently, for rock characterization the porosity concept has to diversify and some classifications were proposed, e.g., geological and engineering classifications. There are two approaches to determine porosity directly and indirectly. The former is conducted on the core samples in the laboratory using various techniques,

while the latter is predominant in the oil and gas industry based on analysis of well log data.

Cross-References

► [Porous Medium](#)

Bibliography

- ASTM D2435 (2020) Standard test methods for one-dimensional consolidation properties of soils using incremental loading. American Society for Testing and Materials (ASTM)
- Fowler A, Yang X (1998) Fast and slow compaction in sedimentary basins. *Appl Math* 59(1):365–385
- Ghanbarian B, Hunt AG, Ewing RP, Sahimi M (2013) Tortuosity in porous media: a critical review. *Soil Sci Soc Am J* 77(5):1461
- Giao PH, Trang PH, Hien DH, Ngoc PQ (2021) Construction and application of an adapted rock physics template (ARPT) for characterizing a deep and strongly cemented gas sand in the Nam Con Son Basin, Vietnam. *J Natl Gas Sci Eng* 94:104117. <https://doi.org/10.1016/j.jngse.2021.104117>
- Glover Paul WJ (2020) Petrophysics MSc course notes, 376 p., Dept. of Geology and Petroleum Geology. University of Aberdeen, UK
- Graton LC, Fraser HJ (1935) Systematic packing of spheres, with particular relation to porosity and permeability. *J Geol* 43:785–909
- Nur A, Mavko G, Dvorkin J, Galmudi D (1998) Critical porosity: a key to relating physical properties to porosity in rocks. *Lead Edge* 17(3): 357–362
- Raymer LL, Hunt ER, and Gardner JS (1980) An improved sonic transit time-to-porosity transform, SPWLA, 21st annual logging symposium, 8–11 July 1980
- Tiab D and Donaldson EC (2015) Petrophysics: theory and practice of measuring reservoir rock and fluid transport properties, 4th edn., 854 p., Gulf Professional Publishing, Elsevier as GPP is an imprint of Elsevier, USA and UK
- Ulusay R (2015) The ISRM suggested methods for rock characterization, testing and monitoring: 2007–2014, Eds. By R. Ulusay, 292 p., Springer International Publishing Switzerland

Porous Medium

Jennifer McKinley
Geography, School of Natural and Built Environment,
Queen's University, Belfast, UK

Definition

A porous medium can be defined as a material that contains spaces, which, when continuous in some way, enable the movement of fluids, air, and materials of different chemical and physical properties. The ability of a porous medium to permit the movement of fluids, air, or a mixture of different fluids is related to porosity and the variability of pore

characteristics in that the presence of interconnected pore spaces, and thus permeability properties of the material, enables the movement of fluids and accompanying materials.

Characteristics of Porous Medium

The characteristics of a porous medium are a result of variability in porosity and permeability, and the spatial continuity of these properties (McKinley and Warke 2006). Porosity as defined by Tucker (1991) is a measure of the amount of pore space in a porous medium and can be characterized as total or absolute porosity and effective porosity. Total or absolute porosity refers to the total void space within a porous medium including void space within grains. Effective porosity can be described as the interconnected pore volume in a porous medium (McKinley and Warke 2006). Effective porosity is more closely related to permeability, which can be defined as the ability of a porous medium to transmit fluids. Permeability is related to the shape and size of pores or voids and pore connections throats), and also on the properties of the fluids involved (i.e., capillary forces, viscosity, and pressure gradient). Permeability as calculated from Darcy's law, in simple terms, is a measure of how easily a fluid of a certain viscosity flows through a porous medium under a pressure gradient (Allen et al. 1988).

Depositional and Diagenetic Characteristics of Geological Porous Media

Geological materials exhibit inherited characteristics from their depositional, compaction, and cementation or crystallization history (McKinley and Warke 2006). This results in individual pores or voids which vary in size, shape, and arrangement, directing the movement of fluids, air, and accompanying materials along preferred pathways and at differential rates. Natural porous medium such as rocks seldom retain their original porosity (McKinley and Warke 2006). The key factors that control porosity and permeability in geological porous medium are primary depositional characteristics (including fabric features) and diagenetic features such as cements and salts (Worden 1998). Primary depositional processes produce fabric characteristics, which in turn may be further modified by processes such as compaction and cementation. This results in primary and secondary porosity. Primary porosity is developed as a sediment is deposited and includes inter- and intraparticle/granular porosity (Tucker 1991). Secondary porosity of a porous medium develops during diagenesis by dissolution or removal of soluble material and through tectonic movements producing fracturing. Fractures or vugs within geological materials may contribute

substantially to the flow capacity (i.e., permeability properties) but contribute little to absolute porosity of a porous medium. Secondary diagenetic precipitation, in the form of cements and salts, has the potential to seal fractures or vugs and reduce the permeability properties of a porous medium. The predominant cements, which affect sandstone porous media, comprise carbonates, clay minerals, and quartz cements. Aspects of carbonate, quartz and clay cementation in sandstones have been comprehensively covered in three special publications (Morad 1998; Worden and Morad 2000, 2001). The variability of porosity in limestone porous media tends to be more erratic in type and distribution than for sandstones (Tucker 1991). Based on the seminal classification scheme by Choquette and Pray 1970, porosity types in carbonate porous media can be defined as fabric selective, depending on whether pores are defined by the fabric (grains and matrix) of the limestone (e.g., intercrystalline), and non-fabric selective, porosity that cuts across the actual rock fabric (e.g., fracture porosity). Stylolites in carbonate porous media can form a type of porosity in terms of acting as conduits for fluid movement or conversely stylolites can produce a reduction in porosity of the porous medium through the accumulation of clays and insoluble residue. Porosity in crystalline porous media, including igneous and metamorphic, occurs generally as a result of fracturing, granular decomposition or dissolution, and may be accentuated by mineral alignment or banding.

Heterogeneity of Porous Medium

No naturally occurring porous medium is homogeneous but some are relatively more homogeneous than others. The problem of nonuniformity or heterogeneity of a porous medium is inherent even at the pore scale, since individual pores vary in size, shape, and arrangement. Small-scale features such as pore geometry or laminae also affect the heterogeneity of porous medium.

By definition a reservoir porous medium, such as sandstone, must contain pores or voids to enable the storage of oil, gas, and water, which must be connected in some way to permit the movement of the same. Fluids of differing physical properties (oil, gas, and water) move through heterogeneous porous media via preferred pathways and at differential rates. Primarily this is due to the variability in porosity, permeability, poro-perm relationship, viscosity, tortuosity, interfacial tension (wettability), and spatial continuity of the porous medium. Knowledge of how the petrophysical properties of porosity and permeability vary throughout materials and identifying the processes responsible for their reduction is important in accurate reservoir modeling of porous media.

The challenge to studies of porous media description, and intrinsically of heterogeneity, can be summarized as to:

- Describe the geology with as much detail and as realistically as possible and in a quantitative approach.
- Assess the spatial distribution of heterogeneities including cements, salts, contaminants, etc., and their effect on fluid flow.
- Subdivide the porous medium into hydraulic or flow units (units that under a given driving force behave differently and in which the variation of properties is less than the hydraulic or flow units above and below).

In addition, studies must go beyond data capture to using this information in the development of mathematical simulation models that depict porous medium behavior accurately and effectively, all of which combines to give a greater understanding of fluid flow processes in a porous medium to optimize recovery and aquifer capabilities. The issues which arise in porous medium description are well documented. The recurrent themes are:

- The detail required to adequately quantify the lateral variability of physical properties far surpasses the detail of sampling (interpolation, extrapolation, and informed interpretation are necessary).
- Many heterogeneities have a spatial distribution much less than the average well or sample spacing.
- Heterogeneity at small scale is homogenized, simplified, or smoothed over within mathematical modeling.
- Averaging of heterogeneity data for modeling purposes can lead to underestimation of their effect on porosity and permeability.

Conclusions

A porous medium contains spaces or voids that, when connected in some way, permit the movement of fluids, air, and associated materials. Porous media tend to be heterogeneous within a larger ordered framework, as a function of scale. At the pore and microscale, the sorting and packing of grains can be markedly variable. Random variations or heterogeneous elements at the pore scale may be sufficiently small to be considered homogeneous at a larger scale, e.g., lamina, stratum, or bedding scale. Porous media, therefore, may appear to be both heterogeneous and homogeneous: depending on the scale of measurement being considered. In view of this it is essential that the scale of measurement being considered be clearly stated. The importance of depositional history in naturally occurring porous media and

diagenetic heterogeneities including the influence of cementation and the diagenetic growth of minerals on porous media need to be fully assessed and recognized in any modeling approach.

Cross-References

► Porosity

Bibliography

- Allen D, Coates G, Ayoub J, Carroll J (1988) Probing for permeability: an introduction to measurements with a Gulf Coast case study. *Tech Rev* 36(1):6–20
- Choquette PW, Pray LC (1970) Geologic nomenclature and classification of porosity in sedimentary carbonates. *AAPG Bull* 54:207–250
- McKinley JM, Warke P (2006) Controls on permeability: implications for stone weathering. *Geol Soc Lond Spec Publ* 271:225–236. <https://doi.org/10.1144/GSL.SP.2007.271.01.22>
- Morad S (ed) (1998) Carbonate cementation in sandstones. International Association of Sedimentologists. Special Publication 26. Blackwell Science, Oxford
- Tucker ME (1991) Sedimentary petrology an introduction. Blackwell Scientific Publications, Oxford
- Worden RH (1998) Dolomite cement distribution in a sandstone from core and wireline data: the Triassic fluvial Chaunoy Formation, Paris Basin. Geological Society, London, Special Publications 136 (1):197–211. <https://doi.org/10.1144/gsl.sp.1998.136.01.17>
- Worden RH, Morad S (eds) (2000) Quartz cementation in sandstones. International Association of Sedimentologists, Special Publication, vol 29. Blackwell Science, Oxford
- Worden RH, Morad S (eds) (2001) Clay cementation in sandstones. International Association of Sedimentologists, Special Publication, vol 34. Blackwell Science, Oxford

Power Spectral Density

Abhey Ram Bansal¹ and V. P. Dimri²

¹Gravity and Magnetic Group, CSIR- National Geophysical Research Institute, Hyderabad, India

²CSIR- National Geophysical Research Institute, Hyderabad, Telangana, India

Definition

The power spectral is used in many branches of earth sciences to carry out various mathematical operations on geophysical data to understand various earth processes and earth structure. The power spectral density is a representation of power of the signal in the frequency domain. The representation of data in the frequency domain found many applications.

Introduction

The geophysical measurements carried out in space/time represent the variations of physical measures in space/time. Exploration geophysics aims to find information about the subsurface distribution of physical properties from the surface/air/ship/satellite measurements. The distribution of physical properties in the subsurface is further interpreted to find the anomalous bodies containing hydrocarbons, minerals, water or to understand the crustal structure of the Earth. These measurements are often carried out at equal intervals along the roads (profile mode) or in a grid pattern (2-D survey), depending on the study's aim. The study of variation with space and time is carried out to understand the sub-crustal structure or seismic source nature.

The transformation from space/time to frequency domain is often useful since many mathematical operations in the frequency domain become easy to carry out and further useful to find the physical properties of the subsurface of the Earth. Space/time series can be decomposed in the sum/integral of the harmonic waves of different frequencies, known as Fourier analysis or Fourier transform (FT). To carry out the FT, the time/space series should be of limited length, periodic, or satisfying the condition of absolute integrability, which can be defined as: if $x(t)$ is a continuous function of space/time, then integration of the square of the time/space series is finite (Bath 1974):

$$\int_{-\infty}^{\infty} |x(t)|^2 dt < \infty \quad (1)$$

A continuous time/space series can be described in the FT:

$$X(k) = \int_{-\infty}^{\infty} x(t) e^{-jkt} dt \quad (2)$$

where k is wavenumber $= 2\pi f$ and f = frequency. The advantage of FT is that we can carry out some operations in the frequency domain, which can be further changed back to space/time series using the formula of inverse Fourier transformation:

$$x(t) = (1/2\pi) \int_{-\infty}^{\infty} X(k) e^{jkt} dk \quad (3)$$

If $x(t)$ is not a continuous function but has a finite number values ($x_0, x_1, x_2, \dots, x_{N-1}$) at finite time $i = 0$ to $i = N-1$ having sampling interval Δt , then the Fourier transform X_m at discrete frequencies $f_m = m/N\Delta t$ is called the discrete Fourier transform (DFT):

$$X_m = (1/N) \sum_{i=0}^{N-1} x_i e^{-j(2\pi mi/N)} \text{ for } m = 0, \dots, N-1 \quad (4)$$

The quantity X_m is a complex number having real and imaginary components. Most of the time, we use power spectral density ($P(k)$), which is defined as the square of amplitude of the DFT:

$$P(k) = (1/N) \sum_{m=0}^{N-1} |X_m|^2 \quad (5)$$

The calculation of power spectral density from the data using the DFT is called the periodogram method. Cooley and Tukey (1965) introduced a fast algorithm to calculate the Fourier transform, called the fast Fourier transform (FFT). The other popular method, known as the Blackman and Tukey method (Blackman and Tukey 1958), calculates the power spectral density from the autocorrelation function based on Wiener-Khintchine theorem. Blackman and Tukey's approach is also useful for estimating power spectral density of random distribution when inequality condition of Eq. (1) is not met. The other high-resolution methods, e.g., Maximum Entropy Method (MEM) (Burg 1967) and Maximum Likelihood Method (MLM) (Capon et al. 1967), and multi-Taper Method (MTM) (Thomson 1982) can also be applied to calculate the power spectral density (Dimri 1992).

Application of Power Spectral Density (PSD)

Correlation of Space/Time Series

The power spectral density is also useful for understanding the behavior of a time/space series. The random and uncorrelated time/space series are also known as white noises. The white noises have contents of all frequencies, e.g., the slope of a fitted straight line in the plot of log frequency/wavenumber versus plot of power spectral density equals zero. The white noise series' autocorrelation has only a finite value at zero lag and zero values at other lags. However, a time/space series are often not white noises but are correlated and known as scaling noises. The power spectral density of scaling noises is frequency-dependent. The slope of a straight line fitted to the log of frequency/wavenumber versus the log of power spectral density defines the series' correlation. The higher the value of slope higher is the correlation of time/space series. Recent measurements from the boreholes have shown the distribution of physical properties in the subsurface as scaling (Bansal and Dimri 2014). The white noise distribution assumption is away from reality and it was assumed so because of mathematical simplicity and unknown information about detailed distribution of physical properties within the

crust. The scaling noises find many applications in the interpretation of geophysical data, e.g., depth estimation from gravity and magnetic data, regional residual separation of gravity and magnetic measurements, inversion of geophysical data, predictive deconvolution, ore deposits, and earthquake distribution in space and time (Turcotte 1997).

The scaling noises' power spectral density is expressed as $P(k) \propto k^\beta$, where $P(k)$, k , and β are power spectral density, wavenumber, and scaling exponent. The value of the scaling exponent equal to zero represents the white noise. The scaling exponent can be further related to the fractal dimension (D), valid for β between 1 and 3.

$$\beta = 5 - 2D \quad (8)$$

The scaling noises can be further related to fractional Gaussian and Brownian walks. The scaling noises are also useful for defining a self-similar or self-affine fractal (Turcotte 1997).

Estimation of Density/Susceptibility Distribution from Gravity and Magnetic Observations

The Fourier series analysis of space/time series is useful to solve some of the mathematical functions like convolution/deconvolution. For a linear system, the convolution operator is useful to find an output ($y(t)$) from the input signal ($x(t)$) and earth's response ($h(t)$). The convolution operator for a discrete case is defined:

$$y_t = \sum_i x_i h_{t-i} \quad (6)$$

These convolution operators exist almost in all geophysical studies. We want to study the output by sending a known input, e.g., estimating gravity/magnetic potential from the density distribution and structural function.

In the frequency domain, the convolution of gravity/magnetic potentials (Eq. 6) is converted in the form of multiplication using the well-known convolution theorem as:

$$U_g(f) = \rho(f) R(f) \quad (7)$$

where $U_g(f)$ is the Fourier transform of the gravity/magnetic potential, $\rho(f)$ is the Fourier transform of the density/susceptibility distribution, and $R(f)$ is the Fourier transform of Geometric function. In other words, the Fourier transform of gravity/magnetic potential is a combination of two multiplicative factors: (1) a function of depth and thickness, and (2) a function of distribution of density/susceptibility (Blakely 1995).

Depth Estimation from Gravity and Magnetic Data

The estimation of the depth of anomalous sources from the gravity/magnetic measurements in spectral methods expressed for a white noise distribution of sources as (Spector and Grant 1970):

$$P(k) = Ae^{-2dk} \quad (9)$$

The white noise distribution of sources provides an over-estimation of anomalous sources' depths (Bansal and Dimri 2001, 2014). The scaling noise distribution of sources resulted in the modification of the above equation as:

$$P(k) = Bk^{-\beta}e^{-2kd} \quad (10)$$

The depth estimations for the scaling distribution of sources are found close to reality. The scaling distribution of sources found many applications in estimating the depth of anomalous sources from gravity and magnetic measurements from different parts of the world.

Upward and Downward Continuation

The gravity and magnetic measurements are carried out on the Earth's surface, at a ship on the sea, air-borne surveys at sea and land, or measured at satellite elevation. These measurements are at different heights. To bring these measurements at the same elevations, frequently upward continuation of the measured data is carried out. The other advantage of upward continuation is attenuating near-surface anomalies and enhancing the more resonant anomalies at the expense of near-surface anomalies. The downward continuation is carried out to enhance the effect of the shallower bodies as compared to the upward continuation. The upward continuation is smoothing operation whereas downward continuation is unsmoothing operation (Blakely 1995). A small noise in the data causes large variations during the downward continuation. The computation of upward continuation is complicated in the space domain where these become multiplication with the exponential term and Fourier transform of the measured field. Again we can bring back the field in space domain by taking inverse transforms.

Summary and Conclusions

The power spectral density found wide applications to understand the correlation of time series, estimation of density/susceptibility distribution within the crust, depth estimation

from gravity and magnetic data, and downward and upward continuation of potential field data.

Bibliography

- Bansal AR, Dimri VP (2001) Depth estimation from the scaling power spectral density of nonstationary gravity profile. *Pure Appl Geophys* 158:799–812. <https://doi.org/10.1007/PL00001204>
- Bansal AR, Dimri VP (2014) Modelling of magnetic data for scaling geology. *Geophys Prospect* 62(2):385–396
- Bath M (1974) *Spectral analysis in geophysics*. Elsevier Scientific, Amsterdam, p 563
- Blackman RB, Tukey JW (1958) *The measurement of power spectra from the point of view of communication engineering*. Dover Publications, New York, p 190
- Blakely RJ (1995) *Potential theory in gravity & magnetic applications*. Cambridge University Press
- Burg JP (1967) Maximum entropy spectral analysis, presented at the 37th annual international SEG meeting November 1, in Oklahoma City
- Capon J, Greenfield RJ, Kolker RJ (1967) Multidimensional maximum-likelihood processing of a large aperture seismic array. *Proc IEEE* 55: 192–211
- Cooley JW, Tukey JW (1965) An algorithm for the machine calculation of complex Fourier series. *Math Comput* 19:297–301
- Dimri VP (1992) Deconvolution and inverse theory: application to geophysical problems. Elsevier, Amsterdam
- Spector A, Grant FS (1970) Statistical model for interpreting aeromagnetic data. *Geophysics* 35(2):293–302
- Thomson DJ (1982) Spectrum estimation and harmonic analysis. *Proc IEEE* 70:1055–1096
- Turcotte DL (1997) *Fractals and chaos in Geology and Geophysics*, 2nd edition, Cambridge University Press

Predictive Geologic Mapping and Mineral Exploration

Frits Agterberg
Central Canada Division, Geomathematics Section,
Geological Survey of Canada, Ottawa, ON, Canada

Definition

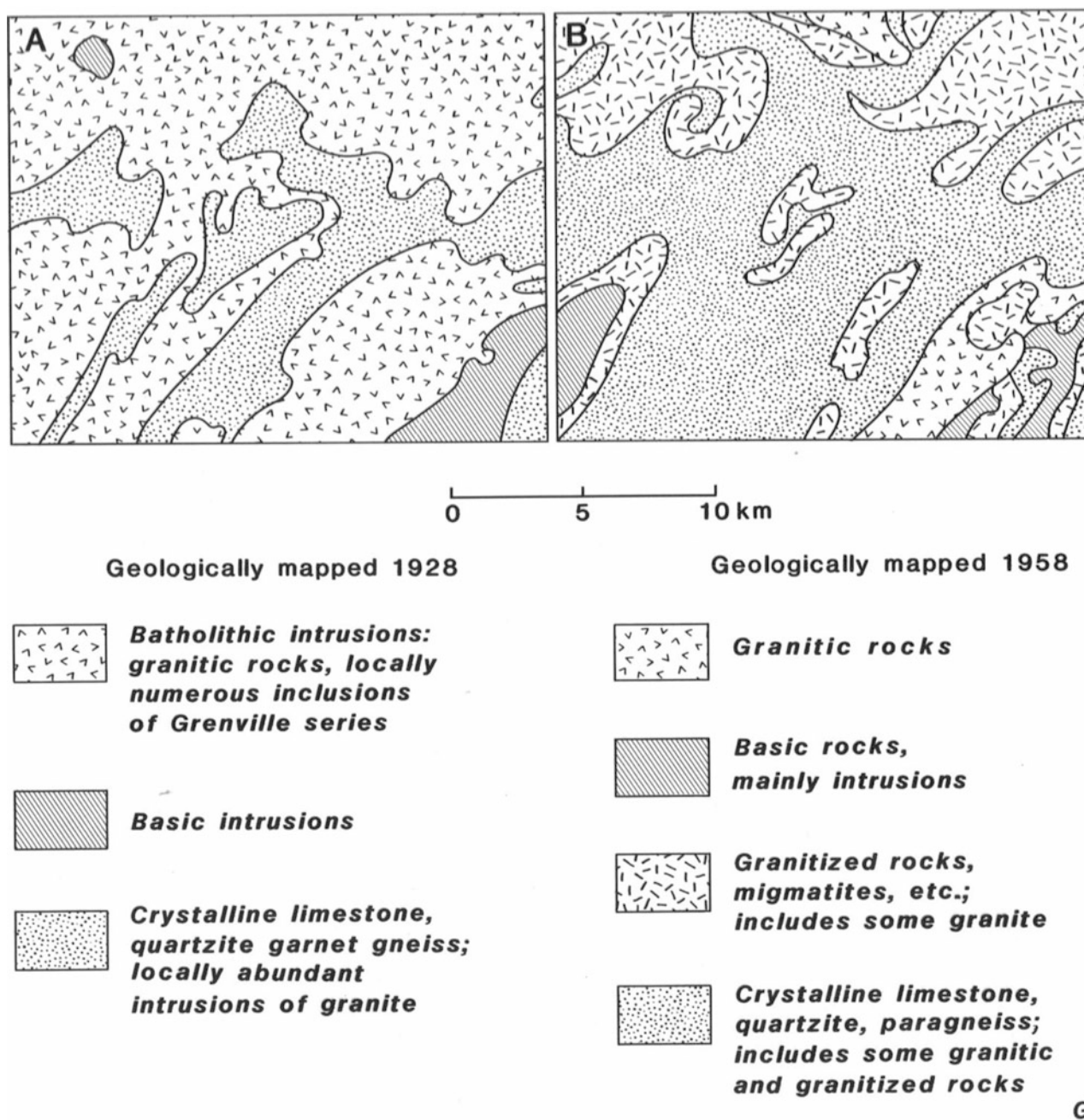
During the nineteenth century and most of the twentieth century, the geology map was the most important tool for mineral exploration. More recently, the emphasis to find new mineral deposits has shifted to the use of geophysical and geochemical methods. Nevertheless, the geologic map augmented by remote sensing at the surface of the Earth and the use of exploratory boreholes remains indispensable for target selection in mineral exploration. In this chapter, the emphasis will be on lesser known but potentially powerful new methods of predictive geologic mapping for mineral exploration, because the subject is also covered in other chapters.

Introduction

The first geologic map was published in 1815. This feat has been well documented by Winchester (2001) in his book: *The Map that Changed the World – William Smith and the Birth of Modern Geology*. According to the concept of stratigraphy, strata were deposited successively in the course of geologic time, and they are often uniquely characterized by fossils such as ammonites that are strikingly different from age to age. Additionally, there are different types of igneous and metamorphic rocks. Geologic maps turned out to be essential tools for mineral exploration. For example, during the nineteenth century, coal became important in England and other countries. Its occurrence is largely restricted to the Carboniferous Period, which occurs in many different countries and is readily recognizable in outcrops at the surface of the Earth, although it covers only a small portion of the total surface area in the regions where it occurs. Today different countries in the world have fairly complete geological maps synchronized according to international standards (*cf.* portal.oneygeology.org).

It is remarkable that the usefulness of geologic maps remained virtually unknown until approximately 1800. There are many other examples of geologic concepts that became only gradually accepted more widely, although they usually had been proposed much earlier in one form or another by individual scientists. The best-known example is plate tectonics: Alfred Wegener (1966) had demonstrated the concept of continental drift fairly convincingly as early as 1912 but this idea only became acceptable in the early 1960s. One reason that this theory initially was rejected by most geoscientists was the lack of a plausible mechanism for the movement of continents. It is important to keep in mind that geologic maps are two-dimensional and our ability to extrapolate downward from the surface of the Earth to create three-dimensional realities is severely limited. Rocks were formed at different times by means of different processes. Thus, a single internationally accepted geologic time scale is important (*cf.* Gradstein et al. 2020) in addition to sound geoscientific theory. The following two examples illustrate some of the difficulties originally and currently experienced in creating 3D geologic realities by downward extrapolation using data observed at the surface of the Earth.

Harrison (1963) has pointed out that, although the outcrops in an area remain more or less the same in the immediate past, a geologic map constructed from them can change significantly over time when new geoscientific concepts become available. A striking example is shown in Fig. 1. Over a 30-year period, the exposed bedrock in this study area on the Canadian Shield that was repeatedly visited by geologists had remained nearly unchanged. However, the geologic maps and their legends created by different geologists are entirely different. These discrepancies reflect



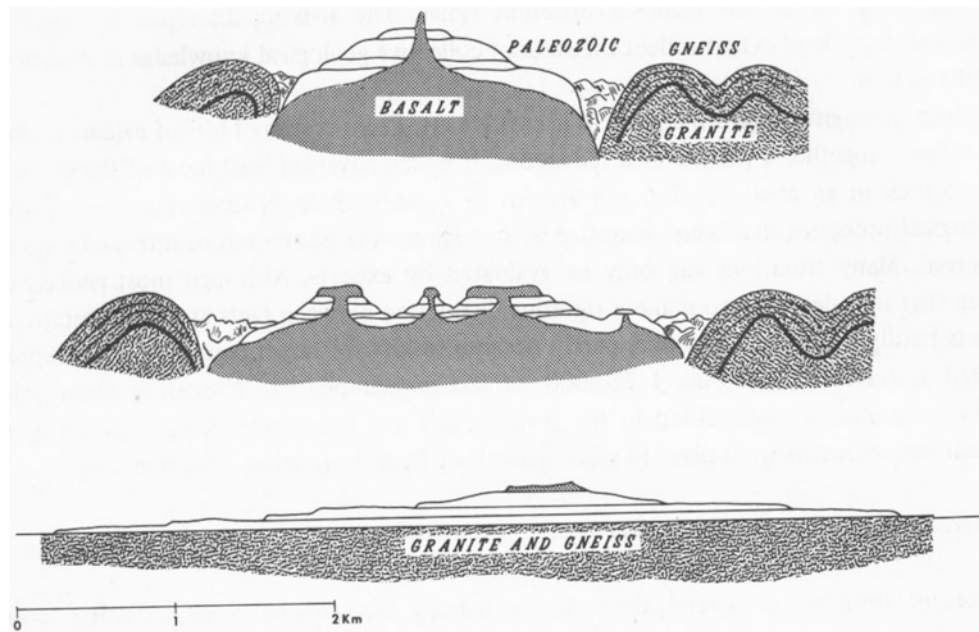
Predictive Geologic Mapping and Mineral Exploration, Fig. 1 Two geologic maps for the same area on the Canadian Shield published at different times. As pointed out by Harrison (1963), between 1928 (a) and 1958 (b) there was development of conceptual ideas about

the nature of metamorphic processes that resulted in geologists constructing different maps based on same observations (Source: Agterberg 2014, Fig. 1)

changes in the state of geologic knowledge at different points in time. In another example (Fig. 2), according to Nieuwenkamp (1968), a typical downward extrapolation problem is shown with results strongly depending on initially erroneous theoretical considerations. In the Kinnehulle area in Sweden, the tops of the hills consist of basaltic rocks; sedimentary rocks are exposed on the slopes and granites and gneisses in the valleys. The first projections into depth for this

area were made by a prominent geologist (Von Buch 1842). It can be assumed that today most geologists would quickly arrive at the third 3D reconstruction (Fig. 2c) by Westergård et al. (1943). At the time of von Buch, it was not yet common knowledge that basaltic flows can form extensive flat plates within sedimentary sequences.

The projections in Fig. 2a and b reflect A.G. Werner's pan-sedimentary theory of "Neptunism" according to which all



Predictive Geologic Mapping and Mineral Exploration, Fig. 2 Geological observations made at the surface of the Earth have to be extrapolated downwards in order to obtain a full 3-dimensional representation of reality. This can be an error prone process as pointed out by Nieuwenkamp (1968). Sections **a** and **b** for an area in Sweden are modified after von Buch (1842) illustrating his genetic interpretation that

rocks on Earth were deposited in a primeval ocean. Nieuwenkamp (1968) demonstrated that this theory was related to the philosophical concepts of F.W.J. Schelling and G.W.F. Hegel. When Werner's view was criticized by other early geologists who assumed processes of change took place during geologic time, Hegel publicly supported Werner by comparing the structure of the Earth with that of a house. One looks at the complete house only, and it is trivial that the basement was constructed first and the roof last. Initially, Werner's conceptual model provided an appropriate and temporarily adequate classification system, although von Buch, who was a student of Werner, rapidly ran into problems during his attempts to apply Neptunism to explain the occurrences of rock formations in different parts of Europe. For the Kinnehulle area (Fig. 2), von Buch assumed that the primeval granite became active later changing sediments into gneisses, while the primeval basalt became a source for hypothetical volcanoes.

Probably the most important usage of geologic maps is that they can be used to delineate target areas for the discovery of new mineral deposits by drilling boreholes in places with a relatively high probability of success in such discovery. In addition to the geologic maps, frequent use is made of remote sensing and geophysical and geochemical data. These types of information also can be used to estimate the mineral potential of larger areas, i.e., total number of undiscovered mineral deposits of different types. In the following sections, the

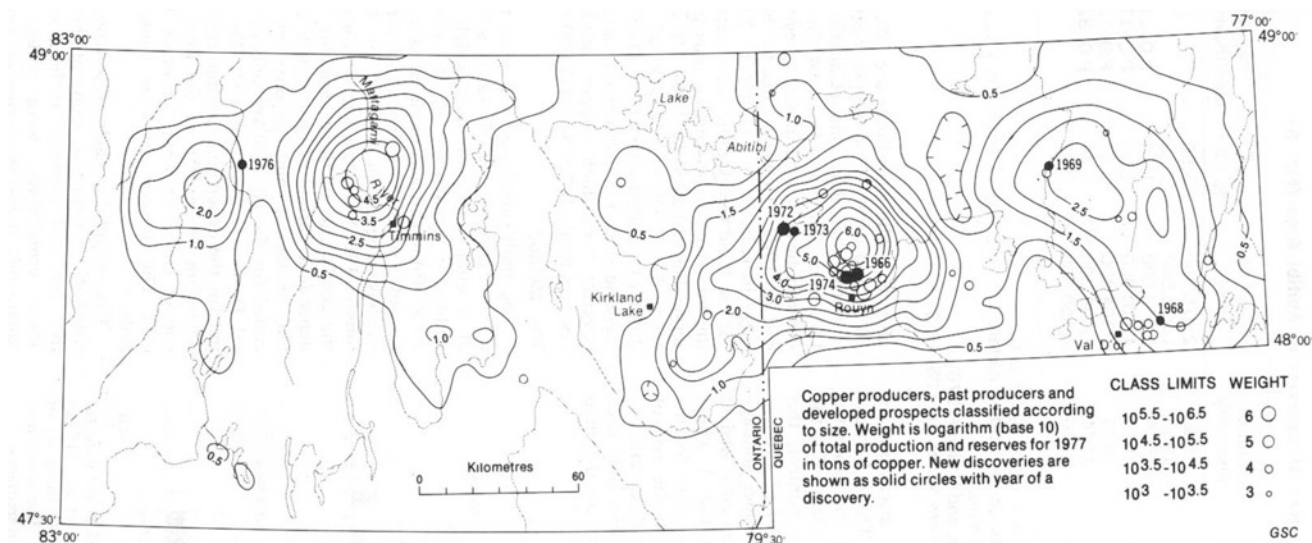
was based on combining Werner's original theory of Neptunism with von Buch's own firsthand knowledge of volcanoes including Mount Etna in Sicily with a basaltic magma chamber. He was not aware of the fact that basaltic flows can be widespread covering large areas, the genetic model used by Westergård et al. (1943) for constructing Section **c** (Source: Agterberg 1974, Fig. 1)

topics to be discussed are (1) the use of multivariate statistical methods to estimate the probability of occurrence of mineral deposits; (2) use of jackknife method to reduce bias instead of defining "control" areas; (3) weights-of-evidence (WofE) method; and (4) singularity analysis. Topics (3) and (4) are well covered in other chapters and will only be briefly discussed in this chapter.

Mineral Potential Maps

An early example of regional resource prediction is shown in Fig. 3 (from Agterberg and David 1979). Amounts of copper in existing mines and prospects had been related to lithological and geophysical data in the Abitibi area on the Canadian Shield to construct copper and zinc potential maps (Agterberg et al. 1972). During the 1970s there was extensive mineral exploration for these metals in this region that resulted in the discovery of seven new large copper deposits shown in black on Fig. 3. This example illustrates some of the advantages and drawbacks of the application of statistical techniques to predict regional mineral potential from geoscience map data. The later discoveries in the Abitibi area fit in with the prognostic contour pattern that was based on the earlier (pre-1968) discovered copper deposits (open circles in Fig. 3).

The prognostic contours in Fig. 3 are for the estimated number of (10 × 10 km) "control cells" with known copper



Predictive Geologic Mapping and Mineral Exploration, Fig. 3 Copper potential map, Abitibi area on the Canadian Shield. Contours and deposit locations are from Agterberg et al. (1972). Contour

deposits within 16 times larger (40×40 km) cells. When the contour map was constructed in 1972, there were 27 control cells. The probability that any ($10 \text{ km} \times 10 \text{ km}$) cell in the study area would contain one or more mineable copper deposits was assumed to satisfy a Bernoulli random variable with parameter p . Any contour value on the map therefore can be regarded as the mean $x = n \cdot p$ of a binomial distribution with variance $n \cdot p \cdot (1-p)$ where $n = 16$. The corresponding amount of copper would be the sum of x values drawn at random from the exceedingly skewed size-frequency distribution for amounts of copper in the 27 control cells. The resulting uncertainty for most contour values therefore was and remains exceedingly large.

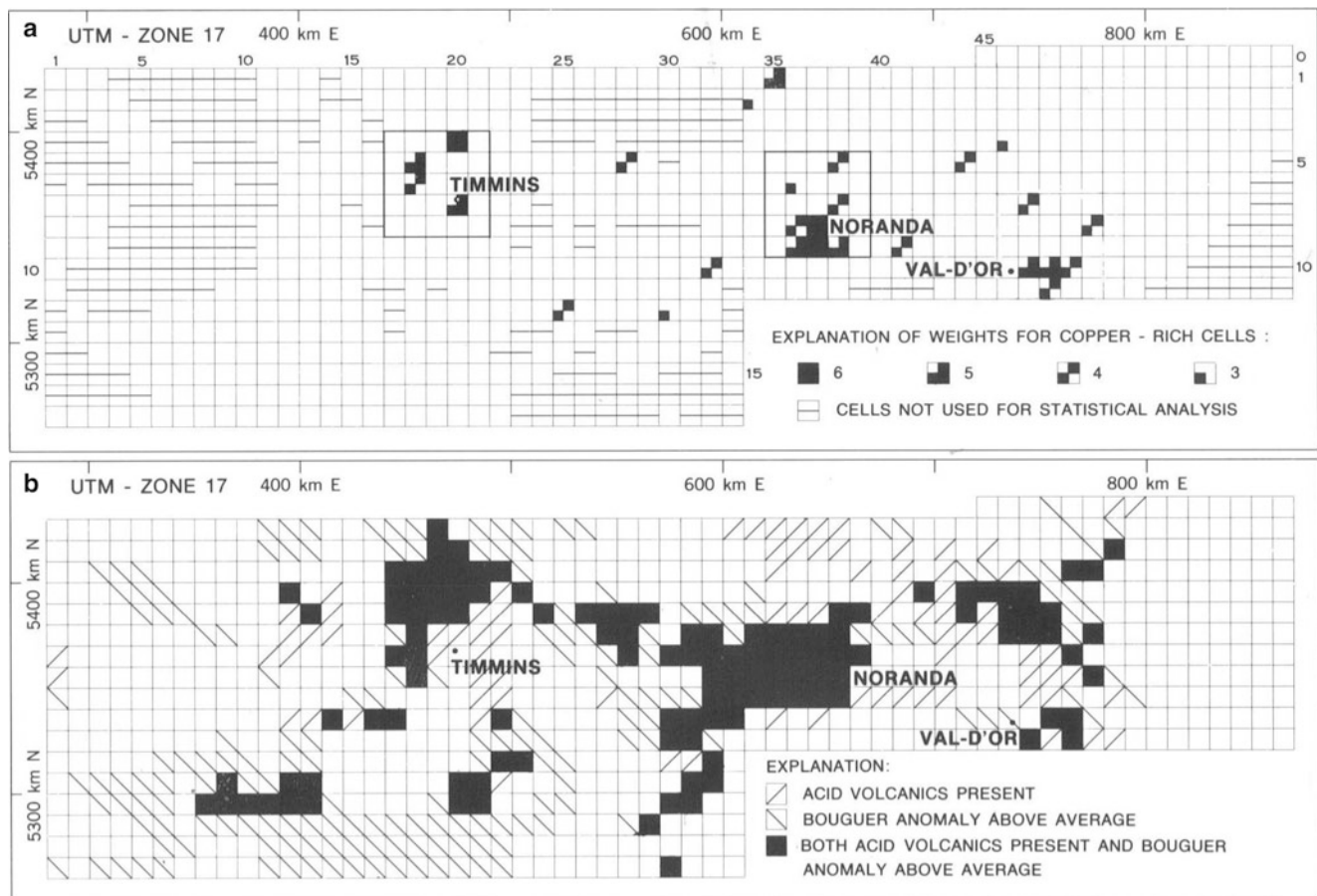
The probabilities p for the 644 ($10 \text{ km} \times 10 \text{ km}$) cells used for the example of Fig. 3 were estimated by stepwise multiple regression (Draper and Smith 1966). The value of the dependent variable for the 27 control cells out of the 644 ($10 \text{ km} \times 10 \text{ km}$) cells used was set equal to the logarithm (base 10) of the total amount of copper per control cell and equal to zero for the ($644 - 27 =$) 617 cells without, known in 1968, copper deposits. A total of 55 geological and geophysical variables were used as independent variables in this application of which 26 were retained after using the stepwise regression method. Figure 4 illustrates that simple binary combinations of the independent variables used can already show a pattern that is positively correlated with the pattern of the control cells.

Initial estimates of the dependent variable were biased because of probable occurrences of undiscovered deposits in many of the ($10 \text{ km} \times 10 \text{ km}$) cells that had been less explored. This bias was corrected by multiplying all initial

value represents expected number of ($10 \text{ km} \times 10 \text{ km}$) cells per ($40 \text{ km} \times 40 \text{ km}$) unit area containing one or more copper ore deposits (Source: Agterberg and David 1979)

estimates by a factor $F = 2.35$ representing the sum of all observed values divided by the sum of the corresponding initial estimates within a well-explored area consisting of 50 cells within the Timmins and Noranda-Rouyn mining camps jointly used as a “control area.” The total amount of mined or mineable copper contained in the entire study area (used to construct Fig. 3) amounted to 3.12 million tons. Multiplication of this amount by $F = 2.35$ gave 7.33 million tons as an estimate of the total amount of copper in the entire study area. After 9 years of intensive exploration subsequent to the construction of the contours in Fig. 3, 5.23 million tons of this hypothetical total had been discovered (*cf.* Agterberg and David 1979).

The metal potential maps published by Agterberg et al. (1972) created a significant amount of discussion in the scientific literature of which three publications will be briefly discussed here. Tukey (1972) suggested that, since probabilities cannot be negative or greater than 1, logistic regression should be considered as an alternative to using the general linear model in studies of this type. This suggestion was followed up in Agterberg (1984) for (A) copper in volcanogenic massive sulfide deposits and (B) nickel in magmatic nickel-copper deposits in a larger part of the Canadian Shield that includes the area used for Fig. 3 (see Fig. 5). The two types of copper deposits (A and B) were analyzed separately because they have different origins (*cf.* Pirajno 2010). Within the smaller Abitibi area, the pattern in Fig. 5a resembles that of the original copper potential map (Fig. 3). This is because the number of volcanogenic massive sulfide deposits significantly exceeds the number of magmatic nickel-copper deposits in this region. The general linear model failed to



Predictive Geologic Mapping and Mineral Exploration, Fig. 4 The Abitibi area on the Canadian Shield in the provinces of Ontario and Quebec had been subdivided into 643 (10 km × 10 km) cells. Top diagram (a) shows 27 copper-deposit rich cells with estimated tonnages of copper (rounded values of logarithm (base 10)). Bottom

diagram (b) shows cells with co-occurrence of acid volcanics and Bouguer anomaly above average. The two patterns are strongly correlated indicating that copper deposits tend to occur in places where acid volcanics occur on top of basic volcanics with relatively high specific gravity (Source: Agterberg 1974, Fig. 119)

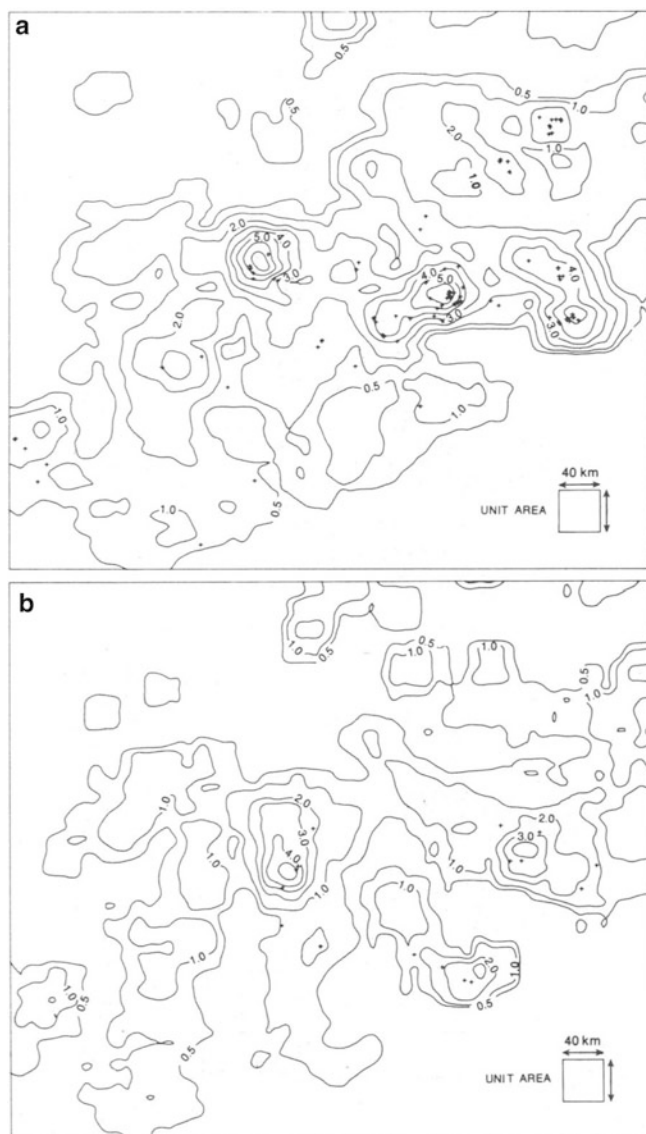
produce meaningful mineral potential contours for the relatively few magmatic nickel-copper deposits that are closely associated with mafic and ultramafic intrusions on the Canadian Shield (Agterberg 1974). This was because relatively many estimated probabilities fell out outside the [0,1] interval. Thus, for relatively rare types of deposits, logistic regression produces better results than ordinary multiple regression.

Wellmer (1983) pointed out that one of the seven new discoveries shown in Fig. 3 is the Timmins nickel-copper deposit with 3,500,000 tons of ore. It is a magmatic nickel-copper deposit and not a volcanogenic massive sulfide deposit discovered in 1976 that occurs on the westernmost peak in Fig. 3 which was delineated in 1972, relatively far away from any earlier discovered copper deposits in the Abitibi area. Since this peak is not clearly delineated in Fig. 5b obtained by logistic regression in 1974, this indicates that the general linear model locally can produce better results than logistic regression. Its final result can be regarded simply as the sum of 27 separate copper potential evaluations each based on

information for a single control cell only. This property of the general linear model is important because it allows the estimation of probable occurrences of mineral deposits that are of different genetic types. Finally, in another discussion paper, Tukey (1984) pointed out that on the peaks of maximal copper potential in Fig. 3, the known copper deposits tend to occur on the flanks of the peaks instead of in their centers. This observation was confirmed by a statistical significance test Agterberg (1984), but a geoscientific explanation of this feature has not yet been found.

Bias and the Jackknife

In order to compare various geoscientific trend surface and kriging applications with one another, Agterberg (1970) had randomly divided the input data set for a study area used for example into three subsets: Two of these subsets were used for control and results derived for the two control sets were



Predictive Geologic Mapping and Mineral Exploration, Fig. 5 Occurrences of (a) copper-zinc deposits, and (b) magmatic nickel-copper deposits in (40 km × 40 km) cells on part of the Precambrian Canadian Shield. was used to estimate The underlying probabilities for (10 km × 10 km) cells were estimated by logistic instead of linear regression for an area larger than that used for Fig. 4. Note the similarity of pattern for the central part of (a) with the pattern of Fig. 4. Logistic regression gave better results than linear regression for relatively few of the nickel-copper deposits of (b) (Source: Agterberg 1984, Fig. 3)

then applied to a third “blind” subset in order to see how well results for the control subsets could predict the values in the third subset. In his comments on this approach, Tukey (1970) stated that this form of cross-validation could indeed be used but a better technique would be to use the then newly proposed Jackknife method. The following brief explanation of this method is based on Efron (1982). In geoscientific mineral potential applications, it can help to eliminate bias due to great variations in the regional intensity of exploration, a problem

that in the previous section was solved by defining a control area consisting of well-explored mining camps (bias factor $F = 2.35$ in Abitibi area).

For cross-validation it has become common practice to leave out one data point (or a small number of data points) at a time and fit the model using the remaining points to see how well the reduced data set does at the excluded point (or set of points). The average of all prediction errors then provides the cross-validated measure of the prediction error. Cross-validation, the jackknife, and bootstrap are three techniques that are closely related. Efron (1982), Chapter 7) discusses their relationships in a regression context pointing out that, although the three methods are close in theory, they generally yield different results in practical applications. The following application was previously described in Agterberg (2021).

For a set of n independent and identically distributed (*iid*) data, the standard deviation of the sample mean ($\bar{x}$) satisfies $\hat{\sigma}(\bar{x}) = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n(n-1)}}$. This result cannot be extended to other estimators such as the median. However, the jackknife and bootstrap can be used to make this type of extension. Suppose $\bar{x}_i = \frac{n\bar{x} - x_i}{n-1}$ represents the sample average of the same data set but with the data point x_i deleted. Let $\bar{x}_{JK}$ represent the mean of the n new values $\bar{x}_i$. The jackknife estimate of the standard deviation then is $\hat{\sigma}_{JK} = \sqrt{\frac{(n-1) \sum_{i=1}^n (\bar{x}_i - \bar{x}_{JK})^2}{n}}$, and it is easy to show that $\bar{x}_{JK} = \bar{x}$ and $\sigma_{JK} = \sigma(\bar{x})$.

The jackknife was originally invented by Quenouille (1949) under another name and with the purpose to obtain a nonparametric estimate of bias associated with some types of estimators. Bias can be formally defined as $BIAS \equiv E_F\{\vartheta(\hat{F})\} - \vartheta(F)$ where E_F denotes expectation under the assumption that n quantities were drawn from an unknown probability distribution F ; $\hat{\vartheta} = \vartheta(\hat{F})$ is the estimate of a parameter of interest with $\hat{F}$ representing the empirical probability distribution. Quenouille’s bias estimate (*cf.* Efron 1982, p. 5) is based on sequentially deleting values x_i from a sample of n values to generate different empirical probability distributions $\hat{F}_i$ each based on $(n-1)$ values resulting in the estimates $\hat{\vartheta}_i = \vartheta(\hat{F}_i)$. Writing $\bar{\vartheta} = \frac{\sum_{i=1}^n \hat{\vartheta}_i}{n}$, Quenouille’s bias estimate then becomes $(n-1) \cdot (\bar{\vartheta} - \hat{\vartheta})$, and the bias-corrected “jackknifed estimate” of ϑ is $\tilde{\vartheta} = n\bar{\vartheta} - (n-1)\hat{\vartheta}$. This estimate is either unbiased or less biased than $\hat{\vartheta}$.

Examples of Jackknife Applications in Regional Mineral Potential Estimation

A very simple example of application to a hypothetical mineral resource estimation problem is as follows. Suppose that half of a study area consisting of 100 equal-area cells is well-explored and that two mineral deposits have been discovered

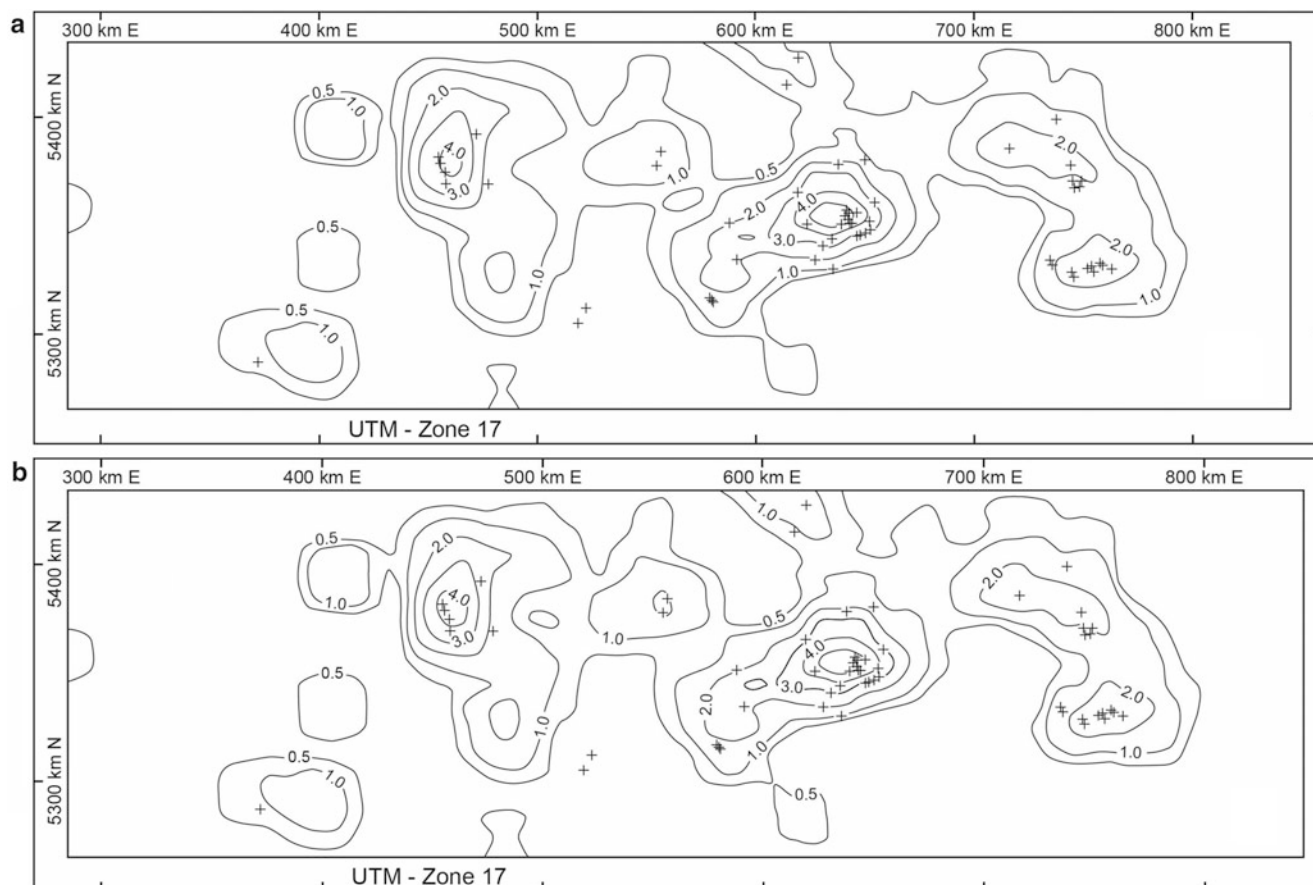
in it. This 50-cell “control area” can be used to estimate the number of undiscovered deposits in the 50-cell “target area” representing the other half of the study area that is relatively unexplored. Suppose further that the mineral deposits of interest are contained within a “favorable” rock type that occurs in 5 cells of the control area as well as in 5 cells of the target area. Because 2 of these 5 cells in the control area are known to contain a known ore deposit of interest, it is reasonable to assume that the target area would contain 2 undiscovered deposits as well. Thus, the entire study area probably contains 4 deposits. Obviously, it would not be good to assume that it contains only two deposits which constitutes a severely biased estimate.

Application of the “leave-one-out” jackknife method to the 50-cell study area results in two biased estimates that are both equal to a single deposit. This also translates into the biased prediction of two deposits in the entire study area. By using the jackknife theory summarized in the preceding section, it is quickly shown that the jackknifed bias in this biased estimate also is equal to two deposits. The use of the jackknife for the

study area, therefore, results in four deposits of which the two undiscovered deposits would occur within the cells with favorable rock type in the target area.

A second, more realistic example is shown in Fig. 6. The top diagram (Fig. 6a) is for the same study area that was used for Fig. 3, but the contour values in Fig. 6 are for smaller ($30 \text{ km} \times 30 \text{ km}$) squares. The bottom diagram (Fig. 6b) shows a result obtained by using the jackknife method that produces nearly the same pattern. For both Figs. 3 and 6a, the contour values were obtained by multiple regression in which the dependent variable was the amount of copper per cell and the explanatory variables consisted of lithological composition and geophysical data as discussed previously.

For the application of the jackknife (Agterberg 1973), the 35 control cells in the study area were divided into 7 groups each consisting of 5 cells, and these groups were deleted successively to obtain the 7 biased estimates required, with the final jackknife estimate shown in Fig. 6b, which is almost the same as Fig. 6a. The fact that two different approaches produced similar results indicates that both



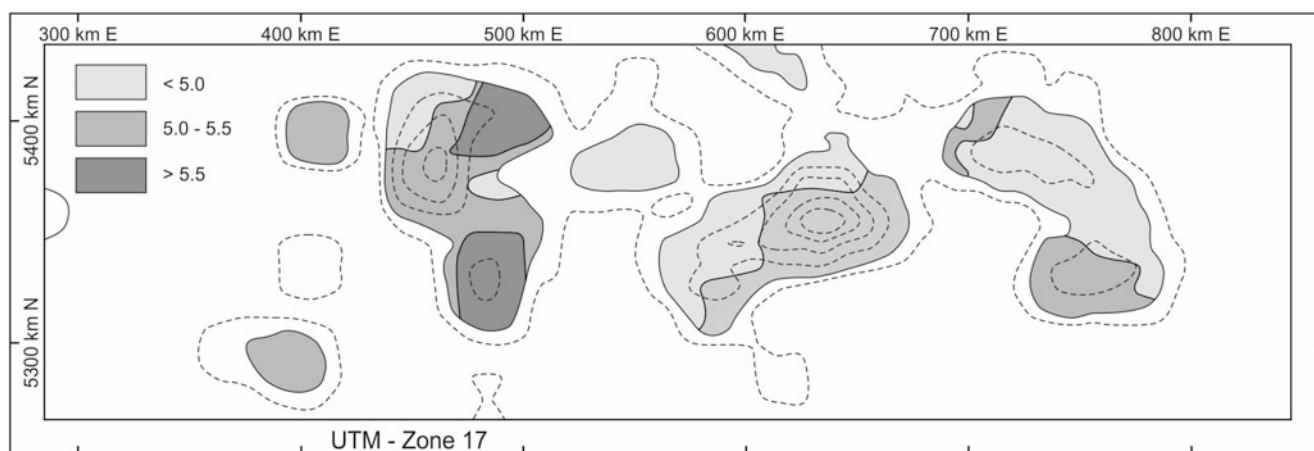
Predictive Geologic Mapping and Mineral Exploration, Fig. 6 (a and b) Comparison of copper potential maps for ($30 \text{ km} \times 30 \text{ km}$) cells in the Abitibi area derived by (a) ordinary multiple regression using a

“control area” with $F = 1.828$, and (b) the jackknife method without the assumption of existence of a “control area” (Source: Agterberg 1973, Fig. 4)

Predictive Geologic Mapping and Mineral Exploration,

Table 1 Comparison of ten (10 km × 10 km) copper cell probabilities in Abitibi area (for location coordinates, see Agterberg et al. 1972)

Location	p(original)	s.d.(1)	p(jackknife)	s.d.(2)
32/62	0.45	0.50	0.44	0.08
16/58	0.33	0.47	0.32	0.04
17/58	0.39	0.49	0.38	0.05
18/58	0.01	0.10	0.00	0.06
16/59	0.35	0.48	0.36	0.04
17/59	0.33	0.47	0.33	0.14
18/59	0.37	0.48	0.40	0.13
16/60	0.43	0.50	0.47	0.04
17/60	0.06	0.24	0.00	0.06
18/60	0.03	0.17	−0.06	0.08



Predictive Geologic Mapping and Mineral Exploration, Fig. 7 Expected total amounts of copper for ore-rich cells predicted in Fig. 6a. Contour values are logarithms (base 10) for (30 km × 30 km) unit cells (Source: Agterberg 1973, Fig. 5a)

methods of prediction of undiscovered copper resources are probably valid. The two methods also yield similar estimates for the probabilities of cells as is illustrated in Table 1 (see Agterberg 2014, for the exact locations of these cells). Standard deviations equal to $\{p \cdot (1 - p)\}^{0.5}$ estimated by the jackknife method are shown in the last column of Table 1. It is noted that one of the estimated jackknife probabilities is negative, although it is not significantly less than 0. Problems of this type can be avoided by using logistic regression. However, the general linear model used to estimate probabilities in this kind of application can have relative advantages (*cf.* Agterberg 2014).

Estimating Total Amounts of Metal from Mineral Potential Maps

In a study described in Agterberg (1973), the Abitibi study area contained 35 (10 km × 10 km) cells with one or more large copper deposits. Two types of multiple regressions were carried out with the same explanatory variables. First the

dependent variable was set equal to 1 in the 35 control cells, and then it was set equal to a logarithmic measure (base 10) of short tons of copper per control cell. Suppose that estimated values for the first regression are written as P_i and those for the second regression as Y_i . Both sets of values were added for overlapping square blocks of cells to obtain estimates of expected values (30 km × 30 km) unit cells. In Fig. 7 the ratio $Y_i^* = Y_i / P_i$ is shown as a pattern that is superimposed on the pattern for the P_i values only. The values of Y_i^* cannot be estimated when Y_i and P_i are both close to zero. Little is known about the precision of Y_i^* for $P_i \geq 0.5$. These values (Y_i^*) should be transformed into estimated amounts of copper per cell here written as X_i . Because of the extreme positive skewness of the size-frequency distribution for amounts of copper per cell (X_i), antilogs (base 10) of the values of Y_i^* as observed in control cells only were multiplied by the constant $c = \sum X_i / \sum 10^{Y_i}$ in order to reduce bias under the assumption of approximate lognormality. The pattern of Fig. 7 is useful as a suggested outline of subareas where the largest volcanogenic massive sulfide deposits are more likely to occur.

Weights-of-Evidence Modeling

Weights-of-evidence modeling is a technique to predict the occurrence of discrete events such as mineral deposits, landslides, or earthquakes from digital geoscience map data. The statistical approach in this method was originally based on the approach taken by Spiegelhalter and Knill-Jones (1984) in their medical GLADYS expert system (*cf.* Agterberg 1989). WofE became widely used by mineral exploration companies and other organizations after the publication of Bonham-Carter (1994)'s book entitled *Geographic information systems for geoscientists: modeling with GIS*. Since the approach is covered in other chapters, the discussion in this section will be restricted to basic principles and a single example of an early application to gold deposits in Nova Scotia (Bonham-Carter et al. 1988).

Suppose that occurrences of mineralization, coded as 1 for presence and 0 for absence in a map pattern called D , are to be related to a single map layer that is available in digital form. When there is a single map pattern B , the odds $O(D|B)$ for the occurrence of mineralization if B is present is given by the ratio of the following two expressions of Bayes' rule:

$$P(D|B) = \frac{P(B|D)P(D)}{P(B)}; P(\bar{D}|B) = \frac{P(B|\bar{D})P(\bar{D})}{P(B)} \text{ where the set } \bar{D}$$

represents the complement of D . Consequently, $\ln O(D|B) = \ln O(D) + W^+$ where the positive weight for presence of B is $W^+ = \ln \frac{P(B|D)}{P(B|\bar{D})}$. The negative weight for the absence of B is $W^- = \ln \frac{P(\bar{B}|D)}{P(\bar{B}|\bar{D})}$. When there are two map patterns: $P(D|B \cap C) = \frac{P(B \cap C|D)P(D)}{P(B \cap C)}$; $P(\bar{D}|B \cap C) = \frac{P(B \cap C|\bar{D})P(\bar{D})}{P(B \cap C)}$. Conditional independence of D with respect to B and C implies:

$$P(B \cap C|D) = P(B|D)P(C|D); P(B \cap C|\bar{D}) = P(B|\bar{D})P(C|\bar{D}).$$

Consequently, $P(D|B \cap C) = P(D) \frac{P(B|D)P(C|D)}{P(B \cap C)}$; $P(\bar{D}|B \cap C) = P(\bar{D}) \frac{P(B|\bar{D})P(C|\bar{D})}{P(B \cap C)}$.

From these two equations, it follows that $\frac{P(D|B \cap C)}{P(\bar{D}|B \cap C)} = \frac{P(D)P(B|D)P(C|D)}{P(\bar{D})P(B|\bar{D})P(C|\bar{D})}$. This expression is equivalent to $\ln O(D|B \cap C) = \ln O(D) + W_1^+ + W_2^+$. The posterior logit on the left of this expression is the sum of the prior logit and the weights of the two map layers. The approach is only valid if the map layers are wholly or conditionally independent, and this hypothesis should be tested in applications. The approach can be extended to include other map layers. Various conditional independence tests have been developed (see, e.g., Agterberg 2014). Approximate variances of the weights can be obtained from:

$$\sigma^2(W^+) = \frac{1}{n(bd)} + \frac{1}{n(\bar{b}\bar{d})}; \sigma^2(W^-) = \frac{1}{n(b\bar{d})} + \frac{1}{n(\bar{b}d)}$$

where the denominators of the terms on the right sides represent numbers $n(\cdot)$ of unit cells with map layer and

deposit present or absent. Spiegelhalter and Knill-Jones (1984) had used similar asymptotic expressions for variances of weights in their GLADYS expert system. Using $p(\cdot)$ to denote probabilities, the following contingency table can be created:

$$P = \begin{bmatrix} p(bd) & p(b\bar{d}) \\ p(\bar{b}d) & p(\bar{b}\bar{d}) \end{bmatrix} \equiv \begin{bmatrix} n(bd)/n & n(b\bar{d})/n \\ n(\bar{b}d)/n & n(\bar{b}\bar{d})/n \end{bmatrix}$$

If there is no spatial correlation between deposit points and map pattern:

$$P = \begin{bmatrix} p(b) \cdot p(d) & p(b) \cdot p(\bar{d}) \\ p(\bar{b}) \cdot p(d) & p(\bar{b}) \cdot p(\bar{d}) \end{bmatrix}$$

where $p(b)$ is the proportion of study area occupied by map layer B and $p(d)$ is the total number of deposits divided by total area, and with similar explanations for the other proportions. A chi-square test for goodness of fit can be applied to test the hypothesis that there is no spatial correlation between deposit points and map using pattern. It is equivalent to using the z -test to be described in the next paragraph. The variance of the "contrast" C is:

$$\sigma^2(C) = \frac{1}{n(bd)} + \frac{1}{n(b\bar{d})} + \frac{1}{n(\bar{b}d)} + \frac{1}{n(\bar{b}\bar{d})}.$$

One of the first WofE applications is 68 gold deposits in a study area of 2,591 km² in Nova Scotia (Bonham-Carter et al. 1988). Weights, contrasts, and their standard deviations for nine map patterns used in Bonham-Carter et al. (1988) for this same study area are shown in Table 2. The last column in this table is the "studentized" value of C , for testing the hypothesis that $C = 0$. Values greater than 1.96 indicate that this hypothesis can be rejected at a level of significance $\alpha = 0.05$. For more information on the validity of the conditional independence tests, see Agterberg (2014). For an alternative approach to solve the same type of problem, see Baddeley et al. (2021).

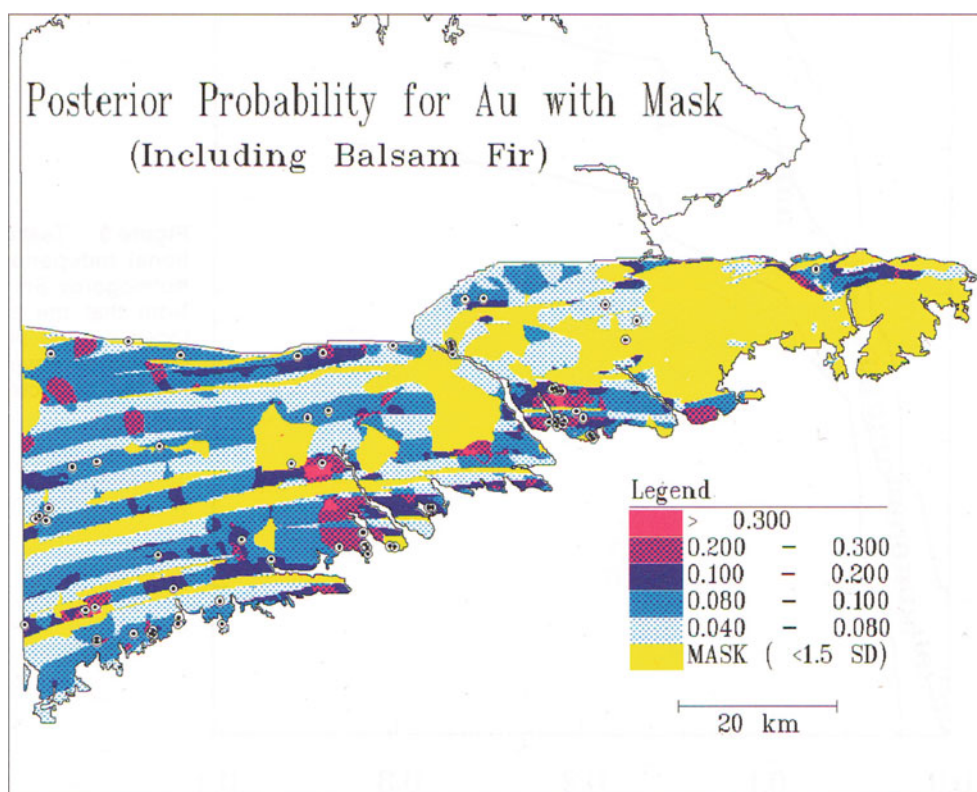
Figure 8 shows the posterior probability map using the weights of Table 2. Some of the areas with the largest posterior probabilities contain known gold deposits but others do not. These other areas are of interest for further exploration, because they may contain undiscovered gold deposits (*cf.* Bonham-Carter et al. 1990). Gold production and reserved figures were available for the 68 gold deposits used as input for this study. Figure 9 is a plot of posterior probability against cumulative area showing gold mines that were producing in 1990 as circles whose radii reflect magnitudes of production in 1990. There appears to be a positive correlation between production and posterior probability.

Predictive Geologic Mapping and Mineral Exploration, Table 2 Weights, contrasts and their standard deviations for predictor map shown in Fig. 8 (Source: Bonham-Carter et al. 1990, Table 5.1)

	W^+	$\sigma(W^-)$	W^-	$\sigma(W^-)$	C	$\sigma(C)$	$C/\sigma(c)$
Goldenville Fm	0.3085	0.1280	-1.4689	0.4484	1.7774	0.4663	3.8117
Anticline axes	0.5452	0.1443	-0.7735	0.2370	1.3187	0.2775	4.7521
Au, biogeochem.	0.9045	0.2100	-0.2812	0.1521	1.1856	0.2593	4.5725
Lake sed. Signature	1.0047	0.3263	-0.1037	0.1327	1.1084	0.3523	3.1462
Golden-Hal contact	0.3683	0.1744	-0.2685	0.1730	0.6368	0.2457	2.5918
Granite contact	0.3419	0.2932	-0.0562	0.1351	0.3981	0.3228	1.2332
NW lineaments	-0.0185	0.2453	0.0062	0.1417	-0.0247	0.2833	0.0872
Halifax Fm.	-1.2406	0.5793	0.1204	0.1257	-1.4610	0.5928	2.4646
Devonian Granite	-1.7360	0.7086	0.1528	0.1248	-1.8888	0.7195	2.6253

Predictive Geologic Mapping and Mineral Exploration,

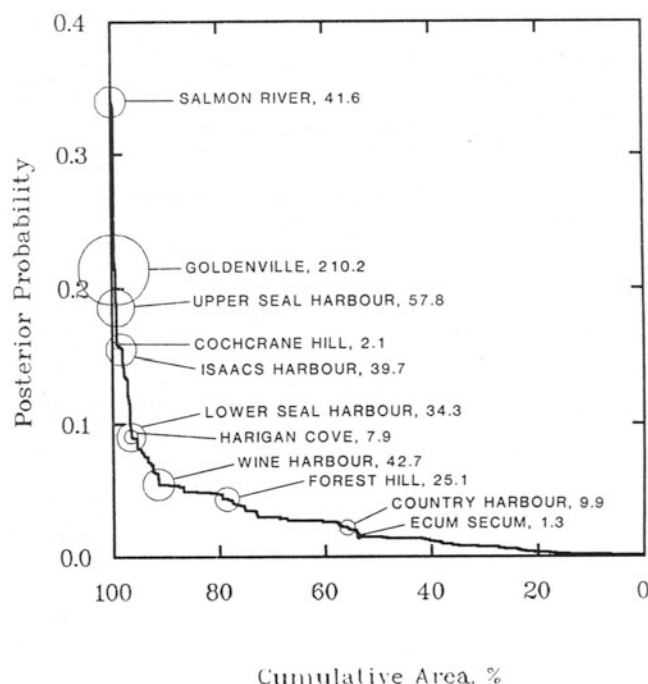
Fig. 8 Posterior probability map corresponding to Table 1 for gold deposits in Meguma Terrane, eastern mainland Nova Scotia (Source: Bonham-Carter et al. 1990, Fig. 1)



Local Singularity Analysis

This last section is devoted to a relatively new method called “local singularity analysis” developed by Cheng (1999, 2005, 2008) for geochemical or other data systematically collected across large study areas. Local singularity analysis, which has become important for helping to predict occurrences of hidden ore deposits, is based on the multifractal equation $x = c \cdot \epsilon^{\alpha-E}$ where x represents the element concentration value, c is a constant, α is the singularity, ϵ is a

normalized distance measure such as square grid cell edge, and E is the Euclidian dimension. The theory of singularity is given in other chapters. Here it will be assumed that $E = 2$ and that chemical element concentration values are available for small square grid cells of which the spatial autocorrelation function (or variogram) is largely determined by its behavior near the origin, which is difficult to establish by earlier statistical or geostatistical techniques. For a relatively simple discussion of singularity analysis, also see de Mulder et al. (2016).



Predictive Geologic Mapping and Mineral Exploration, Fig. 9 Posterior probability as shown in Fig. 8 plotted against cumulative area, with producing gold mines shown as circles with radii reflecting magnitudes of reported production (Source: Bonham-Carter et al. 1990, Fig. 4)

In Cheng's (1999, 2005) original approach, geochemical or other data collected at sampling points within a study area are subjected to two treatments. The first of these is to construct a contour map by any of the methods such as kriging or inverse distance weighting techniques generally used for this purpose. Secondly, the same data are subjected to local singularity mapping. The local singularity α then is used to enhance the contour map by multiplication of the contour value by the factor $\epsilon^{-\alpha/2}$ where $\epsilon < 1$ represents a length measure. In Cheng's (2005) application to predictive mapping, the factor $\epsilon^{-\alpha/2}$ is greater than 2 in places where there has been local element enrichment or by a factor less than 2 where there has been local depletion. Local singularity mapping can be useful for the detection of geochemical anomalies characterized by local enrichment even if contour maps for representing average variability are not constructed (cf. Cheng and Agterberg 2009; Zuo et al. 2009).

Gejiu Mineral District and Northwestern Zhejiang Province Examples

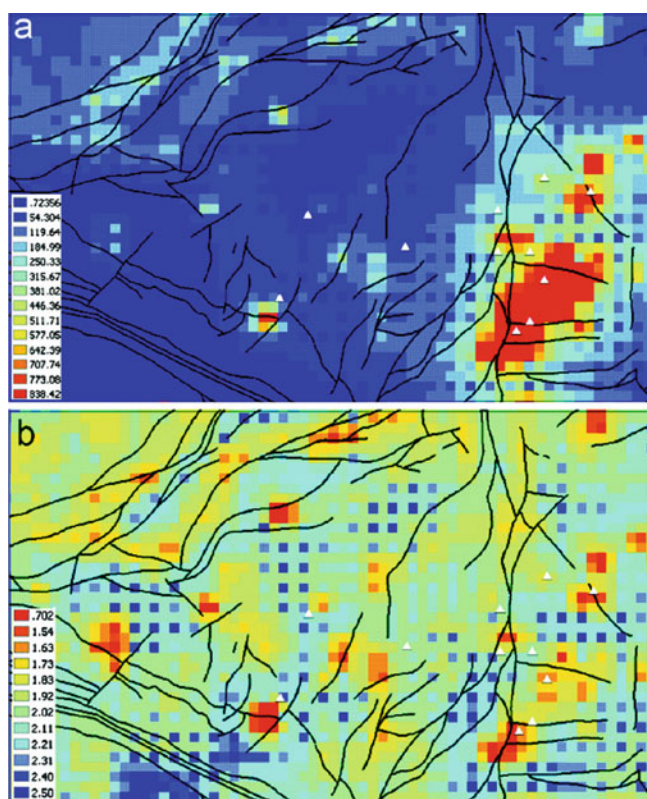
Where the landscape permits this, stream sediments are the preferred sampling medium for reconnaissance geochemical

surveys concerned with mineral exploration (Plant and Hale 1994). During the 1980s and 1990s, government-sponsored reconnaissance surveys covering large parts of Austria, the Canadian Cordillera, China, Germany, South Africa, the UK, and the USA were based on stream sediments (Darnley et al. 1995). These large-scale national projects, which were part of an international geochemical mapping project (Darnley 1995), generated vast amounts of data and continue to be a rich source of information. Cheng and Agterberg (2009) applied singularity analysis to data from about 7800 stream sediment samples collected as part of the Chinese regional geochemistry reconnaissance project (Xie et al. 1997). For illustration, about 1000 stream sediment tin concentration values from the Gejiu area in Yunnan Province were used.

The study in this first example has an area of about 4000 km² and contains 11 large tin deposits. Several of these, including the Laochang and Kafang deposits, which are tin-producing mines with copper extracted as a by-product. These hydrothermal mineral deposits also are enriched in other chemical elements including silver, arsenic, gold, cadmium, cobalt, iron, nickel, lead, and zinc. The application to be described here is restricted to arsenic which is a highly toxic element. Water pollution due to high arsenic, lead, and cadmium concentration values is considered to present one of the most serious health problems especially in underdeveloped areas where mining is the primary industry such as in parts of the Gejiu area. Knowledge of the characteristics of the spatial distribution of ore elements and associated toxic elements in surface media is helpful for the planning of mineral exploration as well as environmental protection strategies.

The Gejiu mineral district (Fig. 10) is located along the suture zone of the Indian Plate and Euro-Asian plates on the southwestern edge of the China subplate, approximately 200 km south of Kunming, capital of Yunnan Province, China. The Gejiu Batholith with outcrop area of about 450 km² is believed to have played an important role in the genesis of the tin deposits (Yu 2002). The ore deposits are concentrated along intersections of NNE-SSW and E-W trending faults. Stream sediment sample locations in the Gejiu area are equally spaced at approximately 2 km in the north-south and east-west directions. Every sample represents a composite of materials from the drainage basin upstream of the collection site (Plant and Hale 1994).

Regional trends were captured in a moving average map of tin concentration values from within square cells measuring 26 km on a side. Several parameters have to be set for use of the inverse distance weighted moving average method (Cheng 2003). In the current application, each square represented the moving average for a square window measuring 26 km on a side with an influence of



Predictive Geologic Mapping and Mineral Exploration, Fig. 10 Map patterns derived from arsenic concentration values in 1,000 stream sediment samples in Gejiu Mineral District, Yunnan Province, China. Large tin deposits are shown as white triangles. Pattern (a) shows distribution of arsenic concentration values using inverse distance weighted moving average; pattern (b) shows distribution of local tin singularities (a). The 3 tin deposits outside the highly polluted mining area shown in pattern (a) are recent discoveries. All 11 large tin deposits occur in places where local anomaly in pattern (b) is less than 2 (Source: Cheng and Agterberg 2009, Fig. 1)

samples decreasing with distance according to a power-law function with exponent set equal to -2 . Original sample locations were 2 km apart both in the north-south and east-west directions. The resulting map (Fig. 10a) shows a large anomaly in the eastern part of the Gejiu area surrounding the most large tin deposits including the mines. The three large tin deposits in the central part of the Gejiu area are recent discoveries that have not yet been mined. To illustrate in more detail how singularities were estimated, see Cheng and Agterberg (2009). The singularities were estimated by fitting straight lines on log-log plots of either the concentration value (x) versus ϵ using $x = c \cdot \epsilon^{-\alpha/2}$ or amount of metal $\mu = x \cdot \epsilon^{-2}$ versus ϵ using $\mu = c \cdot \epsilon^{\alpha}$. Both methods yielded similar estimates of the singularity α . Figure 10b shows results using the first method only. The main difference between the patterns of Fig. 10a and b is that lower and

higher local As singularity values are more evenly distributed across the Gejiu area than the As concentration values themselves. Similar results were obtained for other elements including tin (see Cheng and Agterberg 2009). Clearly, Fig. 10b provides a better guideline for finding unknown large tin deposits than Fig. 10a. On the other hand, Fig. 10a more clearly shows the area in which the streams are highly polluted because of mining pollution over an extended period of time.

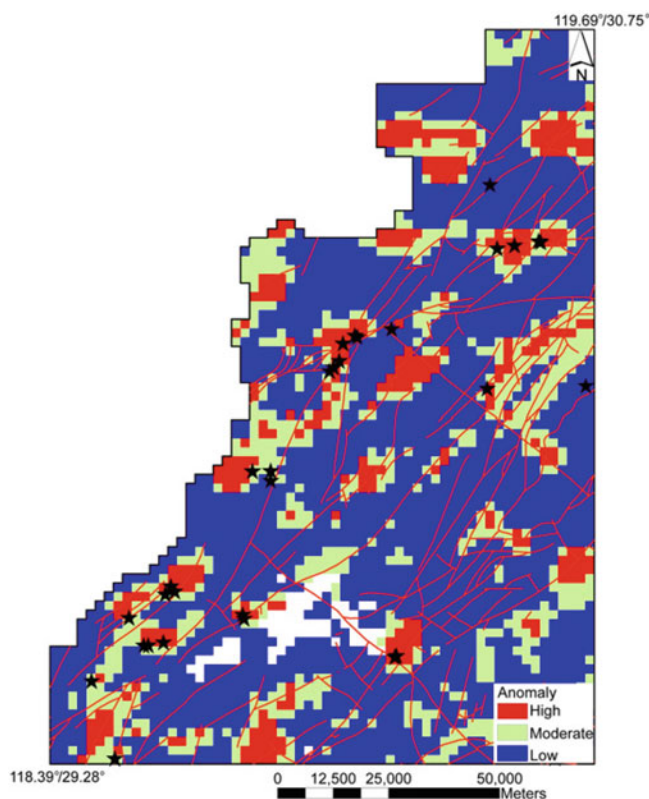
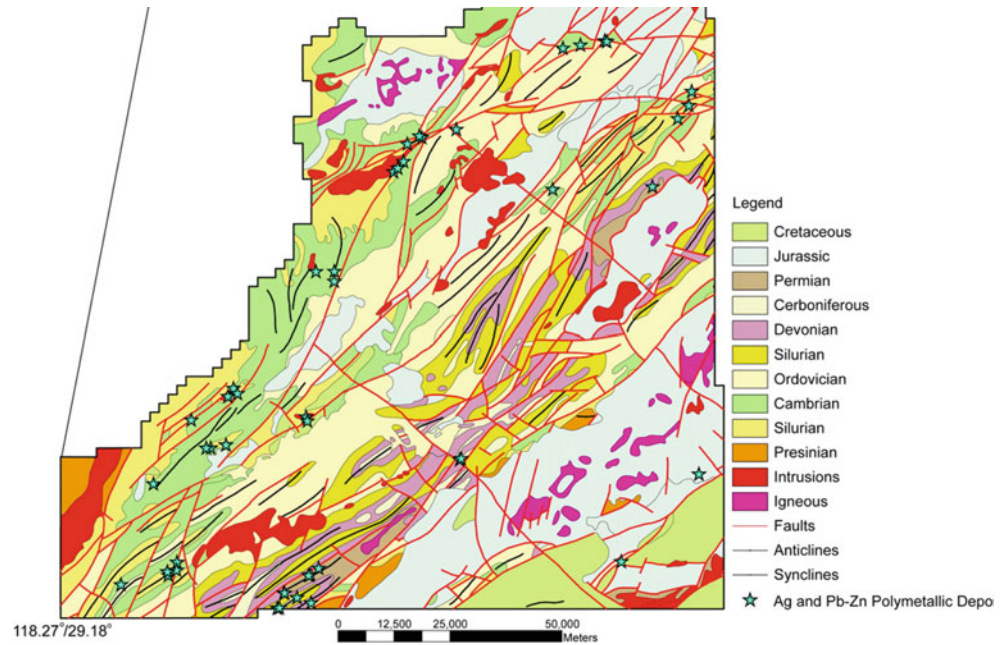
The second example shown in Figs. 11 and 12 (after Xiao et al. 2012) is for lead in northwestern Zhejiang Province, China, that in recent years has become recognized as an important polymetallic mineralization area with the discovery of several moderate to large Ag and Pb-Zn deposits mainly concentrated along the northwestern edge of the study area (Fig. 11), which has been relatively well explored. The pattern of local singularities based on Pb in stream sediment samples is spatially correlated with the known mineral deposits. It can be assumed that similar anomalies in parts of the less explored parts of the area provide new targets for further mineral exploration (Fig. 12). By combining singularities for different chemical elements with one another, spatial correlation between anomalies and mineral deposits can be further increased (Xiao et al. 2012).

Summary and Conclusions

Geologic maps augmented by remote sensing and geophysical and geochemical data provide the main inputs for decision-making in mineral exploration concerned with where to drill boreholes to find hidden ore bodies. In this chapter, the history of predictive geologic mapping with downward extrapolations from the Earth's surface was briefly reviewed. Special attention was paid to the first copper potential map for the Abitibi area on the Canadian Shield published in 1972 for copper-producing mines and prospects publicly available in 1968 using geological maps and geophysical data. During the next 10 years, this area was subjected to intensive prospecting for copper-containing mineral deposits permitting an evaluation of the original copper potential maps. Various statistical methods including logistic regression and the jackknife were used for assessing the uncertainties associated with predicting locations and amounts of copper in square cells of different sizes. In the final sections of this chapter, the relatively well-known method of weights of evidence (WofE) and the more recently developed method of local singularity analysis were briefly discussed with examples of the application using geologic maps and geochemical data.

Predictive Geologic Mapping and Mineral Exploration,

Fig. 11 Geological map of northwestern Zhejiang Province, China (Source: Xiao et al. 2012, Fig. 1)



Predictive Geologic Mapping and Mineral Exploration,
Fig. 12 Raster map for Fig. 11 showing target areas for prospecting for Ag and Pb-Zn deposits delineated by comprehensive singularity anomaly method (cell size is 2 km × 2 km). Known Ag and Pb-Zn deposits are indicated by stars (Source: Xiao et al. 2012)

Cross-References

- Digital Geological Mapping
- Earth Surface Processes
- Fuzzy Inference Systems for Mineral Exploration
- Local Singularity Analysis
- Logistic Regression
- Logistic Regression, Weights of Evidence, and the Modeling Assumption of Conditional Independence
- Multivariate Analysis

Bibliography

- Agterberg FP (1970) Autocorrelation functions in Geology. In: Merriam DF (ed) Geostatistics. Plenum, New York, pp 113–142
- Agterberg FP (1973) Probabilistic models to evaluate regional mineral potential. In: Proceedings of the symposium on mathematical methods in the geosciences held at Příbram, Czechoslovakia, pp 3–38
- Agterberg FP (1974) Geomathematics. Elsevier, Amsterdam, 596 p
- Agterberg FP (1984) Reply to comments by Professor John W. Tukey. Math Geol 16:595–600
- Agterberg FP (1989) Computer programs for mineral exploration. Science 245:76–81
- Agterberg FP (2014) Geomathematics: theoretical foundations, applications and future developments. Springer, Heidelberg, 553 p
- Agterberg FP (2021) Aspects of regional and worldwide mineral resource production. <https://doi.org/10.1007/s12583-020-1397-4>
- Agterberg FP, David M (1979) Statistical exploration. In: Weiss A (ed) Computer methods for the 80's. Society of Mining Engineers, New York, pp 90–115

- Agterberg FP, Chung CF, Fabbri AG, Kelly AM, Springer J (1972) Geomathematical evaluation of copper and zinc potential in the Abitibi area on the Canadian Shield. Geological Survey of Canada, Paper 71–11
- Baddeley A, Brown W, Milne R, Nair G, Rakshit S, Lawrence T, Phatek A, Fu SC et al (2021) Optimum thresholding of predictors in mineral prospectively analysis. *Nat Res* 30:923–969
- Bonham-Carter GF (1994) Geographic information systems for geoscientists: modelling with GIS. Pergamon, Oxford, p 398
- Bonham-Carter GF, Agterberg FP, Wright DF (1988) Integration of geological data sets for gold exploration in Nova Scotia. *Photogram Eng Remote Sens* 54:1585–1592
- Bonham-Carter GF, Agterberg FP, Wright DF (1990) Weights of evidence modelling: a new approach to mapping mineral potential. *Geol Surv Can Pap* 89-9:171–183
- Cheng, Q. (1999). Multifractality and spatial statistics. *Comput Geosci* 10:1–13
- Cheng Q (2003) GeoData Analysis System (GeoDAS) for mineral exploration and environmental assessment, user's guide. York University, Toronto, 268 p
- Cheng Q (2005) A new model for incorporating spatial association and singularity in interpolation of exploratory data. In: Leuangthong D, Deutsch CV (eds) *Geostatistics Banff 2004*. Springer, Dordrecht, pp 1017–1025
- Cheng Q (2008) Non-linear theory and power-law models for information integration and mineral resources quantitative assessments. In: Bonham-Carter GF, Cheng Q (eds) *Progress in geomathematics*. Springer, Heidelberg, pp 195–225
- Cheng Q, Agterberg FP (2009) Singularity analysis of ore-mineral and toxic trace elements in stream sediments. *Comput Geosci* 35: 234–244
- Darnley AG (1995) International geochemical mapping – a review. *J Geochem Explor* 55:5–10
- Darnley AG, Garrett RG, Hall GEM (1995) A global geochemical database for environmental and resource management: recommendations for international geochemical mapping. Final Rep IGCP Project 259. UNESCO, PA
- de Mulder E, Chen Q, Agterberg F, Goncalves M (2016) New and game-changing developments in geochemical exploration. *Episodes*, pp 70–71
- Draper NR, Smith H (1966) *Applied regression analysis*. Wiley, New York, 280 p
- Efron B (1982) The Jackknife, the Bootstrap and other resampling plans. SIAM, Philadelphia, 93 p
- Gradstein FM, Ogg J, Schmitt MD, Ogg G (eds) (2020) *Geological time scale 2020*, vol 1 & 2. Elsevier, Amsterdam, 1357 p
- Harrison JM (1963) Nature and significance of geological maps. In: Albritton CC Jr (ed) *The fabric of geology*. Addison-Wesley, Cambridge, MA, pp 225–232
- Nieuwenkamp W (1968) *Natuurfilosofie en de geologie van Leopold von Buch*. K. Ned Akad Wet Proc Ser B 71(4):262–278
- Pirajno F (2010) *Hydrothermal processes and mineral systems*. Springer, Heidelberg, 1250 p
- Plant J, Hale M (eds) (1994) *Handbook of exploration geochemistry* 6. Elsevier, Amsterdam
- Quenouille M (1949) Approximate tests of correlation in time series. *J R Stat Soc Ser B* 27:395–449
- Spiegelhalter DJ, Knill-Jones RP (1984) Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. *J R Stat Soc A* 147(1):35–77
- Tukey JW (1970) Some further inputs. In: Merriam DF (ed) *Geostatistics*. Plenum, New York, pp 163–174
- Tukey JW (1972) Discussion of paper by F. P. Agterberg and S. C. Robinson. *Intern Stat Inst Bull* 38:596
- Tukey JW (1984) Comments on “use of spatial analysis in mineral resource evaluation”. *Math Geol* 16:591–594
- Von Buch L (1842) Ueber Granit und Gneiss, vorzüglich in Hinsicht der äusseren Form, mit welcher diese Gebirgsarten auf Erdoberfläche erscheinen. *Abh K Akad Berlin IV/2(18840)*:717–738
- Wegener A (1966) *The origin of continents and oceans*, 4th edn. Dover, New York
- Wellmer FW (1983) Neue Entwicklungen in der Exploration (II), Kosten, Erlöse, Technologien, *Erzmetall* 36(3):124–131
- Westergård AH, Johansson S, Sundius N (1943) *Beskrivning till Kartbladet Lidköping*. Sver Geol Unders Ser Aa 182
- Winchester S (2001) *The map that changed the world*. Viking, Penguin Books, London, 412 p
- Xiao F, Chen J, Zhang Z, Wang C, Wu G, Agterberg FP (2012) Singularity mapping and spatially weighted principal component analysis to identify geochemical anomalies associated with Ag and Pb-Zn polymetallic mineralization in Northwest Zhejiang, China. *J Geochem Explor* 122:90–100
- Xie X, Mu X, Ren T (1997) Geochemical mapping in China. *J Geochem Explor* 60:99–113
- Yu C (2002) Complexity of earth systems – fundamental issues of earth sciences. *J China Univ Geosci* 27:509–519. (in Chinese; English abstract)
- Zuo R, Cheng Q, Agterberg FP, Xia Q (2009) Application of singularity mapping technique to identify local anomalies using stream sediment geochemical data, a case study from Gangdese, Tibet, western China. *J Geochem Explor* 101:225–235

Principal Component Analysis

Alessandra Menafoglio

MOX-Department of Mathematics, Politecnico di Milano, Milano, Italy

Synonyms

Karhunen-Loève decomposition; Principal orthogonal decomposition

Definition

Principal component analysis (PCA) is a statistical technique aimed to explore and reduce the dimensionality of a multivariate dataset. It is based on the key idea of finding a representation of the p -dimensional data through a smaller set of $k < p$ new variables, defined as linear combinations of the original variables. These are obtained as to explain at best the data variability. PCA was first introduced by Pearson (1901); nowadays its formulation has been developed for varied types of Euclidean data, including compositional data and functional data. Extensions to non-Euclidean data are available as well.

PCA as an Optimization Problem

It is given a dataset made of n observations of p variables, organized in a data matrix $\mathbb{X} = [x_{ij}]$ in $\mathbb{R}^{n \times p}$, whose columns are denoted by $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})$. It is assumed that the variables have null sample mean ($\sum_{j=1}^n x_{ij} = 0$); otherwise, the data can be centered with respect to their sample mean prior to apply the PCA. The goal of the PCA is to define a new set of uncorrelated variables $\mathbf{z}_1, \dots, \mathbf{z}_p$, named *principal components* (PCs), linearly related with the original variables $\mathbf{x}_1, \dots, \mathbf{x}_p$, and suitable to effectively represent the data through a reduced representation based on $k < p$ PCs only – namely, through a new data matrix $\mathbb{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_k]$ in $\mathbb{R}^{n \times k}$. The generic principal component $\mathbf{z}_{\cdot j} = (z_{1j}, \dots, z_{nj})^T$ is thus expressed as

$$z_{ij} = \xi_{j1}x_{i1} + \xi_{j2}x_{i2} + \dots + \xi_{jp}x_{ip}, \quad (1)$$

where the z_{ij} is named *principal component score* (or simply *score*) and represents the i -th observation of the j -th principal component, and the weights ξ_{ji} can be organized in a vector $\boldsymbol{\xi}_j = (\xi_{j1}, \dots, \xi_{jp})^T$, which is named *loading vector*. As exemplified in Fig. 1, the vector $\boldsymbol{\xi}_j$ identifies a direction in $\mathbb{R}^p$ along which the data are projected (obtaining the scores z_{ij}); hence, $\boldsymbol{\xi}_j$ is conventionally set to be a unit vector ($\|\boldsymbol{\xi}_j\|^2 = \sum_{\ell=1}^p \xi_{j\ell}^2 = 1$, with $\|\cdot\|$ the Euclidean norm). The PCA thus ultimately identifies k directions in $\mathbb{R}^p$, which together form a new representation space for the data.

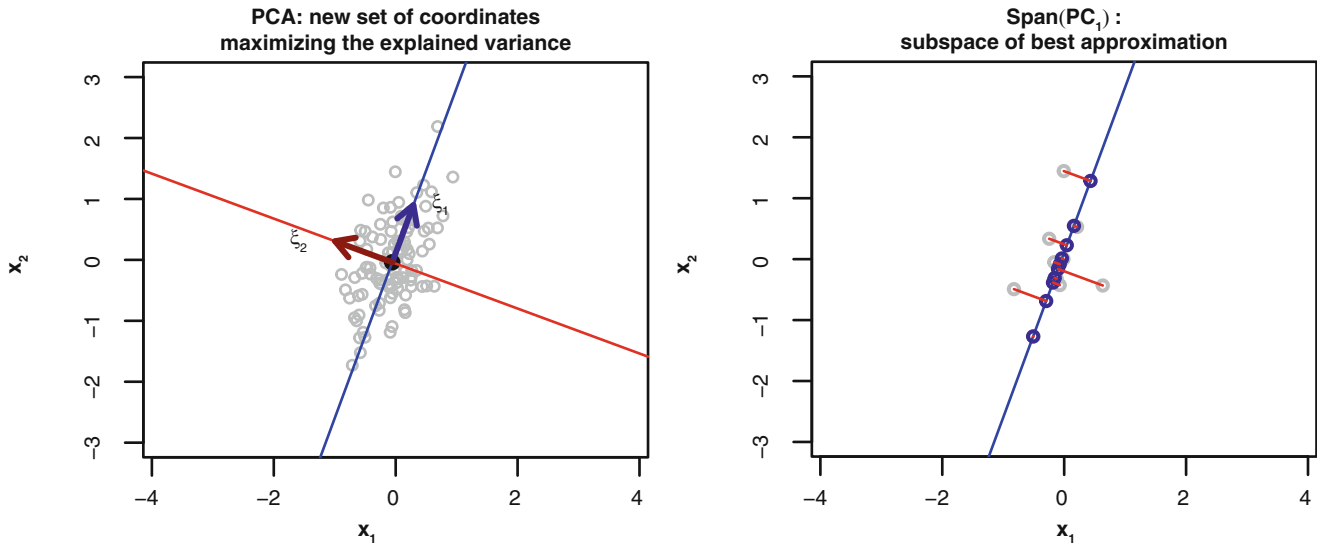
Two requirements underlie the identification of the loading vectors $\boldsymbol{\xi}_j$: (1) the directions are set to be orthogonal to each other; and (2) the variability expressed by the data when projected along these directions is maximized. This leads to find the j -th loading vector as the solution of the optimization problem

$$\arg \min_{\boldsymbol{\xi}_j \in \mathbb{R}^p} \sum_{i=1}^n z_{ij}^2 \quad \text{subject to} \quad \|\boldsymbol{\xi}_j\| = 1, \quad (2)$$

$$\langle \boldsymbol{\xi}_j, \boldsymbol{\xi}_\ell \rangle = 0, \ell < j,$$

where the last orthogonality constraint $\langle \boldsymbol{\xi}_j, \boldsymbol{\xi}_\ell \rangle = 0$ (the symbol $\langle \cdot, \cdot \rangle$ denoting the Euclidean inner product) ensures the first requirement, and it is only imposed for $j > 1$ (i.e., not on the first PC). It can be proved that the optimization problem has a unique solution, which is found from the eigen-decomposition of the covariance matrix of the data $\mathbb{S} = \frac{1}{n-1} \mathbb{X}^T \mathbb{X}$, or, analogously, from the singular value decomposition of the data matrix $\mathbb{X}$ (see Johnson and Wichern (2002)). Here, the variances of the PCs, $\text{Var}(\mathbf{z}_{\cdot j})$, are found as the (ordered) eigenvalues of $\mathbb{S}$, while the loading vectors are the (ordered) eigenvectors of $\mathbb{S}$ (equivalent to the right singular vectors of $\mathbb{X}$).

A strong linear correlation among variables reflects on the presence of PCs associated with very low variability. In this case, the last PCs have sample variance approaching 0. In case of perfect linear relation among variables, only a set of $k < p$ PCs are associated to non-null variance (i.e., non-null eigenvalues of $\mathbb{S}$). In this case, the data matrix $\mathbb{X}$, as well as the covariance matrix $\mathbb{S}$, is not of full rank.



Principal Component Analysis, Fig. 1 Left: Finding a new system of coordinates through PCA to represent a set of bivariate data. Data are reported as gray symbols; directions of the PCs are depicted as straight lines; arrows indicate the loading vectors. The first PC is colored in blue, the second PC in red. Right: Interpretation of PCA as space of best

approximation, for a bivariate dataset (gray symbols). The direction of the first PC ($\text{span}\{\xi_1\}$) is depicted as a blue straight line; projections of the data as blue symbols; reconstruction errors are depicted as red segments. The MSE_1 is the sum of the squares of the lengths of the red segments

Alternative Geometrical Interpretation

The PCA can be alternatively interpreted as a method to build nested spaces of best approximation for the data. Indeed, one can prove that the linear space generated by the loading vectors of the first k PCs, $\text{span}\{\xi_1, \dots, \xi_k\}$, is the k -dimensional space of best approximation for the data, in the sense of minimizing the mean square error of approximation

$$MSE_k = \sum_{i=1}^n \left\| \mathbf{x}_i - \sum_{j=1}^k \langle \mathbf{x}_i, \xi_j \rangle \xi_j \right\|^2, \quad (3)$$

where $\mathbf{x}_i$ denotes the i -th row of the data matrix $\mathbb{X}$, and $\sum_{j=1}^k \langle \mathbf{x}_i, \xi_j \rangle \xi_j$ is the projection of $\mathbf{x}_i$ on $\text{span}\{\xi_1, \dots, \xi_k\}$.

In other terms, having at one's disposal k dimensions to reconstruct the dataset, the first k PCs offer the coordinate system minimizing the reconstruction error (in terms of Euclidean distance MSE_k). Note that, in general, the best reconstruction of dimension 0, in the sense of the MSE, is the sample mean $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$. This alternative viewpoint to

PCA is often taken to generalize PCA to non-Euclidean spaces (e.g., in the principal geodesic analysis, Fletcher et al. (2004)).

Dimensionality Reduction

The PCA can be used to perform a dimensionality reduction of the dataset, by retaining only k out of the p PCs, and then using the reduced data matrix $\mathbb{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_k]$ in $\mathbb{R}^{n,k}$ for further statistical processing. This is typically done by looking for an elbow in the *scree plot*, which depicts, as a function of k , the cumulative variance along the first k PCs, compared to the total variance of the data $\text{Var}_{tot} = \frac{1}{n} \sum_{j=1}^p \sum_{i=1}^n x_{ij}^2$

(an example is given in Fig. 2b). Note that, by construction, the variance of the j -th PCs, $\text{Var}(\mathbf{z}_j)$, is larger than the variances of successive PCs, i.e., $\text{Var}(\mathbf{z}_j) \geq \text{Var}(\mathbf{z}_\ell)$, $\ell > j$, and that $\text{Var}_{tot} = \sum_{j=1}^p \text{Var}(\mathbf{z}_j)$. A full representation of the total vari-

ance Var_{tot} can only be attained by retaining all the p PCs; however, in many cases, a set of $k < p$ PCs may be sufficient to explain a large amount of the data variability, particularly if the original variables are highly correlated. The number k of PCs to retain can then be set by thresholding the proportion of variance explained by the PCs, typical thresholds being 80%, 90%, or 95%.

Interpretation of the principal components is key for their use in practice. For this purpose, a useful representation is provided by the *biplot* (Gabriel 1971), which displays the

projection of the original variables in the Cartesian space defined by the first and second PCs (see Johnson and Wichern (2002)). An example of biplot corresponding to the PCA of a multivariate dataset with $p = 5$ is given in Fig. 2c.

The dimensionality reduction offered by PCA is useful to face the curse of dimensionality affecting data analyses as the number p of measured variables increases. In particular, if $p > n$ (a.k.a. *large p, small n problem*) several statistical techniques cannot be directly applied. In these cases, PCA can be used prior to further statistical processing, to single out a reduced data matrix $\mathbb{Z}$ of dimension $k \leq n$. These include, e.g., regression analysis in the presence of multicollinearity (principal component regression), where PCA is applied to the set of predictors. Other techniques which may benefit from dimensionality reduction to stabilize estimations or speed up computations are, in a spatial context, co-kriging and stochastic co-simulation (Chiles and Delfiner 2009).

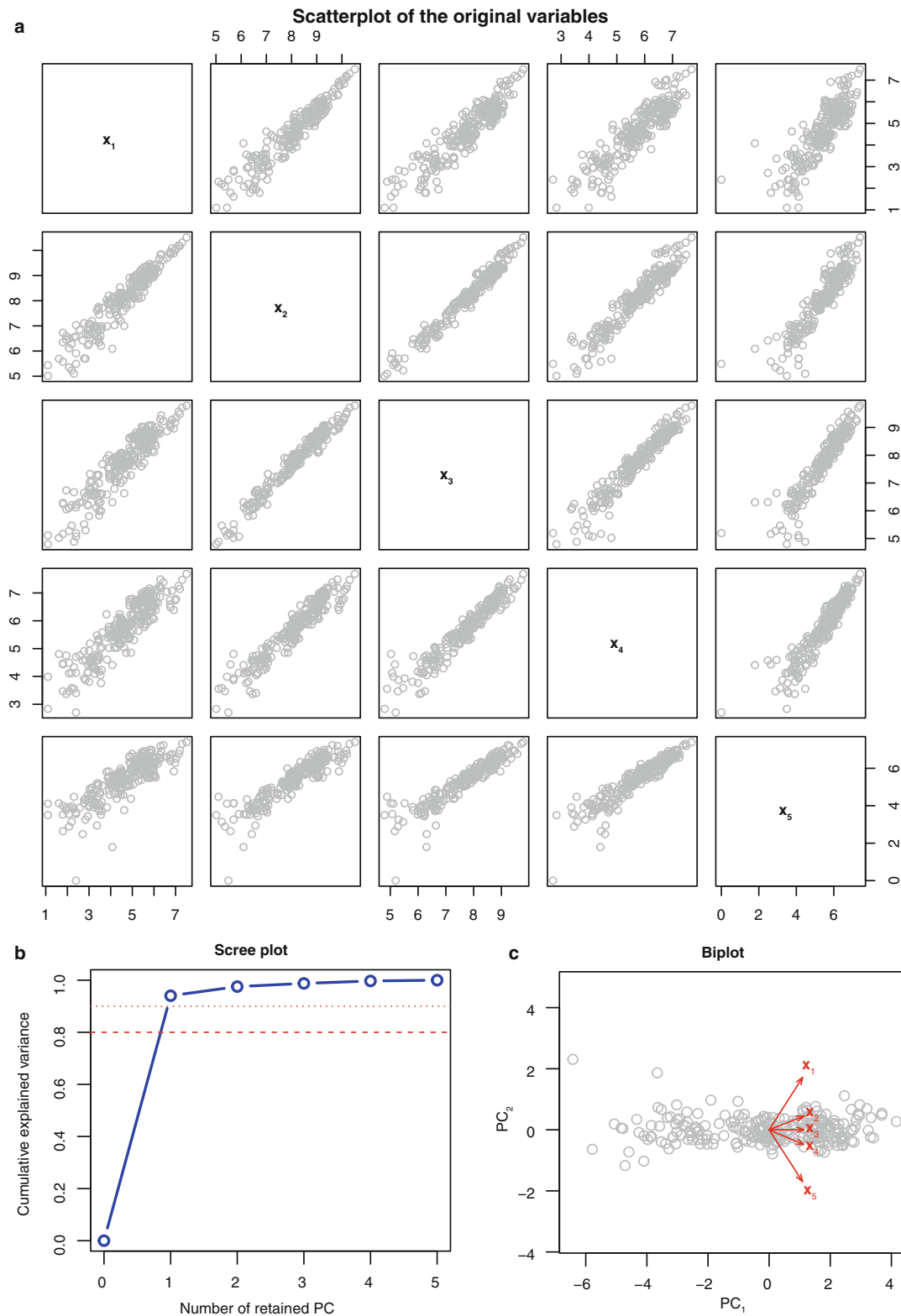
Extensions

Several extensions of PCA to linear and nonlinear spaces are available. These include, among others, the principal component analysis for compositional datasets based on the Aitchison geometry (Pawlowsky-Glahn et al. 2015); the functional principal component analysis for functional data (e.g., curves or images) when these are embedded in Hilbert spaces (Ramsay and Silverman 2005); and the principal geodesic analysis for shape data (Fletcher et al. 2004).

Different representations can be used to perform dimensionality reduction of a multivariate dataset, while attaining slightly different maximization tasks. Among these, we mention independent component analysis (ICA; see Hyvriinen (2013)), which identifies independent variables; Fisher's canonical decomposition (see Johnson and Wichern (2002)), which identifies uncorrelated directions of maximum discrimination among groups in a classification setting; and multi-dimensional scaling (MDS; see Johnson and Wichern (2002)), which identifies a best approximation space in the sense of respecting the mutual distances among data points. PCA, however, unlike these alternatives, returns a *bi-orthogonal decomposition*, in the sense that the loading vectors are orthogonal, and the scores are uncorrelated, by virtue of the Karhunen-Loeve theorem. PCA is part of a broader class of statistical methods for the identification of latent factors, named factor analysis (see Johnson and Wichern (2002)).

Summary

Principal component analysis (PCA) allows one to reduce the dimensionality of a multivariate dataset, by finding directions of maximum variability of the data or, equivalently, nested



Principal Component Analysis, Fig. 2 Example of PCA of a multivariate dataset ($p = 5$). **(a)** Scatterplot of the original variables, showing significant correlation among variables. **(b)** Scree plot. Typical thresholds set at 80% and 90% of explain variability (red horizontal lines)

suggest to select $k = 1$. **(c)** Biplot. Data are reported in the coordinate system of the first and second PC; red arrows indicate the directions of the original variables

spaces of best approximation. The representation offered by PCA can be useful for further statistical processing, to face the curse of dimensionality affecting data analysis as the number p of measured variables increases. The geometrical perspective upon which the PCA is rooted enables its generalization to complex situations, pertaining constrained data (compositional, directional, spherical data) or infinite-dimensional data (functional, distributional data).

Cross-References

- [Compositional Data](#)
- [Correlation Coefficient](#)
- [Eigenvalues and Eigenvectors](#)
- [Exploratory Data Analysis](#)
- [Independent Component Analysis](#)
- [Multidimensional Scaling](#)
- [Q-Mode Factor Analysis](#)
- [R-Mode Factor Analysis](#)
- [Variance](#)

Bibliography

- Chiles J-P, Delfiner P (2009) Geostatistics: modeling spatial uncertainty, vol 497. Wiley
- Fletcher P, Lu C, Pizer S, Joshi S (2004) Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Trans Med Imaging* 23:995–1005
- Gabriel KR (1971) The biplot graphic display of matrices with application to principal component analysis. *Biometrika* 58(3):453–467
- Hyvärinen A (2013) Independent component analysis: recent advances. *Philos Trans R Soc A Math Phys Eng Sci* 371(1984):20110534
- Johnson R, Wichern D (2002) Applied multivariate statistical analysis, 5th edn. Prentice Hall
- Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015) Modeling and analysis of compositional data. Wiley, Chichester
- Pearson KF (1901) LIII. On lines and planes of closest fit to systems of points in space. *Lond Edinb Dublin Philos Mag J Sci* 2(11):559–572
- Ramsay J, Silverman BW (2005) Functional data analysis. Springer, New York

Probability Density Function

Abraão D. C. Nascimento
Universidade Federal de Pernambuco, Recife, Brazil

Synonyms

Density; Joint probability density function

Definition

This chapter aims to furnish the reader with knowledge on the probability density function (pdf). It is assumed as a prerequisite a course in elementary calculus and set theories.

Probability Measure

First consider the definition of two concepts: a σ -field of a set and a probability measure. Let Ω be the set of all outcomes formulated from a real random phenomenon, e.g., measurement of vegetation index from remote sensing radar (Maus et al. 2019) and description of intensities in synthetic aperture radar imagery (Ferreira and Nascimento 2020; Yue et al. 2021). A nonempty collection of subsets drawn from Ω , say $\mathcal{A}$, is said to be a σ -field of Ω if $\mathcal{A}$ satisfies two conditions (Hoel et al. 1973):

- (i) If $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$.
- (ii) If $A_i \in \mathcal{A}$ for $i = 1, 2, \dots$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$ and $\bigcap_{i=1}^{\infty} A_i \in \mathcal{A}$, where A^c is the complement of a set A (i.e., the elements not in A) and $\cup$ and $\cap$ represent the union and intersection operations, respectively.
- (iii) A probability measure, say $\Pr(\cdot)$, on a σ -field of subsets $\mathcal{A}$ of a set Ω is a real-valued function having domain $\mathcal{A}$ satisfying the following properties (Hoel et al. 1973):
- (iv) $\Pr(\Omega) = 1$.
- (v) $\Pr(A) \geq 0$ for all $A \in \mathcal{A}$.
- (vi) If $\{A_n; n = 1, 2, \dots\}$ are mutually disjoint sets in $\mathcal{A}$, then

$$\Pr\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{i=1}^{\infty} \Pr(A_n).$$

From the previous system, various properties can be derived, e.g.:

- (vii) (Boole's inequality) If $\{A_i; i = 1, 2, \dots, n\}$ are any sets, then

$$\Pr\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n \Pr(A_i).$$

- (viii) $\Pr(A_1 \cap A_2 \cap \dots \cap A_n) = \Pr(A_1)\Pr(A_2|A_1)\Pr(A_3|A_1 \cap A_2) \dots \times \Pr(A_n|A_1 \cap \dots \cap A_{n-1})$,

$\forall A_1, \dots, A_n \in \mathcal{A}$, and $\forall n = 1, 2, \dots$, where $\Pr(A_i|B) := \Pr(A_i \cap B)/\Pr(B)$ for $\Pr(B) > 0$ is the conditional probability, one of the most important concepts of probability theory. From the latter, one can define the notion of *independence of two sets* that says: two sets A and $B \in \mathcal{A}$ are independent if $\Pr(A \cap B) = \Pr(A)\Pr(B)$.

- (ix) (Bayes' formula) Let $\{A_i; i = 1, 2, \dots\}$ be a sequence of events that represent a partition of Ω , then

$$\Pr(A_i|B) = \frac{\Pr(A_i)\Pr(B|A_i)}{\sum_{j=1}^{\infty} \Pr(A_j)\Pr(B|A_j)}, \quad \forall B \in \mathcal{A}.$$

From on now, a *probability space*, denoted by $(\Omega, \mathcal{A}, \Pr)$, is a system that approaches a set of events Ω , a σ -field of subsets $\mathcal{A}$, and a probability measure $\Pr : \mathcal{A} \mapsto (0, 1)$.

Random Variables and Vectors

A mapping $X : \Omega \mapsto \mathcal{X} \subset \mathbb{R}$ is said to be a *random variable* (RV) if:

- (x) It is a real-valued function defined on a sample space on a collection of subsets having defined probability.
 (xi) For every Borel set $B \subset \mathbb{R}$, $\{w \in \Omega : X(w) \in B\} \in \mathcal{A}$, where Borel set represents any set in a topological space which is obtained from union, intersection, and complement operations of open or closed sets. Such sets are crucial in the measure theory because any measure formulated on the open or closed sets of a pre-specified space needs to be defined on all Borel sets of such space.

As an extension of X , $\mathbf{X} = (X_1, \dots, X_p)^T : \Omega \mapsto \mathcal{X} \subset \mathbb{R}^p$ is a random vector if X_i is random variable for $i = 1, \dots, p$, where $(\cdot)^T$ is the transposition operator.

In practical terms, an RV represents a numerical value defined from the outcome on the random experiment. Mathematically, it requires $\Pr(X \leq x)$ for all $x \in \mathbb{R}$ to be well-defined or, equivalently, $\{w \in \Omega : X(w) \leq x\}$ belongs to $\mathcal{A}$. Thus, it is necessary to define the cumulative distribution function (cdf) of an RV. Let X be an RV defined on $(\Omega, \mathcal{A}, \Pr)$, its cdf, say $F_X : \mathcal{X} \mapsto [0, 1]$, is given by

$$F_X(x) = \Pr(X \leq x).$$

The cdf is nondecreasing and continuous from the right, satisfying $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$. The cdf plays a crucial role to compute probability of events, e.g., $\Pr(a < X \leq b) = F_X(b) - F_X(a)$ for $a \leq b$. For random vectors, the joint cumulative distribution function (jcdf) of $\mathbf{X}$, say $F_X : \mathcal{X} \mapsto [0, 1]$, is given by

$$F_X(x_1, \dots, x_p) = \Pr(X_1 \leq x_1, \dots, X_p \leq x_p).$$

The jcdf is non-decreasing for x_i (for all $i = 1, \dots, p$) in $(x_1, \dots, x_{i-1}, x_i, \dots, x_p)$, i.e., if $x_{i_1} < x_{i_2}$, then

$$F_X(x_1, \dots, x_{i_m}, \dots, x_p) \leq F_X(x_1, \dots, x_{i_2}, \dots, x_p).$$

The jcdf is continuous from the right for x_i , i.e., if $x_{i_1} \downarrow x_{i_m}$ for $m \rightarrow \infty$,

$$F_X(x_1, \dots, x_{i_m}, \dots, x_p) \downarrow F_X(x_1, \dots, x_{i_1}, \dots, x_p) \text{ for}$$

$$m \rightarrow \infty.$$

Further

$$\lim_{\forall i, x_i \rightarrow -\infty} F_X(x_1, \dots, x_i, \dots, x_p) = 0$$

$$\text{and } \lim_{\forall i, x_i \rightarrow \infty} F_X(x_1, \dots, x_i, \dots, x_p) = 1.$$

Probability Density Function

The nature of the under study phenomenon determines the kind of variable to be used, discrete, or continuous. Let X be a real-valued RV on $(\Omega, \mathcal{A}, \Pr)$, then

- (xii) X is a discrete RV if it has as both domain Ω and range $\mathcal{X}$ a finite or countably infinite subset $\{x_1, x_2, \dots\} \subset \mathbb{R}$ such that $\{w \in \Omega : X(w) = x_i\} \in \mathcal{A}$ for all $i = 1, 2, \dots$.
 (xiii) X is a continuous RV if it has $\mathcal{X} \subset \mathbb{R}$ such that $\Pr(\{w \in \Omega : X(w) = x\}) = 0$ for $-\infty < x < \infty$.

From now on, this discussion will focus on the continuous case.

The cdf of a continuous RV is often defined in terms of its probability density function (pdf). The pdf of X is a nonnegative function, say $f_X : \mathcal{X} \mapsto [0, \infty)$, such that

$$\int_{-\infty}^{\infty} f_X(x) dx = \int_{\mathcal{X}} f_X(x) dx = 1.$$

Note that $F_X(x)$ and $\Pr(X \in B)$ can be rewritten as

$$F_X(x) = \int_{-\infty}^x f_X(t) dt \text{ and } \Pr(X \in B) = \int_B f_X(t) dt,$$

respectively. The pdf can also be deduced from $f_X(x) = dF_X(x)/dx$. For random vectors, the joint probability density function (jpdf) of $\mathbf{X}$ is a nonnegative function, say $f_X : \mathcal{X} \mapsto [0, \infty)$, such that

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_X(\mathbf{t}) dt_1 \dots dt_p = \int_{\mathcal{X}} f_X(\mathbf{t}) dt_1 \dots dt_p.$$

Note that $F_X(x)$ and $\Pr(\mathbf{X} \in \mathbf{B})$ for $\mathbf{B} \subset \mathcal{X}$ can be explicated as

$$F_X(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_p} f_X(\mathbf{t}) dt_1 \cdots dt_p \quad \text{and}$$

$$\Pr(X \in \mathbf{B}) = \int \cdots \int_{\mathbf{B}} f_X(\mathbf{t}) dt_1 \cdots dt_p.$$

Analogous to the univariate counterpart, the jpdf arises from $f_X(\mathbf{x}) = d^p F_X(\mathbf{x})/dx_1 \cdots dx_p$.

Some Statistical and Computational Essays

Now we are in a position of illustrating the pdf shape of some univariate and multivariate distributions. Hundreds of continuous univariate distributions have been extensively used for modeling phenomena in various areas of science (cf. Cordeiro et al. 2020). The most employed one is the normal (or Gaussian) distribution, having its univariate version support in $\mathbb{R}$ and as parameters the mean $\mu \in \mathbb{R}$ and standard deviate $\sigma > 0$ and its p variate counterpart domain in $\mathbb{R}^p$ and parameters, the mean vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance matrix $\boldsymbol{\Sigma} \succcurlyeq \mathbf{0}$, where $\succcurlyeq \mathbf{0}$ represents the nonnegative definite condition. Their pdfs are presented in Table 1. This case is denoted as $X \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the multivariate case and $X \sim N(\mu, \sigma^2)$ for the univariate. An alternative distribution to the N_p law is the t -student distribution. Similar to the normal, this law is symmetric having pdf with heavy tails driven by the parameter ν called degree of freedom (df), and it is denoted as $X \sim t_\nu$. The smaller the parameter ν is, the heavier the tail is. Figure 1a illustrates the relation between $N(\mu, \sigma^2)$ and t_ν distributions. Note that the latter tends to the former when increasing ν and the t_2 law has the most heavy tail. From the relation between these two distributions, the most used t and Hotelling's T^2 tests for the null hypotheses $\mathcal{H}_{01}: \mu = [\mu_0 \in \mathbb{R}]$ and $\mathcal{H}_{02}: \boldsymbol{\mu} = [\boldsymbol{\mu}_0 \in \mathbb{R}^p]$ arise from, respectively:

- (Univariate case: the t test) Let $X_1, \dots, X_n$ be an independent and identically distributed (iid) sample of size n

drawn from $X \sim N(\mu, \sigma^2)$; then the appropriate test statistic (called t statistic) for $\mathcal{H}_{01}$ is

$$S_t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \mathcal{H}_{01} t_{n-1} \quad \text{or} \quad S_t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1},$$

where $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ is the sample mean, $S^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ is the sample variance, and “ $\sim \mathcal{H}_{01}$ ” represents a variable distributed as a distribution law under the hypothesis $\mathcal{H}_{01}$. As a decision rule, we reject $\mathcal{H}_{01}$ when $|S_t|$ is large.

- (Multivariate case: the Hotelling's T^2 test) Let $X_1, \dots, X_n$ be an iid sample of size n obtained from $X \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; then the appropriate test statistic for $\mathcal{H}_{02}$ (named Hotelling's T^2 statistic) is

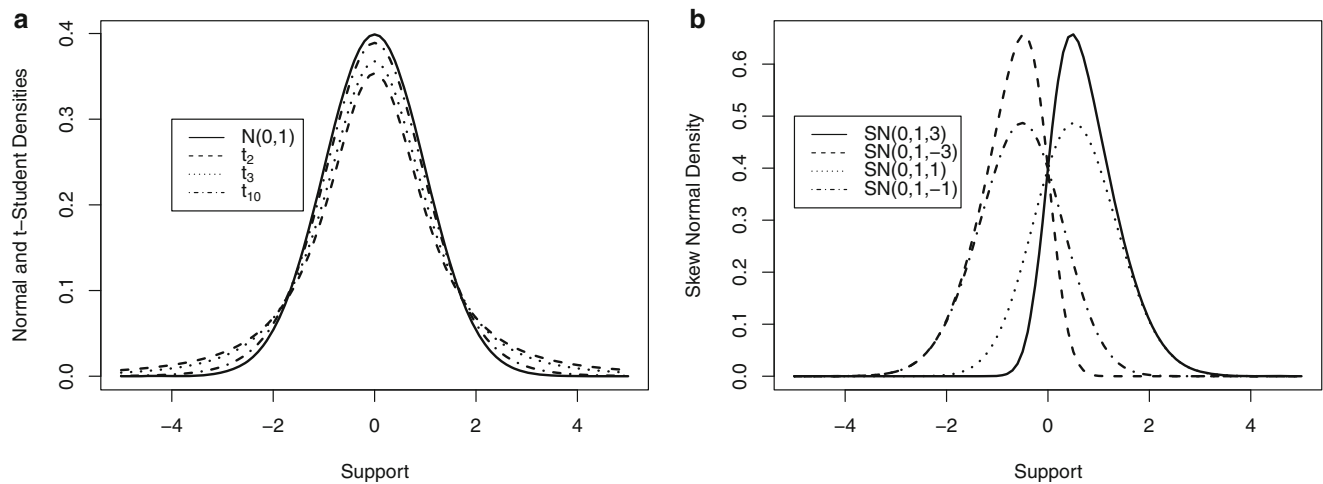
$$T^2 = n(\bar{X} - \boldsymbol{\mu}_0)^T \mathbf{S}^{-1} (\bar{X} - \boldsymbol{\mu}_0) \sim \frac{(n-1)p}{(n-p)} \mathcal{F}_{p, n-p},$$

where $\bar{X} = n^{-1} \sum_{i=1}^n X_i$, $\mathbf{S} = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T$ and $\mathcal{F}_{p, n-p}$ represents the F-Snedecor with degree of freedom p and $n-p$ (which is defined as the ratio of two independent random variables following the chi-square distribution divided by their dfs, termed χ_k^2 in Table 1). Thus, if T^2 is too large, we have evidence to reject H_{02} .

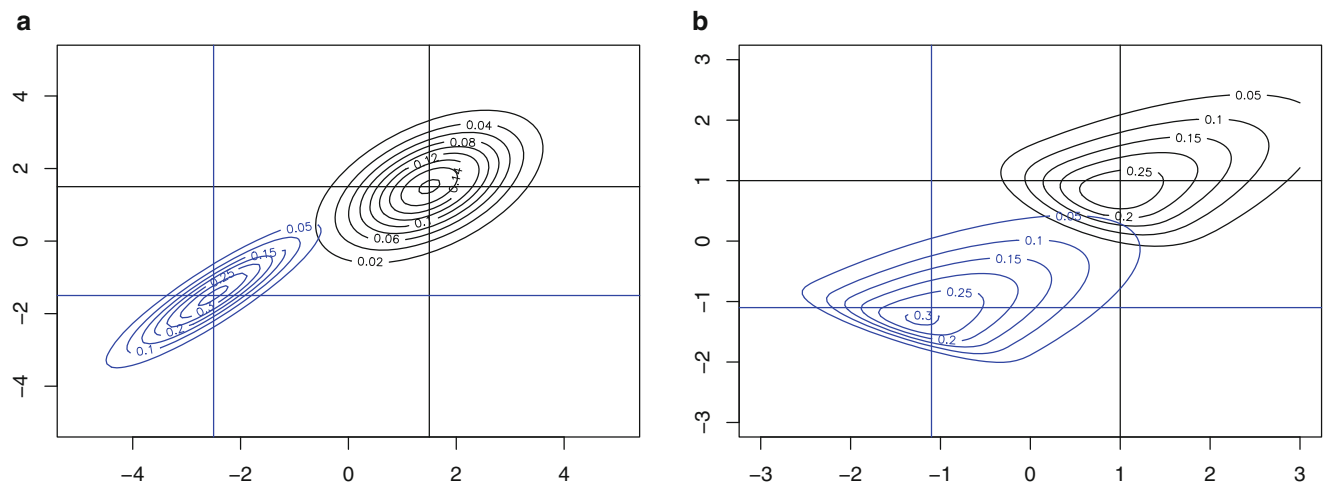
Emery (2008) showed that t and Hotelling's T^2 tests are crucial to check the fit quality in geoscience applications. The N_p distribution (analogous to the t -student law) is symmetric. Figure 2a displays two groups of level curves for this law. To extend the N_p distribution, Azzalini and Valle (1996) pioneered the skew-normal (SN) distribution having pdf given in Table 1. Figure 2b illustrates that the SN level curves can describe the skewness multivariate database, while Fig. 1b confirms this property for a univariate perspective.

Probability Density Function, Table 1 Some pdfs of random variables and vectors. $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal pdf and cdf. $\phi_p(\cdot)$ is the p -variate standard normal pdf. $|\cdot|$ and $(\cdot)^{-1}$ are the matrix determinant and inverse operations

Distributions	pdfs	jpdfs
$N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	$(2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$	$ 2\pi\boldsymbol{\Sigma} ^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$
Skew normal	$\frac{2}{\delta} \phi\left(\frac{x-\mu}{\delta}\right) \Phi\left(\alpha \frac{x-\mu}{\delta}\right)$	$2\phi_p(\mathbf{x}; \boldsymbol{\Omega}) \Phi(\boldsymbol{\alpha}^T \mathbf{x})$
t_ν -student	$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi} \Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$	$\frac{\Gamma((\nu+p)/2)}{\Gamma(\nu/2) \nu^{p/2} \pi^{p/2} \boldsymbol{\Sigma} ^{1/2}} \left[1 + \frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^{-(\nu+p)/2}$
χ_k^2	$\frac{1}{2^{k/2} \Gamma(k/2)} x^{k/2-1} \exp(-x/2)$	-----
		$\frac{(x_1-\gamma_1)^{\alpha_1-1}}{\beta^{\alpha_1} \prod_{i=1}^k \Gamma(\alpha_i)} (x_2-x_1-\gamma_2)^{\alpha_2-1} \cdots (x_p-x_{p-1}-\gamma_p)^{\alpha_p-1}$
$\Gamma(\alpha, \beta)$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$	$\exp\{[x_p - (\gamma_1 + \cdots + \gamma_p)]/\beta\}$,
		$x_{i-1} + \gamma_i < x_i$ (for $i = 2, \dots, p$), $\gamma_1 < x_1$



Probability Density Function, Fig. 1 Curves for normal, t -student, and SN pdfs



Probability Density Function, Fig. 2 Curves for multivariate normal and SN jpdfs

Finally, one of the most used positive distributions is the gamma law that has two parameters: one of shape, $\alpha > 0$, and another scale, $\beta > 0$. This case is denoted by $\Gamma(\alpha, \beta)$ and its pdf is presented in Table 1. The Γ distribution is known as the mother law for describing the speckle noise in synthetic aperture radar system (Ferreira and Nascimento 2020). Mathai and Moschopoulos (1992) proposed a multivariate gamma version having jpdf given in Table 1.

Probabilistic Approach in Geoscience Problems

Many geoscience applications can be understood as random experiments, e.g., on resource characterization (Ma 2019) and measurements of features obtained from remote sensing data (Maus et al. 2019) or SAR imagery (Yue et al. 2021). Thus, it is required to use a probability approach and pdf to describe

geoscience data. Particularly, Ma (2009) showed evidence that the previous axioms iii, iv, and v are crucial to deal with depositional facies analysis, while (Tolosana-Delgado and van den Boogaart 2013) addressed the importance of them to analyze mineral and geochemical elements. Finally, Ma (2019) have pointed out that the use of (a) probability dilemmas and the conditional probability concept and (b) pdfs and jpdfs in the probabilistic mixture are vital to understanding the physical conditions and statistical properties of geoscience phenomena.

Conclusion

In this chapter, we have presented the main concepts for understanding the probability density function (pdf). Some particular jpdfs and pdfs were approached and related to t and

Hotelling's T^2 statistical tests. The latter are important tools in geoscience applications. In summary, the probability approach consists of a base to deal with several geoscience problems, like resource evaluation and mapping by remote sensing sources.

Bibliography

- Azzalini A, Valle AD (1996) The multivariate skew-normal distribution. *Biometrika* 83(4):715–726
- Cordeiro GM, Silva RB, Nascimento ADC (2020) Recent advances in lifetime and reliability models, 1st edn. Bentham Science Publishers, The Netherlands
- Emery X (2008) Statistical tests for validating geostatistical simulation algorithms. *Comput Geosci* 34(11):1610–1620
- Ferreira JA, Nascimento ADC (2020) Shannon entropy for the G_I^0 model: a new segmentation approach. *IEEE J Sel Top Appl Earth Obs Remote Sens* 13:2547–2553
- Hoel PG, Port SC, Stone CJ (1973) Introduction to Probability Theory, Houghton Mifflin, Boston
- Ma YZ (2009) Propensity and probability in depositional facies analysis and modeling. *Math Geosci* 41:737–760
- Ma YZ (2019) Quantitative geosciences: data analytics, geostatistics, reservoir characterization and modeling, 1st edn. Springer, Switzerland
- Mathai AM, Moschopoulos PG (1992) A form of multivariate gamma distribution. *Ann Inst Stat Math* 44(1):97–106
- Maus V, Câmara G, Appel M, Pebesma E (2019) dtwSat: time-weighted dynamic time warping for satellite image time series analysis in R. *J Stat Softw* 88(5):1–31
- Tolosana-Delgado R, van den Boogaart KG (2013) Joint consistent mapping of high dimensional geochemical surveys. *Math Geosci* 45:983–1004
- Yue D-X, Xu F, Frery AC, Jin Y-Q (2021) Synthetic aperture radar image statistical modeling: part one-single-pixel statistical models. *IEEE Geosci Remote Sens Mag* 9(1):82–114

Proximity Regression

Jaya Sreevalsan-Nair
Graphics-Visualization-Computing Lab, International
Institute of Information Technology Bangalore, Electronic
City, Bangalore, Karnataka, India

Definition

Geological features include mineral deposits, intrusions, faults, etc. It is known that the distance to the geological feature, i.e., distance to mineralization, is influenced by several geochemical variables or entities. These variables include the trace elements and the major oxides. The goal is to identify multi-element signatures (Bonham-Carter and Grunsky 2018) in the proximity to the geological features, and locate other regions with similar signatures. To

accomplish such data analysis, linear regression models using *proximity* are used.

Proximity is defined as a function related to distance, just as similarity, where proximity is inversely proportional to distance (Bonham-Carter and Grunsky 2018). The rate of decay of proximity with distance can be considered to be a constant. Thus, by integrating the rate value, we get proximity to be an exponential function of distance. While this can be used for modeling proximity, say at half distance, it can also be predicted using known values of geochemical variables. Such a prediction is done using a linear regression model. Thus, *proximity regression* is the linear regression model of predicting proximity, as the response or dependent variable, using the geochemical variables as the predictors or the independent variables (Bonham-Carter and Grunsky 2018). For proximity Y , expressed as a real value in $[0,1]$, the corresponding distance from feature Z ranges from infinity to zero. Using constant function, $\frac{dY}{dZ} = -\alpha$, which upon integration gives: $Y(Z) = Y(0)e^{-\alpha Z}$.

At half distance $Z_{0.5}$, $\frac{Y(Z_{0.5})}{Y(0)} = 0.5$, thus $\alpha = \frac{-\ln 0.5}{Z_{0.5}} \Rightarrow Y(Z) = \exp\left(\frac{\ln 0.5}{Z_{0.5}} \cdot Z\right)$.

Overview

A dispersion halo is a physical phenomenon that is observed in the country rocks, where a region around a mineral deposit exhibits higher metal values compared to those of the deposit and considerably higher than background values. Usually geochemical sampling and testing are used to delineate the dispersion halo. For computations, proximity is one such measure used for modeling the dispersion halo. While distance to the geological feature is the direct observation, proximity is preferred over distance to be regressed on the geochemical variables for modeling the dispersion halo around a mineral deposit. This is because of the decay of the halo following an exponential function or a power law. Hence, using a measure, such as proximity, that is modeled as a similar function of distance in a linear regression model, improves its prediction. Overall, proximity is modeled as an inversely proportional function of distance (Bonham-Carter and Grunsky 2018).

For using proximity regression, proximity values in close range to the selected feature are used. For dependent variables, the centered logratio (CLR) transformed values of the element values are used in the form of a matrix X of size $m \times n$ for m geochemical variables and n samples or observations. Thus, the column vector of response variables of size n , used in a linear regression model, is given by $Y = X\beta + \varepsilon$, where β is the coefficient (column) vector of size n , and ε is the error (column) vector of size n . This overdetermined equation, as

$m \leq n$, is solved using the least squares method or the conjugate gradient method.

For studying geochemistry of soil samples, compositional data analysis is routinely used, owing to the constant sum constraint (Aitchison 1982). A linear transformation of the Aitchison geometry or simplex, which is the representation of the data points in the compositional data, is used to convert the data to real space. The centered logratio (CLR) is one such linear transformation that is effectively used for the geochemical variables, to avoid the closure problem (Bonham-Carter and Grunsky 2018). The closure problem arises from the standard multivariate analysis to compositional data, where there is a negative bias in the product-moment correlation between the constituents of a geochemical composition (Pawlowsky-Glahn et al. 2007). CLR transformation is also used in applications involving principal component analysis for dimensionality reduction (Pawlowsky-Glahn et al. 2007). CLR transformation involves dividing samples by the geometric mean of its values, and then using its logarithm values. While all components are treated symmetrically in CLR transformation, the transformed data may exhibit collinearity, which is not applicable for methods using full rank data matrices (Filzmoser et al. 2009). In cases of collinearity in the CLR data, the isometric logratio transformation (ILR) is alternatively used (Egozcue et al. 2003; Filzmoser et al. 2009).

Applications

Proximity regression has been applied in the study of litho-geochemistry of volcanic rocks in the Ben Nevis township, Ontario, Canada. Usually auto- and cross-correlations of the geochemical variables are useful in computations of spatial factors for factor analysis methods used for distinguishing litho-geochemical trends from geological processes (Grunsky and Agterberg 1988). It is observed that the anisotropically distributed variables are delineated better than isotropically distributed ones, when using the spatial factor maps. Such a multivariate analysis helps in identifying the dispersion halo that reflects mineralization. Expanding this and other similar works to identifying specific mineral occurrences in Canagau Mines deposit and Croxall property in the Abitibi Greenstone Belt, proximity regression has been used (Bonham-Carter and Grunsky 2018). This method also predicts the Croxall property using the Canagau Mines deposit, using 26 geochemical variables, and 278 samples within 3 km of Canagau Mines

deposit, for training. In addition to identifying these regions as “bulls-eye,” the method has enabled in identifying the predictors or independent variables that identify Croxall property based on the training on the Canagau Mines deposit.

Future Scope

Prospectivity mapping can be considered to be an organic successor to proximity regression-based methods. This can be applied by constructing the proximity measure to the nearest mineral deposit. The problem can be then reframed as predicting the proximity to the nearest mineral occurrence, instead of the occurrence of a mineral deposit (Bonham-Carter and Grunsky 2018).

Different from other spatial regression methods, where proximity and its analogues are used as independent variables, proximity regression is one of the methods where proximity is the response variable. The methodology for prediction can itself be improved through machine learning using neural networks. Alternative to linear regression, logistic regression can be used to strictly constrain the range of proximity to (0,1) (Bonham-Carter and Grunsky 2018). ILR can also be considered as an alternative to CLR for transforming the independent variables (Filzmoser et al. 2009).

Overall, proximity regression is a multivariate predictive method used in multivariate compositional data.

Bibliography

- Aitchison J (1982) The statistical analysis of compositional data. *J R Stat Soc Ser B Methodol* 44(2):139–160
- Bonham-Carter G, Grunsky E (2018) Two ideas for analysis of multivariate geochemical survey data: proximity regression and principal component residuals. In: *Handbook of mathematical geosciences*. Springer, Cham, pp 447–465
- Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barcelo-Vidal C (2003) Isometric logratio transformations for compositional data analysis. *Math Geol* 35(3):279–300
- Filzmoser P, Hron K, Reimann C (2009) Principal component analysis for compositional data with outliers. *Environmetrics* 20(6):621–632
- Grunsky E, Agterberg F (1988) Spatial and multivariate analysis of geochemical data from metavolcanic rocks in the Ben Nevis area, Ontario. *Math Geol* 20(7):825–861
- Pawlowsky-Glahn V, Egozcue JJ, Tolosana Delgado R (2007) Lecture notes on compositional data analysis. <https://dugi-doc.udg.edu/bitstream/handle/10256/297/?sequence=1>. Accessed online last on May 14, 2022

Q

Q-Mode Factor Analysis

Norman MacLeod

Department of Earth Sciences and Engineering, Nanjing University, Nanjing, Jiangsu, China

Definition

Q-Mode Factor Analysis – a statistical modelling procedure whose purpose is to find a small set of latent linear variables that estimate the influence of the factors controlling the structure of case similarities/dissimilarities in a multivariate dataset and are optimized to preserve as much of the inter-case similarity or dissimilarity structure as appropriate given an estimate of the number of causal influences.

Introduction

Whereas a linear, multivariate data-analysis procedure whose purpose was to summarize the covariance/correlation structure among a set of variables (*R*-mode) was first introduced by Karl Pearson (1901), the development of linear procedures whose purpose was to summarize the similarity/dissimilarity structure among sets of cases (e.g., objects, samples or events) (*Q*-mode) is of more recent vintage. This type of multivariate data analysis was introduced into the earth-science literature by John Imbrie (1963) based on previous applications in the fields of psychology and biology. Both principal component analysis (PCA) and component-based, *R*-mode factor analysis (FA) can be used to create mathematical ordination spaces into which cases described by the variables can be projected. Cases that plot close to each other in such spaces can be regarded as having similar variable values and cases that plot at some distance from others can be regarded as having different variable values. But in *R*-mode analyses the spaces created by these methods are referenced to the structure of

covariance or correlation relations among the variables, not similarity relations among the cases themselves. The *Q*-mode methods of principal coordinates analysis (PCoord) and component-based *Q*-mode factor analysis (QFA) were created to redress this deficiency.

Estimating Similarity

Just as the core of the PCA and FA lies in the manner in which pairwise structural relations among variables are portrayed, the core of QFA lies in the manner in which pairwise structural relations among cases are portrayed. But while in *R*-mode methods this usually comes down to a choice between two quite distinct alternatives – covariance or correlation – in *Q*-mode methods a wide variety of similarity and dissimilarity indices are available. Moreover, across the earth sciences it is often the case that variable values are reported as proportions or percentages, a practice that adds complexity to QFA calculations (Table 1).

The similarity coefficients or indices used in QFA can be classified into several types: binary association indices, distance indices, and proportional similarity indices. Binary association indices, such as the coefficients of Jaccard, Dice, Simpson, and Otsuka (Table 1; see Cheetham and Hazel 2012 for a review), are used exclusively for presence-absence datasets and have been applied most extensively in the context of *Q*-mode cluster analysis. They do, however, produce “similarity” matrices than can be operated on by any number of multivariate data-analysis procedures, including QFA. Owing to the count of either mutual presences, or mutual absences, appearing in the numerators of these ratios, the selection of particular binary association indices will focus the analysis on the structure of similarity associations, dissimilarity associations, or some combination of both (Table 2).

Distance-based indices are the most readily understandable family of *Q*-mode similarity coefficients because they

Q-Mode Factor Analysis,
Table 1 A selection of
six *Q*-mode binary
similarity coefficients

$S_{Jaccard} = \frac{C}{N_1 + N_2 - C}$	$S_{SimpleMatching} = \frac{C+A}{N_1+A}$
$S_{Dice} = \frac{2C}{N_1 + N_2}$	$S_{Simpson} = \frac{C}{N_1}$
$S_{Kluczynski2} = \frac{C(N_1 + N_2)}{2(N_1 N_2)}$	$S_{Otsuka} = \frac{C}{\sqrt{N_1 N_2}}$

C Present in both cases, *A* Absent on both cases, *N_i* Total present in case 1, *N₂* Total present in case 2, *N_t* Total present in cases 1 and 2

Q-Mode Factor Analysis,
Table 2 A selection of
five *Q*-mode distance
(dissimilarity)
coefficients

Euclidean Distance	Squared Euclidean Distance
$d_{ij} = \sqrt{\sum_{k=1}^p (x_{i,k} - x_{j,k})^2}$	$d_{ij}^2 = \sum_{k=1}^p (x_{i,k} - x_{j,k})^2$
Manhattan Distance	Mahalanobis Distance
$D_{ij} = \vec{x}_i - \vec{x}_j $	$D_{ij} = (\vec{x}_i - \vec{x}_j)^T S^{-1} (\vec{x}_i - \vec{x}_j)$
Gower Distance	
$g_{ij} = \frac{\sum_{k=1}^p x_{i,k} - x_{j,k} / range_k}{p}$	

x_i = The i^{th} case. x_j = The j^{th} case. p = no. of variables, S = covariance/correlation matrix for the p variables

represent case similarity and difference as a spatial construct. Euclidean, squared Euclidean, Manhattan, and Mahalanobis distances (Table 2) are all examples of distance indices that can be used to express the structure of relations between cases by summarizing pairwise differences across all variables. Because the trace of a pairwise, n by n distance matrix will always be occupied by zeros (since the distance between any object, sample or event with itself = 0.0), distance-based indices are often said to record “dissimilarity”.

Distance indices are applied typically to sets of interval or ratio data values and in situations where it is desirable to retain the magnitude or range of each variables’ values in the analysis. However, if a dataset contains a broad mixture of variable types, or if the data analyst wishes to minimize the influence of different variables’ numerical ranges, Gower’s distance (Gower 1971) may be used. When using Gower’s distance it should be kept in mind that atypical variable values can inflate a variable’s range and so have a disproportionate effect on the resulting pairwise similarity estimate. Ideally, outliers should be removed prior to the calculation of Gower distances and the number of different variable types balanced insofar as possible.

Weights may also be employed to control the relative influence of different variables on the overall measure of similarity though, as pointed out by Gower (1971), justification of an appropriate set of weights in the context of particular investigations is usually quite difficult. Here, the temptation to devise a weighting scheme that will either produce a particular ordination or speed calculations will be ever-present. Such temptations should be resisted.

The final type of similarity index often employed in QFA is Imbrie and Purdy’s (1962) index of proportional similarity.

$$\cos \theta_{ij} = \frac{\sum_{k=1}^p x_{i,k} x_{j,k}}{\sqrt{\sum_{k=1}^p x_{i,k}^2 \sum_{k=1}^p x_{j,k}^2}} \quad (1)$$

In this equation x_i is the i^{th} case; x_j = is the j^{th} case; and p = is the no. of variables. This is by far the similarity index most commonly employed in QFA and often constitutes the default index in public-domain and commercial statistics and data-analysis software packages. The proportional similarity index is often referred to as the “cosine θ ” index because it conceptualizes cases as vectors in the variable space and represents similarity as the angular difference between all pairwise combinations of case vectors, taking no account of these vectors’ lengths. Accordingly, the cosine θ index is a bounded index, constrained to vary between 0 and 1. The cosine θ , or proportional similarity, index is also a true similarity index insofar the trace of the resulting matrix is occupied by 1 s. If the variables used to compute the cosine θ index are standardized prior to similarity estimation the resulting values will be identical to those estimated from a calculation of the “correlation” across variables.

The Method

Once a similarity/distance matrix has been obtained the QFA procedure conforms closely to that of FA. The linear model used in QFA derives from Spearman’s (1904) original formulation and separates the observed n by n similarity or distance structure among cases, calculated across all variables, into a set of a priori designated generalized factors with a residual, specific factor (e) associated with each of the n cases.

$$\begin{aligned}
X_1 &= a_{1,1}F_1 + a_{1,2}F_2 \cdots + a_{1,m}F_m + e_1 \\
X_2 &= a_{2,1}F_1 + a_{2,2}F_2 \cdots + a_{2,m}F_m + e_2 \\
X_3 &= a_{3,1}F_1 + a_{3,2}F_2 \cdots + a_{3,m}F_m + e_3 \\
&\vdots \\
X_j &= a_{j,1}F_1 + a_{j,2}F_2 \cdots + a_{j,m}F_m + e_j
\end{aligned} \quad (2)$$

In this expression a_i is a standardized score the i^{th} case, F is the “factor” value for that case across all measured or observed variables, and e_i is the part of the distance or similarity not accounted for by the factor structure. Q -mode factor analysis attempts to find a set of m linear factors (F , where $m < p$) that underly the sample’s similarity or distance structure and, hopefully, provide a good estimate of the similarity/distance structure of the larger population from which the sample was drawn.

Choosing and Extracting the Factors

In a manner analogous to its R -mode counterpart, QFA begins with a principal coordinate analysis of a properly selected basis matrix whose structure reflects those aspects of between-case similarity, distance or association most appropriate for the investigation at hand. The decision of which similarity, distance or association index to use is not a trivial one as it will influence all aspects of the subsequent analysis. In addition, the number of factors the investigator wishes to construct their model around also needs to be specified at the outset of analysis. The guidance for factor number selection offered in the presentation of R -mode FA can also be applied to QFA. As always, there is no substitute for having a detailed understanding of the system being modeled. If there is uncertainty regarding the number of factors controlling the dataset’s distance/similarity structure, QFA may proceed, but its results should be interpreted with caution.

Like R -mode FA, QFA scales the loadings of the retained principal coordinates by their eigenvalues, according to the following expression.

$$a_{i,m} = \sqrt[2]{\lambda_m} b_{i,m} \quad (3)$$

Here, λ_m is the eigenvalue associated with the m^{th} principal coordinate with m being set to the number of retained factors and b the eigenvector loading of the i^{th} case on the m^{th} principal coordinate (= factor). This scaling, along with the reduction in the number of factors being considered, represents a basic difference between PCoord and QFA.

Next, the “communality” values of the scaled factors are calculated as the sum of squares of the factor loading values

(a) across all factors. This summarizes the proportion of similarity, distance or association provided by each case to each factor. The quantity ‘1- the sum of communality values for each case, then, expresses the proportion of the case’s distance, similarity, or association structure attributable to the error term (e) of the factor model. If the correct number of factors has been chosen, all the summed, squared factor values should exhibit a high communality for each case with a small residual error.

Once the QFA factor equations have been determined, the factor loading values constitute the positions of objects, samples or events projected into the space formed from the eigenvalue-scaled orthogonal factors. These projection scores ($\hat{X}$) may then be tabulated, plotted, inspected and interpreted in the manner normal for PCoord ordinations. Table 3 and Fig. 1 compare and contrast results of a PCoord and two-factor QFA solution for the trilobite data listed in MacLeod (2006).

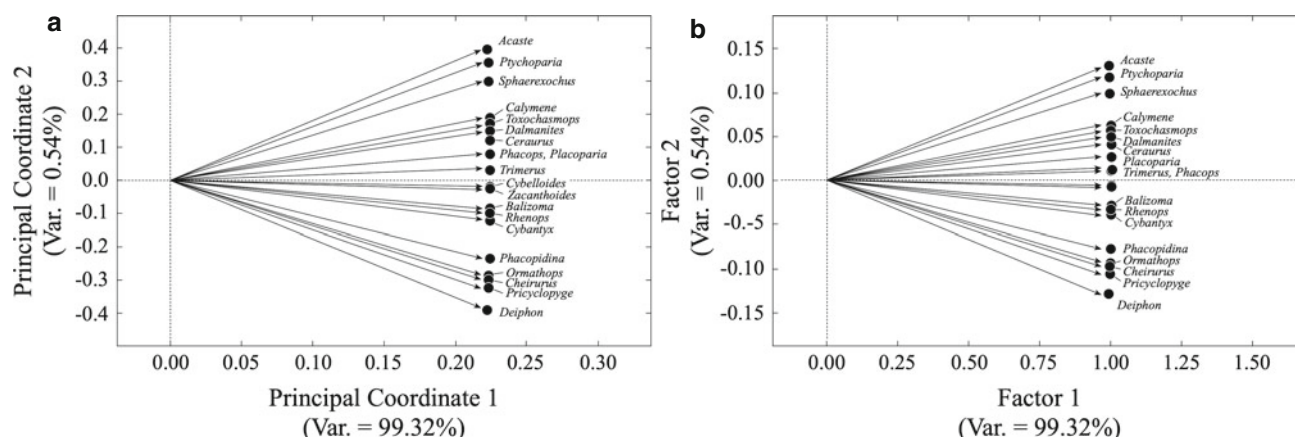
Note that, despite the cosine θ basis for this analysis being a square 20 x 20 matrix, only three principal coordinates can be extracted from this small, demonstration dataset. This limitation stems from the fact that the data from which the similarity matrix was calculated included only three variables: body length, glabella length and glabella width. For this simple dataset both the first principal component and first factor would be interpreted as being consistent with generalized size variation owing to their uniformly positive loading coefficients. Similarly, for both analyses the second component/factor would be interpreted as reflecting localized shape variation owing to the contrast between signs among the different taxa.

Figure 1 illustrates the distribution of these trilobite body shapes in the ordination space formed by the first two principal coordinate axes and the two retained factor axes. While the relative positions of the projected data points are also similar, and discontinuities in the gross form distributions are evident in both plots, the scaling of the PCoord and FA axes differ reflecting the additional eigenvalue-based scaling calculation that is part of component-based QFA. Irrespective of this correspondence, an important change has taken place in the ability of the QFA vectors to represent the structure of the original cosine θ similarity matrix (Table 4).

From these results it is evident that the two-factor QFA has been much more successful in reproducing the entire structure of the original cosine θ similarity matrix than principal coordinate analysis. If all 20 principal coordinates had been used in these calculations the trace of the original similarity matrix would be reproduced exactly, but none of the off-diagonal elements. This simple demonstration illustrates how the eigenvalue-scaling operation, that stands at the core of both FA and QFA, render these approaches to data analysis more

Q-Mode Factor Analysis, Table 3 Principal coordinate loadings, Q-mode factor loadings, and factor communalities for the example trilobite morphometric variables

Parameter/ Variable	Principal Coordinates			Factors		
	1	2	3	1	2	Communality
Eigenvalue	19.865 (99.33%)	0.108 (0.540%)	0.027 (0.134%)	19.865 (99.33%)	0.108 (0.540%)	
<i>Acaste</i>	0.222	0.400	0.090	0.991	0.131	1.000
<i>Balizoma</i>	0.224	-0.084	0.032	1.000	-0.028	1.000
<i>Calymene</i>	0.224	0.192	0.009	0.998	0.063	1.000
<i>Ceraurus</i>	0.224	0.124	-0.022	0.999	0.041	1.000
<i>Cheirurus</i>	0.223	-0.294	0.326	0.994	-0.097	0.997
<i>Cybantyx</i>	0.224	-0.119	-0.137	0.999	-0.039	0.999
<i>Cybeloides</i>	0.224	-0.016	0.075	1.000	-0.005	1.000
<i>Dalmanites</i>	0.224	0.152	-0.277	0.998	0.050	0.998
<i>Deiphon</i>	0.222	-0.388	0.344	0.990	-0.128	0.997
<i>Ormathops</i>	0.223	-0.283	-0.209	0.995	-0.093	0.999
<i>Phacopidina</i>	0.224	-0.233	-0.055	0.997	-0.077	1.000
<i>Phacops</i>	0.224	0.033	0.444	0.997	0.011	0.995
<i>Placopania</i>	0.224	0.083	-0.059	1.000	0.027	1.000
<i>Priscyclopyge</i>	0.223	-0.321	-0.211	0.994	-0.106	0.999
<i>Ptychoparia</i>	0.223	0.359	-0.202	0.992	0.118	0.999
<i>Rhenops</i>	0.224	-0.097	-0.422	0.997	-0.032	0.995
<i>Sphaerexochus</i>	0.223	0.301	0.170	0.995	0.099	0.999
<i>Toxochasmops</i>	0.224	0.174	0.316	0.997	0.057	0.997
<i>Trimerus</i>	0.224	0.035	-0.064	1.000	0.012	1.000
<i>Zacanthoides</i>	0.224	-0.019	-0.144	1.000	-0.006	0.999



Q-Mode Factor Analysis, Fig. 1 Ordination spaces formed by the first two principal coordinates (A) and a two-factor extraction (B) from the three-variable trilobite dataset. Note the similarity of these results in

terms of the relative positions of forms projected into the PCoord and factor spaces and the difference in terms of the axis scales

information-rich than their PCA and PCoord counterparts, especially if a dramatic reduction in causal-factor dimensionality is required. The validity of applying either the FA or QFA models to a dataset is predicated, however, on the dataset exhibiting a factor-based variation/similarity structure.

Aside from the use of different indices to quantify the structure of pairwise relations among cases, QFA also differs

from R-mode FA in a number of other, more subtle ways. For example, it is standard practice to mean center each variable prior to analysis in PCA and FA so the ordination space created as a result of those procedures will be centered on the origin of the coordinate system within which the ordination results will be displayed. This operation is typically not performed in QFA (or PCoord) for either the set of variables

or the set of cases. In this way, the ordination actually expresses the orientation of eigenvectors all of whose tails emanate from the coordinate system origin, irrespective of whether the shafts or bodies of the eigenvectors are drawn in the Q -mode ordination space plots. It would also make little

sense to mean center the case rows across a set of variables of very different types and magnitude ranges.

When QFA ordinations are examined the loadings (= ordination scores) on the first factor (F1) will often all exhibit comparably high positive values. This is usually is a

Q-Mode Factor Analysis, Table 4 Original cosine θ similarity matrix (upper), cosine θ matrix reproduced on the basis of the first to principal coordinate (middle), and cosine θ matrix reproduced by the two-factor

FA solution (lower). Note the bottom two matrices are based on the component/loading values shown in Table 3

Original Similarity Matrix																				
Genus	<i>Acaste</i>	<i>Balizoma</i>	<i>Calymene</i>	<i>Ceraurus</i>	<i>Cheirurus</i>	<i>Cybantyx</i>	<i>Cybeloides</i>	<i>Dalmanites</i>	<i>Deiphon</i>	<i>Ormathops</i>	<i>Phacopidina</i>	<i>Phacops</i>	<i>Placopania</i>	<i>Pricyclopyge</i>	<i>Pychoptaria</i>	<i>Rhenops</i>	<i>Sphaerexochus</i>	<i>Toxochasmops</i>	<i>Trimerus</i>	<i>Zacanthoides</i>
<i>Acaste</i>	1.000	0.987	0.998	0.996	0.973	0.985	0.991	0.995	0.966	0.974	0.978	0.991	0.994	0.971	0.999	0.983	0.999	0.997	0.992	0.990
<i>Balizoma</i>	0.987	1.000	0.996	0.998	0.996	1.000	1.000	0.996	0.994	0.997	0.999	0.997	0.996	0.996	0.989	0.997	0.992	0.995	0.999	0.999
<i>Calymene</i>	0.998	0.996	1.000	1.000	0.986	0.994	0.998	0.999	0.980	0.987	0.990	0.996	0.999	0.985	0.998	0.993	0.999	0.999	0.999	0.997
<i>Ceraurus</i>	0.996	0.998	1.000	1.000	0.989	0.997	0.999	0.999	0.984	0.991	0.993	0.997	1.000	0.989	0.997	0.995	0.998	0.998	1.000	0.999
<i>Cheirurus</i>	0.973	0.996	0.986	0.989	1.000	0.995	0.995	0.984	1.000	0.996	0.998	0.994	0.990	0.996	0.973	0.990	0.981	0.988	0.992	0.993
<i>Cybantyx</i>	0.985	1.000	0.994	0.997	0.995	1.000	0.999	0.996	0.993	0.998	0.999	0.994	0.998	0.998	0.999	0.989	0.989	0.993	0.999	0.999
<i>Cybeloides</i>	0.991	1.000	0.998	0.999	0.995	0.999	1.000	0.997	0.991	0.995	0.997	0.998	0.999	0.994	0.991	0.996	0.994	0.997	1.000	0.999
<i>Dalmanites</i>	0.995	0.996	0.999	0.999	0.984	0.996	0.997	1.000	0.979	0.990	0.991	0.992	0.999	0.988	0.998	0.996	0.996	0.995	0.999	0.998
<i>Deiphon</i>	0.966	0.994	0.980	0.984	1.000	0.993	0.991	0.979	1.000	0.995	0.997	0.990	0.986	0.996	0.966	0.988	0.974	0.983	0.988	0.989
<i>Ormathops</i>	0.974	0.997	0.987	0.991	0.996	0.998	0.995	0.990	0.995	1.000	1.000	0.989	0.992	1.000	0.978	0.998	0.980	0.985	0.994	0.996
<i>Phacopidina</i>	0.978	0.999	0.990	0.993	0.998	0.999	0.997	0.991	0.997	1.000	1.000	0.993	0.995	0.999	0.981	0.997	0.984	0.989	0.996	0.997
<i>Phacops</i>	0.991	0.997	0.996	0.997	0.994	0.994	0.998	0.992	0.990	0.989	0.993	1.000	0.996	0.987	0.989	0.989	0.995	0.999	0.997	0.995
<i>Placopania</i>	0.994	0.998	0.999	1.000	0.990	0.998	0.999	0.999	0.986	0.992	0.995	0.996	1.000	0.991	0.996	0.996	0.997	0.998	1.000	0.999
<i>Pricyclopyge</i>	0.971	0.996	0.985	0.989	0.996	0.998	0.994	0.988	0.996	1.000	0.999	0.987	0.991	1.000	0.975	0.997	0.977	0.983	0.993	0.995
<i>Pychoptaria</i>	0.999	0.989	0.998	0.997	0.973	0.988	0.991	0.998	0.966	0.978	0.981	0.989	0.996	0.975	1.000	0.988	0.998	0.995	0.994	0.992
<i>Rhenops</i>	0.983	0.997	0.993	0.995	0.990	0.999	0.996	0.996	0.988	0.998	0.997	0.989	0.996	0.997	0.997	1.000	0.987	0.989	0.997	0.999
<i>Sphaerexochus</i>	0.999	0.992	0.999	0.998	0.981	0.989	0.994	0.996	0.974	0.980	0.984	0.995	0.997	0.977	0.998	0.987	1.000	0.999	0.995	0.993
<i>Toxochasmops</i>	0.997	0.995	0.999	0.998	0.988	0.993	0.997	0.995	0.983	0.985	0.989	0.999	0.998	0.983	0.995	0.989	0.999	1.000	0.997	0.995
<i>Trimerus</i>	0.992	0.999	0.999	1.000	0.992	0.999	1.000	0.999	0.988	0.994	0.996	0.997	1.000	0.993	0.994	0.997	0.995	0.997	1.000	1.000
<i>Zacanthoides</i>	0.990	0.999	0.997	0.999	0.993	0.999	0.999	0.998	0.989	0.996	0.997	0.995	0.999	0.995	0.992	0.999	0.993	0.995	1.000	1.000

Reproduced Similarity Matrix (PCoord)																				
Genus	<i>Acaste</i>	<i>Balizoma</i>	<i>Calymene</i>	<i>Ceraurus</i>	<i>Cheirurus</i>	<i>Cybantyx</i>	<i>Cybeloides</i>	<i>Dalmanites</i>	<i>Deiphon</i>	<i>Ormathops</i>	<i>Phacopidina</i>	<i>Phacops</i>	<i>Placopania</i>	<i>Pricyclopyge</i>	<i>Pychoptaria</i>	<i>Rhenops</i>	<i>Sphaerexochus</i>	<i>Toxochasmops</i>	<i>Trimerus</i>	<i>Zacanthoides</i>
<i>Acaste</i>	0.209	0.016	0.127	0.099	-0.068	0.002	0.043	0.111	-0.106	-0.063	-0.043	0.063	0.083	-0.079	0.193	0.011	0.170	0.119	0.064	0.042
<i>Balizoma</i>	0.016	0.057	0.034	0.040	0.075	0.060	0.052	0.037	0.083	0.074	0.070	0.047	0.043	0.077	0.020	0.058	0.025	0.035	0.047	0.052
<i>Calymene</i>	0.127	0.034	0.087	0.074	-0.007	0.027	0.047	0.079	-0.025	-0.004	0.005	0.056	0.066	-0.012	0.119	0.032	0.108	0.083	0.057	0.047
<i>Ceraurus</i>	0.099	0.040	0.074	0.066	0.014	0.036	0.048	0.069	0.002	0.015	0.021	0.054	0.061	0.010	0.094	0.038	0.087	0.072	0.055	0.048
<i>Cheirurus</i>	-0.068	0.075	-0.007	0.014	0.136	0.085	0.055	0.005	0.164	0.133	0.118	0.040	0.025	0.144	-0.056	0.078	-0.039	-0.001	0.040	0.056
<i>Cybantyx</i>	0.002	0.060	0.027	0.036	0.085	0.064	0.052	0.032	0.096	0.084	0.078	0.046	0.040	0.088	0.007	0.062	0.014	0.029	0.046	0.053
<i>Cybeloides</i>	0.043	0.052	0.047	0.048	0.055	0.052	0.051	0.048	0.056	0.055	0.054	0.050	0.049	0.055	0.044	0.052	0.045	0.047	0.050	0.051
<i>Dalmanites</i>	0.111	0.037	0.079	0.069	0.005	0.032	0.048	0.073	-0.009	0.007	0.015	0.055	0.063	0.001	0.104	0.035	0.096	0.077	0.056	0.047
<i>Deiphon</i>	-0.106	0.083	-0.025	0.002	0.164	0.096	0.056	-0.009	0.200	0.160	0.140	0.037	0.017	0.174	-0.090	0.087	-0.067	-0.018	0.036	0.057
<i>Ormathops</i>	-0.063	0.074	-0.004	0.015	0.133	0.084	0.055	0.007	0.160	0.130	0.116	0.041	0.026	0.141	-0.052	0.077	-0.035	0.001	0.040	0.055
<i>Phacopidina</i>	-0.043	0.070	0.005	0.021	0.118	0.078	0.054	0.015	0.140	0.116	0.104	0.042	0.031	0.125	-0.034	0.073	-0.020	0.009	0.042	0.055
<i>Phacops</i>	0.063	0.047	0.056	0.054	0.040	0.046	0.050	0.055	0.037	0.041	0.042	0.051	0.053	0.039	0.062	0.047	0.060	0.056	0.051	0.050
<i>Placopania</i>	0.083	0.043	0.066	0.061	0.025	0.040	0.049	0.063	0.017	0.026	0.031	0.053	0.057	0.023	0.080	0.042	0.075	0.065	0.053	0.049
<i>Pricyclopyge</i>	-0.079	0.077	-0.012	0.010	0.144	0.088	0.055	0.001	0.174	0.141	0.125	0.039	0.023	0.153	-0.065	0.081	-0.047	-0.006	0.039	0.056
<i>Pychoptaria</i>	0.193	0.020	0.119	0.094	-0.056	0.007	0.044	0.104	-0.090	-0.052	-0.034	0.062	0.080	-0.065	0.178	0.015	0.158	0.112	0.063	0.043
<i>Rhenops</i>	0.011	0.058	0.032	0.038	0.078	0.062	0.052	0.035	0.087	0.077	0.073	0.047	0.042	0.081	0.015	0.059	0.021	0.033	0.047	0.052
<i>Sphaerexochus</i>	0.170	0.025	0.108	0.087	-0.039	0.014	0.045	0.096	-0.067	-0.035	-0.020	0.060	0.075	-0.047	0.158	0.021	0.140	0.102	0.061	0.044
<i>Toxochasmops</i>	0.119	0.035	0.083	0.072	-0.001	0.029	0.047	0.077	-0.018	0.001	0.009	0.056	0.065	-0.006	0.112	0.033	0.102	0.080	0.056	0.047
<i>Trimerus</i>	0.064	0.047	0.057	0.055	0.040	0.046	0.050	0.056	0.036	0.040	0.042	0.051	0.053	0.039	0.063	0.047	0.061	0.056	0.052	0.050
<i>Zacanthoides</i>	0.042	0.052	0.047	0.048	0.056	0.053	0.051	0.047	0.057	0.055	0.055	0.050	0.049	0.056	0.043	0.052	0.044	0.047	0.050	0.051

(continued)

Q-Mode Factor Analysis, Table 4 (continued)

Genus	<i>Acaste</i>	<i>Balizoma</i>	<i>Calymene</i>	<i>Ceraurus</i>	<i>Cheirurus</i>	<i>Cybantyx</i>	<i>Cybeloides</i>	<i>Dalmanites</i>	<i>Deiphon</i>	<i>Ormathops</i>	<i>Phacopidina</i>	<i>Phacops</i>	<i>Placopania</i>	<i>Prisiclypyge</i>	<i>Psychoparia</i>	<i>Rhenops</i>	<i>Sphaerexochus</i>	<i>Toxochasmops</i>	<i>Trimerus</i>	<i>Zacanthoides</i>
<i>Acaste</i>	1.000	0.987	0.998	0.996	0.972	0.985	0.990	0.996	0.965	0.974	0.978	0.990	0.994	0.971	0.999	0.984	0.999	0.996	0.993	0.990
<i>Balizoma</i>	0.987	1.000	0.996	0.998	0.996	1.000	1.000	0.996	0.993	0.997	0.999	0.997	0.998	0.996	0.989	0.998	0.992	0.995	0.999	0.999
<i>Calymene</i>	0.998	0.996	1.000	1.000	0.986	0.995	0.998	0.999	0.980	0.987	0.990	0.996	0.999	0.985	0.998	0.993	0.999	0.999	0.999	0.997
<i>Ceraurus</i>	0.996	0.998	1.000	1.000	0.989	0.997	0.999	0.999	0.984	0.990	0.993	0.997	1.000	0.989	0.996	0.995	0.998	0.999	1.000	0.999
<i>Cheirurus</i>	0.972	0.996	0.986	0.989	0.997	0.997	0.994	0.987	0.997	0.998	0.998	0.990	0.991	0.998	0.975	0.994	0.979	0.985	0.993	0.994
<i>Cybantyx</i>	0.985	1.000	0.995	0.997	0.997	0.999	0.999	0.995	0.994	0.998	0.999	0.996	0.997	0.997	0.987	0.997	0.990	0.994	0.998	0.999
<i>Cybeloides</i>	0.990	1.000	0.998	0.999	0.994	0.999	1.000	0.997	0.991	0.995	0.997	0.997	0.999	0.994	0.992	0.997	0.994	0.997	1.000	1.000
<i>Dalmanites</i>	0.996	0.996	0.999	0.999	0.987	0.995	0.997	0.998	0.982	0.988	0.991	0.996	0.999	0.986	0.996	0.993	0.997	0.998	0.998	0.997
<i>Deiphon</i>	0.965	0.993	0.980	0.984	0.997	0.994	0.991	0.982	0.997	0.997	0.997	0.986	0.986	0.998	0.968	0.991	0.972	0.980	0.989	0.991
<i>Ormathops</i>	0.974	0.997	0.987	0.990	0.998	0.998	0.995	0.988	0.997	0.999	0.999	0.991	0.992	0.999	0.977	0.995	0.981	0.987	0.994	0.995
<i>Phacopidina</i>	0.978	0.999	0.990	0.993	0.998	0.999	0.997	0.991	0.997	0.999	1.000	0.993	0.995	0.999	0.980	0.997	0.984	0.990	0.996	0.997
<i>Phacops</i>	0.990	0.997	0.996	0.997	0.990	0.996	0.997	0.996	0.986	0.991	0.993	0.995	0.997	0.990	0.991	0.994	0.993	0.995	0.997	0.997
<i>Placopania</i>	0.994	0.998	0.999	1.000	0.991	0.997	0.999	0.999	0.986	0.992	0.995	0.997	1.000	0.991	0.995	0.996	0.997	0.998	1.000	0.999
<i>Prisiclypyge</i>	0.971	0.996	0.985	0.989	0.998	0.997	0.994	0.986	0.998	0.999	0.999	0.990	0.991	0.999	0.974	0.994	0.978	0.985	0.992	0.994
<i>Psychoparia</i>	0.999	0.989	0.998	0.996	0.975	0.987	0.992	0.996	0.968	0.977	0.980	0.991	0.995	0.974	0.999	0.986	0.999	0.996	0.994	0.991
<i>Rhenops</i>	0.984	0.998	0.993	0.995	0.994	0.997	0.997	0.993	0.991	0.995	0.997	0.994	0.996	0.994	0.986	0.995	0.989	0.992	0.997	0.997
<i>Sphaerexochus</i>	0.999	0.992	0.999	0.998	0.979	0.990	0.994	0.997	0.972	0.981	0.984	0.993	0.997	0.978	0.999	0.989	0.999	0.997	0.996	0.994
<i>Toxochasmops</i>	0.996	0.995	0.999	0.999	0.985	0.994	0.997	0.998	0.980	0.987	0.990	0.995	0.998	0.985	0.996	0.992	0.997	0.997	0.998	0.996
<i>Trimerus</i>	0.993	0.999	0.999	1.000	0.993	0.998	1.000	0.998	0.989	0.994	0.996	0.997	1.000	0.992	0.994	0.997	0.996	0.998	1.000	1.000
<i>Zacanthoides</i>	0.990	0.999	0.997	0.999	0.994	0.999	1.000	0.997	0.991	0.995	0.997	0.997	0.999	0.994	0.991	0.997	0.994	0.996	1.000	0.999

reflection of the “size” or combined magnitudes of all the variables included in the analysis. Since there often seems to be little difference among the cases along F1 it is often regarded as an irrelevant or “nuisance” factor, especially when compared to the oftentimes more informative distinctions between cases revealed in the plots of higher-level QFA ordination spaces. Nonetheless, as uniformly high correlations among variables are not characteristic of all earth science datasets, it is always a good idea to inspect the F1 vs F2 and (if necessary) F1 vs F3 ordination plots to determine whether any useful insight can be gained into the structure of between-case similarity/distance relations from an interpretation of F1.

In some instances the variables being analyzed might be expressed as either compositional proportions or percentages such that the sum of each row is constrained to add up to 1.0 or 100.0, respectively. Such datasets are said to have been artificially “closed”. In such cases, application of a centered log-ratio transformation prior to analysis is usually recommended in order to mitigate the effects of the closure constraint.

Factor Rotation

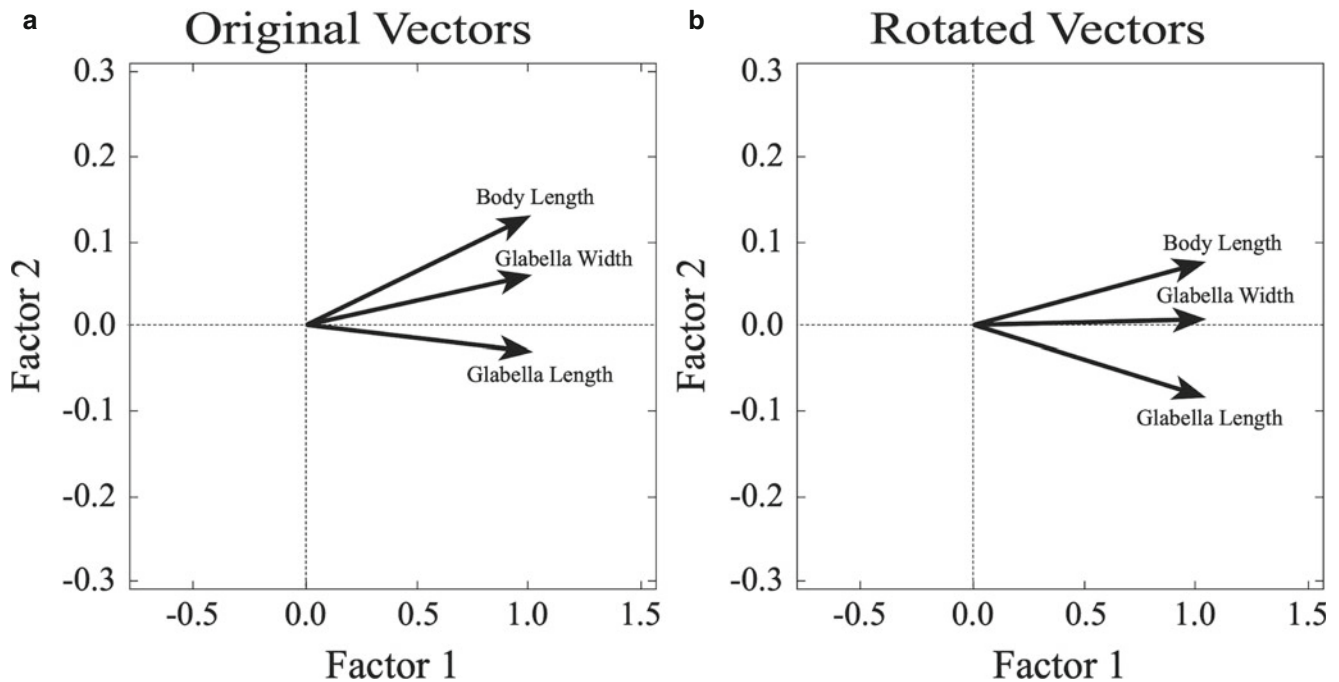
As with FA, QFA factor axes may be rotated relative to the variable axes in order to improve their interpretability. Axis rotation is less commonly applied in QFA because the emphasis is often on identifying the structure of similarity relations

among cases rather than interpretation of the factors in terms of the original variables. Nonetheless, such interpretations are possible and factor-axis rotations often make this interpretation easier. Operationally this amounts to adjusting the set of factor loadings, rigidly in a geometric sense, until all variables exhibit loadings at or as close to 0.0 or ± 1.0 across all factors as possible.

Neuhaus and Wrigley’s (1954) quartimax, and Kaiser’s (1958) varimax orthogonal axis-rotation procedures can be used to perform this operation. If even more simplification is desired the constraint of the factor axes maintaining orthogonal relations can be relaxed using procedures such as Carroll’s (1953) quartimin or Imbrie’s (1963) oblique axis rotation, in which case the factor axes will coincide with the orientation of the extreme variable vectors. Figure 2 illustrates the unrotated and varimax-rotated solutions for the example trilobite dataset.

Conclusion

Q-mode factor analysis is often used to array cases along a spectrum defined by opposing case extrema. In some instances the purpose is to identify these extreme end-members while, in others, it is to gauge which of these spectral end-members particular cases are closest (= most similar) to. Such analyses can also be accomplished, to some extent, by Q-mode cluster analysis. However, in most



Q-Mode Factor Analysis, Fig. 2 Original (a) and rotated (b) variable-axis orientations with respect to the fixed, orthogonal factor axes for the three-variable trilobite data. The rotation of these variables is centered on

Factor 1 owing to their high mutual correlations (reflected in the high Factor 1 eigenvalue, see Table 3). More complex datasets will typically exhibit larger variable-vector rotations

cases cluster analysis represents the structure of between-case similarity relations as a hierarchy despite the fact that the generative processes that give rise to between-case structural relations might not be organized hierarchically. For systems in which similarity-distance relations are known, or suspected, not to be structured hierarchically, QFA may be the more appropriate conceptual model to employ in addition to one that will minimize the distortions that often accompany attempts to portray non-hierarchically structured data using hierarchical models.

Q-mode factor analysis has proven especially attractive to geochemists, mineralogists petrologists, hydrologists, paleo-ecologists, and biostratigraphers who often deal with compositional data and find themselves in need of a procedure to non-hierarchically define both end-members and gradient orderings of cases in which a number of influences are mixed in varying proportions. In perhaps one of the more scientifically significant applications of the method, Imbrie and Kipp (1971) employed QFA to develop a sets of linear transfer functions whereby sea surface temperatures for ancient ocean basins could be inferred from the relative abundances of Cenozoic planktonic foraminifer species. Sepkoski (1981) also used QFA to summarize the diversity history marine life across the Phanerozoic and used the results generated through its application to recognize the three great evolutionary faunas that, together, constitute fundamental macroevolutionary historio-structural units of life on Earth. These are but two prominent examples of QFAs potential for

making important data analysis-based advances in the study of earth history.

Cross-References

- [Cluster Analysis and Classification](#)
- [Correlation and Scaling](#)
- [Eigenvalues and Eigenvectors](#)
- [Grain Size Analysis](#)
- [Inverse Distance Weight](#)
- [Multivariate Data Analysis in Geosciences, Tools](#)
- [Principal Component Analysis](#)
- [R-Mode Factor Analysis](#)
- [Shape](#)
- [Statistical Outliers](#)
- [Variance](#)

Bibliography

- Carroll JB (1953) An analytical solution for approximating simple structure in factor analysis. *Psychomet Theory* 18:23–38
- Cheetham AH, Hazel JE (2012) Binary (presence-absence) similarity coefficients. *J Paleontol* 43:1130–1136
- Davis JC (2002) *Statistics and data analysis in geology*, 3rd edn. Wiley, New York
- Gould SJ (1967) Evolutionary patterns in pelycosaurian reptiles: a factor analytic study. *Evolution* 21:385–401

- Gower JC (1971) A general coefficient of similarity and some of its properties. *Biometrics* 27:857–871
- Imbrie J (1963) Factor and vector analysis programs for analyzing geologic data. Office of Naval Research, Technical Branch, Report 6:83
- Imbrie J, Kipp NG (1971) A new micropaleontological method for quantitative paleoclimatology: application to a Pleistocene Caribbean core. In: Turekian KK (ed) *The Late Cenozoic Glacial Ages*. Yale University Press, New Haven, pp 72–181
- Imbrie J, Purdy E (1962) Classification of modern Bahamian carbonate sediments. *AAPG Mem* 7:253–272
- Kaiser HF (1958) The varimax criterion for analytic rotations in factor analysis. *Psychometrika* 23:187–200
- MacLeod N (2006) Minding your Rs and Qs. *Palaeontological Assoc Newsl* 61:42–60
- Neuhaus JO, Wrigley C (1954) The quartimax method: an analytical approach to orthogonal simple structure. *Br J Stat Psychol* 7:81–91
- Pearson K (1901) On lines and planes of closest fit to a system of points in space. *Philos Mag* 2:557–572
- Sepkoski JJ (1981) A factor analytic description of the Phanerozoic marine fossil record. *Paleobiology* 7:36–53
- Spearman C (1904) “General intelligence”, objectively determined and measured. *Am J Psychol* 15:201–293

Quality Assurance

N. Caciagli

Metals Exploration, BHP Ltd, Toronto, ON, Canada
Barrick Gold Corp, Toronto, ON, USA

Synonyms

QAQC; Quality Control; Quality Management; Quality Systems

Definition

Quality assurance is the act of giving confidence or the state of being certain that a product or service is fit for use or conforms to specifications.

The ISO 9000 standard, clause 3.2.11 defines Quality Assurance as: “A part of quality management focused on providing confidence that quality requirements will be fulfilled”. (ref ISO 9000)

Quality assurance (QA) encompasses all the planned and systematic activities implemented to provide confidence that a product or service will fulfill requirements for quality. Quality control (QC) typically refers to the operational techniques and activities used to fulfill requirements for quality (ASQ/ANSI 2008). Often “quality assurance” and “quality control” are used interchangeably, referring to all the actions performed to ensure the quality of a product, service, or process.

Why QAQC

Quality assurance can be viewed as proactive and process oriented and is focused on the actions that produce the product or ensure the conditions a service must provide. In the geosciences, quality assurance considers the requirements of the data and often deals with the selection of appropriate analytical methods, standards to monitor and document the QC of the analyses, and regular check sampling by a third-party analytical laboratory, including a description of the pass/fail criteria, and the actions taken to address results that are outside of the pass/fail limits (CIM 2018).

Securities regulators and financial institutions now demand that full QAQC procedures are used for all resource programs. There are various regulations and codes depending on where in the world the company has their primary stock exchange listing, such as SAMREC (South Africa), JORC (Australia) and NI 43-101 (Canada). These codes are all based on the standard definitions and globally consistent reporting requirements as determined by the Council of Mining and Metallurgical Institutions (CMMI) and the Committee for Mineral Reserves International Reporting Standards (CRIRSCO). These guidelines and standards are published and adopted by the relevant professional bodies around the world. They detail the mandatory requirements for disclosure of exploration results, reserves, resources, mineral properties, and require a summary of the nature and extent of all QC procedures employed, check assay, and other analytical and testing procedures utilized, including the results and corrective actions taken, in other words a Quality Assurance Program.

The purpose of these international reporting codes is to ensure that fraudulent information relating to mineral properties is not published and to reassure investors that projects have been assessed in a scientific manner. The impetus for developing these codes came from the Busang or Bre-X scandal (Indonesia) in 1997 (Wilton 2017).

Bre-X was a Canadian mining company based in Calgary, Alberta. In October 1995, they announced the discovery of one of the biggest gold deposits in the world. The estimates for this deposit climbed from 2 million ounces to a peak of 70 million ounces in 1997 with speculation of up to 200 million ounces, sending its stock price soaring from a penny share to a peak of CAD\$286.50 a share.

In March 1997 the project geologist, Michael de Guzman, fell to his death from a helicopter over the Indonesian jungle within days of Freeport-McMoran, a potential Busang project partner, announcing that its due diligence revealed only insignificant amounts of gold at the property. An independent third-party, Strathcona Minerals, was brought in to make its own analysis. On May 4th they published their results: the grade and tonnage of the deposit was

unsubstantiated and not based on scientific principles. The reporting was fraudulent, and the shares became worthless. It wiped out billions of dollars for many investors, which included major Canadian pension plans as well as individual investors who had committed their entire retirement funds, their children's college funds, and even mortgaged their homes. It turned out to be the biggest mining scandal of the century (Wilton 2017).

QA, Accuracy, Standards, and Control limits

Accuracy is a qualitative term referring to whether there is agreement between a measurement made and its accepted or reference value. Accuracy is determined using standards, also known as certified reference materials (CRMs) or standard reference materials (SRMs).

Within the geosciences CRMs and SRMs are manufactured materials, either from natural ores or blends of barren material and ore-concentrates, with a known concentration of the elements of interest. The “known” concentration is determined from a round robin involving replicate analyses of multiple aliquots from various labs. This consensus approach results in a recommended value determined from the means of accepted replicate values as well as confidence limits obtained by calculation of the variance of the recommended value (mean of means). These confidence limits are key in determining the upper control (UCL) and lower control limits (LCL) for the process control charts.

The selection of appropriate standards and blanks, and the accuracy and tolerance requirements that the analysis of those standards and blanks must fulfill as well as the precision of the replicate analyses is determined by the QA program.

A standard must be of a material as similar as possible to the samples being assayed to match the ore body mineralogy, ideally custom manufactured from in-house material, and

have a value that sits within the range used for classifying the material (i.e., typical waste values, ore-waste boundary, typical ore values). Given the deposit characteristics (i.e., oxide, sulfide, or various mineralization styles) and material processing requirements (i.e., leach, milling, etc.) several standards may be needed. A minimum of three is typically recommended.

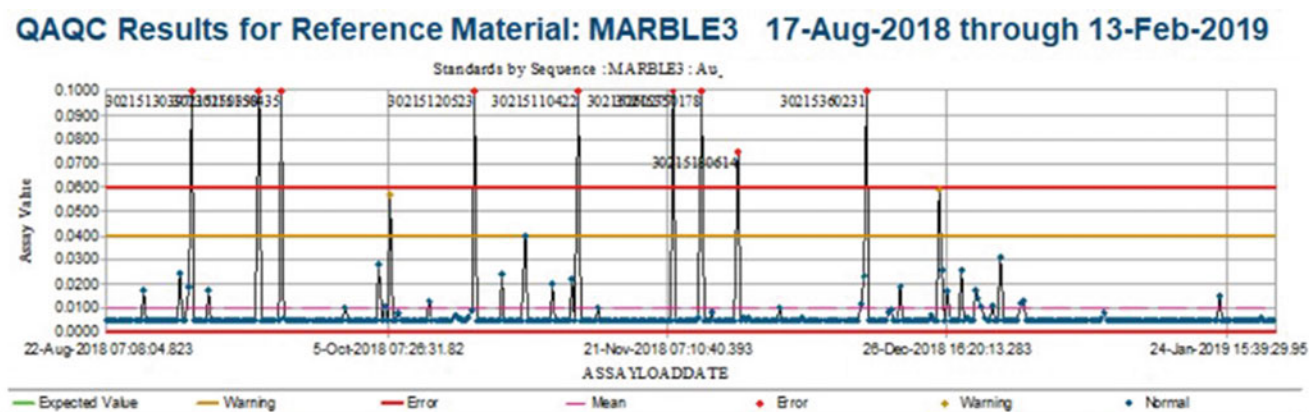
Statistics and QA intersect around the description of pass/fail criteria of the quality controls that are put in place as part of the QA program to ensure accuracy and precision. Schewart Control Charts (Fig. 1.) are used to track the variation in repeated measurements of standard reference materials that is due to a defect or error in the process (i.e., contamination, equipment calibration, deviation from established operating procedures).

Definition of control limits or the natural boundaries of an analysis within specified confidence levels depends on the detection of outliers. When dealing with natural materials, which are expected to be inherently variable, the determination of control limits and detection of outliers in these materials can be a rather circular process. Additionally, identifying an observation as an outlier depends on the underlying distribution of the data, which is often assumed rather than known a priori.

An additional complication faced in the geosciences is that precision is a function of concentration and detection limit and, decreases as concentrations approach the detection limit of a given method. This will also impact what control limits can realistically be achieved for concentrations approaching the detection limits.

QA, Precision and Duplicates

Precision is defined as the amount of variation that exists in the values of multiple measurements of the same material/



Quality Assurance, Fig. 1 Example of a Schewart Control Chart for a standard reference material used to monitor variation in the analysis

parameter and is often expressed as the percent relative variation at the two-standard deviation (95%) confidence level. In this definition, the larger the precision number, the less precise the sampling and analysis. Precision is measured utilizing multiple analysis of standard reference material. Often precision is expressed as percent relative standard deviation (RSD) which is determined from the mean standard deviation of multiple analysis and the mean value of the analysis. Typically, a geological analysis requires $<10\%$ precision, that is that repeated measurements have no more than $\pm 10\%$ error.

In the context of QA precision is examined to assess measurement error and determine realistic control limits and provide a measure of analytical quality (Stanley and Lawie 2008). Statistical analysis of analytical duplicates, such as Thompson-Howarth plots (Fig. 2), are often used to help determine how the variation that is intrinsic to the process (i.e., normal lab variability, natural variability of the material) is a function of concentration and ensure the data is fit for purpose. The Thompson-Howarth method plots the mean of the duplicate pairs against an estimate of the standard deviation of the pair. Depending on the size of the datasets, there is a short method and a long method. From this a linear regression can be determined that expresses standard deviation as a

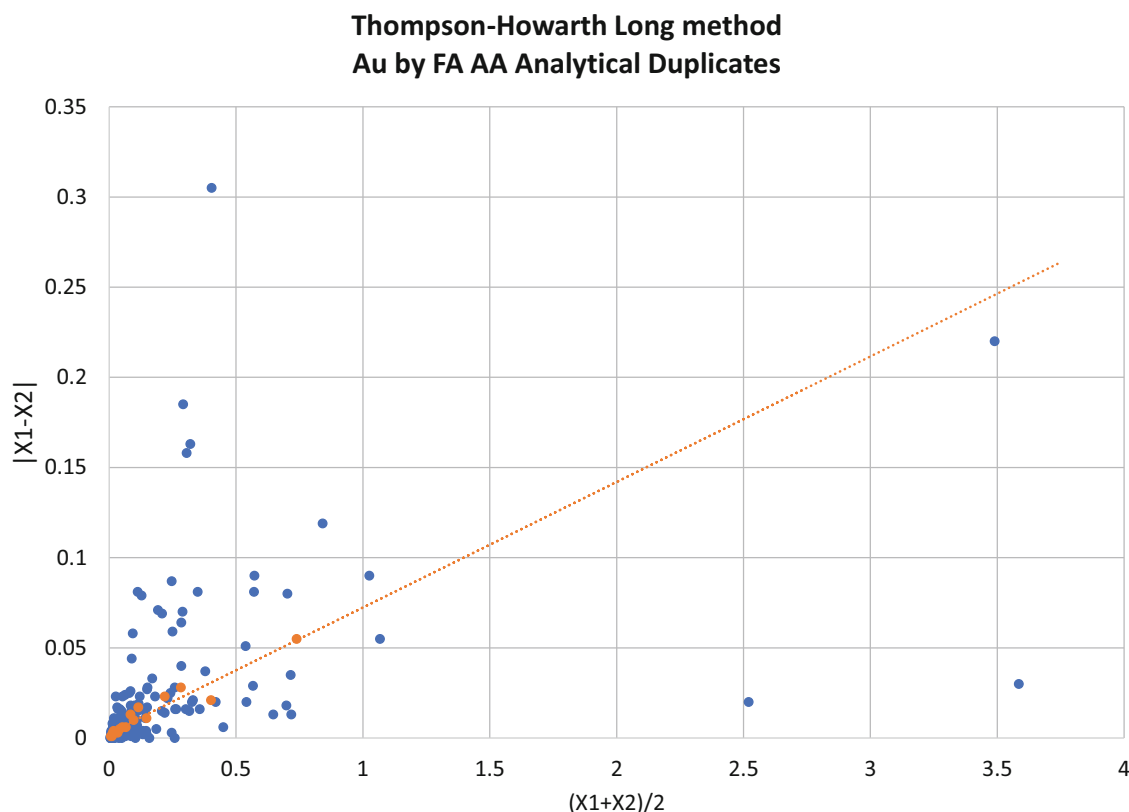
function of concentration, which can then be used to quantify precision at a given concentration (Thompson and Howarth 1978).

Challenges in QA Programs

One of the challenges that commonly arises during the establishment of any QAQC program involving geological systems is the inherent variability found in these systems, which is not known a priori, nor guaranteed to remain constant throughout the life of the project. With respect to this inherent variability outliers will often be encountered; these outliers are hard to recognize and avoid, particularly if the data is noisy or has poor precision to begin with.

Another challenge arises from the distribution of the data that is being examined. In fact, normally distributed errors are the exception rather than the rule in ore deposits and exploration datasets. Many of the common methods of error analysis assume these errors are normally distributed and can result in biased results and poor understanding of data quality when they are not (Stanley and Lawie 2008)

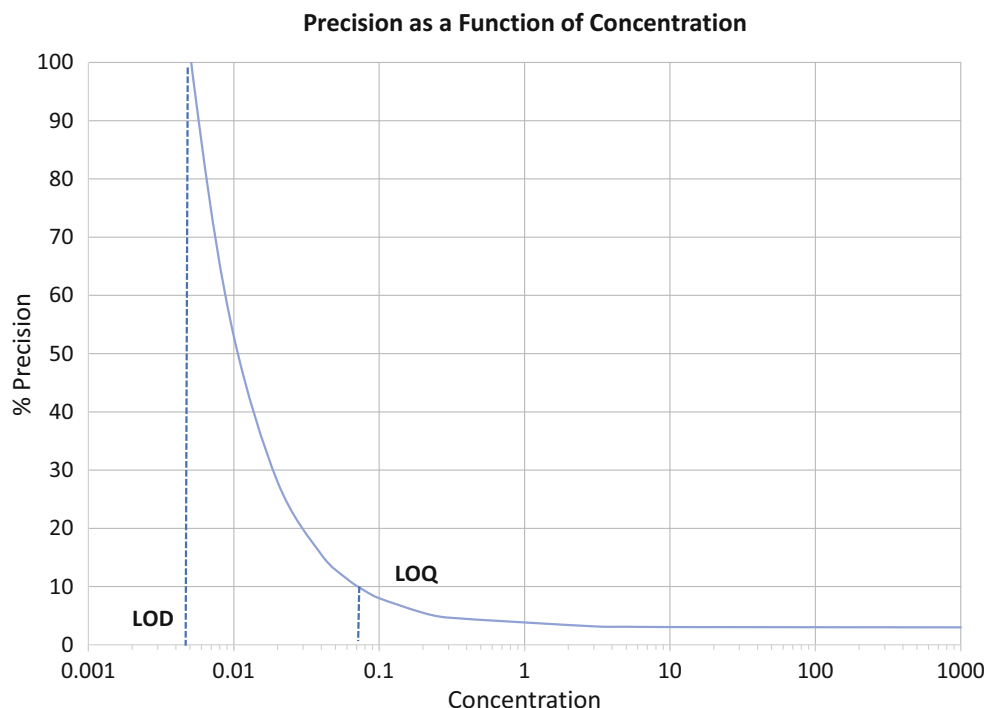
With the improvement of analytical methodology lower limits of detection are constantly being reached; however, in



Quality Assurance, Fig. 2 Example of a Thompson-Howarth long method plot describing error as a function of concentration

Quality Assurance,

Fig. 3 Example of a Thompson-Howarth long method concentration-precision plot that can be used to visualize the precision as a function of concentration. The LOD that is often quoted in analytical brochures reflects the $\pm 100\%$ precision, whereas the LOQ is what is important



geochemical exploration, geochemists are still pushing against these limits with both the materials we seek to characterize and the reference materials we choose to monitor QAQC with. Quite often geologists focus on the method detection limit (LOD), without understanding the limit of quantification (LOQ), that is the concentration above the background where a result is considered quantifiable and trusted (generally the 90th percentile for precision). Because precision is a function of concentration care must be taken to assure that the analytical method chosen results in data that is fit for purpose at the concentrations likely to be important for decision points, geological interpretations, and assumptions (Fig. 3). As a rule of thumb, this can be approximated as 10x the LOD quoted in the analytical brochures.

Summary

Quality assurance in geoscientific reporting centers around the selection of suitable labs and methods of analysis and the selection of appropriate standards and control limits for those standards. It defines the strategies in place for the management of quality as well as a description of the control limits, and the actions taken to address results that are outside of these limits.

Cross-References

- [Howarth, Richard J.](#)
- [Lognormal Distribution](#)
- [Quality Control](#)
- [Standard Deviation](#)
- [Statistical Outliers](#)
- [Statistical Quality Control](#)

Bibliography

- ASQ/ANSI (2008) A3534-2, Statistics, “Vocabulary and Symbols” Statistical Quality Control, American Society of Quality Control. Available at <https://asq.org/quality-resources/quality-glossary>. Accessed 20 Apr 2021
- CIM (2018) CIM Mineral Exploration Best Practice Guidelines. Canadian Institute of Mining, Metallurgy and Petroleum
- International Organization for Standardization (1992) ISO 9000: international standards for quality management. International Organization for Standardization, Genève
- Stanley C, Lawie D (2008) Geochemistry exploration environment. *Analysis* 7:1–10
- Thompson M, Howarth RJ (1978) A new approach to the estimation of analytical precision. *J Geochem Explor* 9(1):23–30
- Webpage: Wilton S (2017) Bre-X: the real story and scandal that inspired the movie Gold. *Calgary Herald*. <https://calgaryherald.com/news/local-news/bre-x-the-real-story-and-scandal-that-inspired-the-movie-gold>. Accessed 27 Nov 2020

Quality Control

N. Caciagli

Metals Exploration, BHP Ltd, Toronto, ON, Canada
Barrick Gold Corp, Toronto, ON, USA

Synonyms

QAQC; Quality control; Quality management; Quality systems

Definition

Quality control can be defined as “part of quality management focused on fulfilling quality requirements” (ASQ/ANSI 2008) by testing a sample of the output against the required tolerances or specifications as determined by the quality assurance requisites.

Often “quality assurance” (QA) and “quality control” (QC) are used interchangeably, referring to all the actions performed to ensure the quality of a product, service, or process. However, QA deals specifically with the setup of the systems and procedures that occur before a batch of samples is sent to the laboratory for analysis, and QC refers to the procedures and tools used to monitor these requirements for quality.

Why QAQC

Securities regulators and financial institutions now demand that full QAQC procedures are used for all resource programs. There are various regulations and codes depending on where in the world the company has their primary stock exchange listing, such as SAMREC (South Africa), JORC (Australia), and NI 43-101 (Canada). These codes are all based on the standard definitions and globally consistent reporting requirements as determined by the Council of Mining and Metallurgical Institutions (CMMI) and the Committee for Mineral Reserves International Reporting Standards (CRIRSCO). These guidelines and standards are published and adopted by the relevant professional bodies around the world. They detail the mandatory requirements for disclosure of exploration results, reserves, resources, and mineral properties and require a summary of the nature and extent of all QC procedures employed, including standards to monitor the quality of the analyses, and regular check sampling by a third-party analytical laboratory, including a description of the pass/fail criteria, and the actions taken to address results that are outside of the pass/fail limits in other words a Quality Assurance and Control Program (CIM 2018).

The purpose of these international reporting codes is to ensure that fraudulent information relating to mineral properties is not published and to reassure investors that projects have been assessed in a scientific manner. The impetus for developing these codes came from the Bre-X scandal in 1997 (Wilton 2017; see Quality Assurance). Bre-X, a Canadian company based in Calgary Alberta, announced the discovery of a large gold deposit in Busang Indonesia and triggering a rise in the stock price from pennies a share to CAD\$286.50 a share. The grade and tonnage of the deposit was discovered to be unsubstantiated and not based on data. It was the biggest mining scandal of the century (Wilton 2017) and wiped out billions of dollars for its shareholders, individuals, as well as Canadian pension plans.

QC Elements to Measure and Monitor Accuracy and Bias

In the geosciences, quality controls include replicate analyses of appropriate standards, blanks to measure accuracy and detect contamination, and regular monitoring of assays by an independent laboratory to identify bias.

Standards

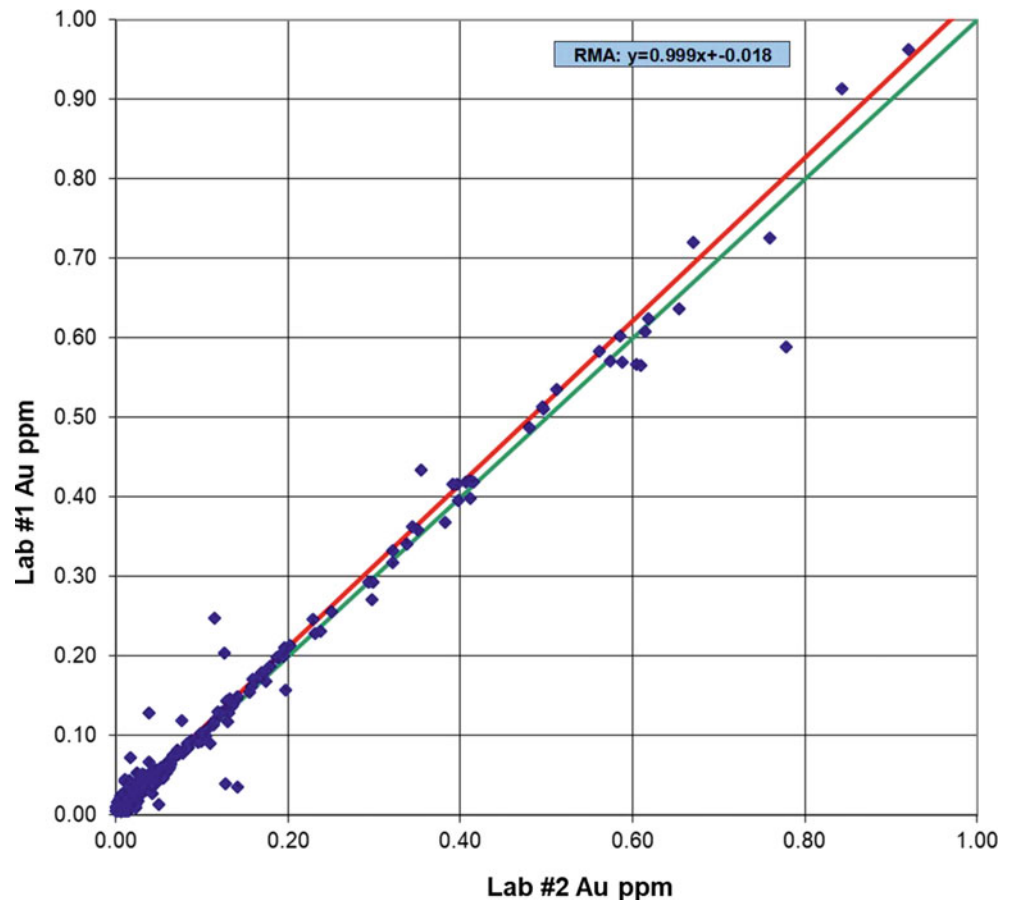
Standards or standard reference materials (SRM) are certified materials inserted into the work order and used to assess the accuracy of the reported data. To be effective, these materials need to match the host rock and the mineralization style (matrix-matched), and efforts must be made to use various concentrations or grades (low, mid, and high) to provide relevant data over the range of reported assays.

Shewhart control charts (see Quality Assurance Fig. 1) are used to track the variation in repeated measurements of standard reference materials that is due to a defect or error in the process (i.e., contamination, equipment calibration, and deviation from established operating procedures). A control chart is plotted for each SRM in use, and assay data are plotted in time order along with the mean, an upper line for the upper control limit, and a lower line for the lower control limit. These lines are determined from a round robin involving replicate analyses of multiple aliquots from various labs. Typically, an SRM falling outside of 3 standard deviations from the mean signals a failure of control in the assay or measurement process. Two out of three successive measurements or assays of SRMs that are on the same side of the mean falling outside of 2 standard deviations will signal bias, another failure of control.

Blanks

Blanks are barren samples with an expected low concentrations or grade relative to the elements being evaluated and are used to check for contamination during sample preparation or

Quality Control, Fig. 1 XY-Scatter plot of primary assay data and umpire assay data showing RMA regression



assaying of elements. The ideal blank material is a hard, chunky barren material such as quartzite. Soft material is prepped more easily in the lab and will not thoroughly scour the equipment for smearing and residue of prior samples. If the material is too small (e.g., pea sized), it will not be crushed effectively and not be a thorough test for smearing of softer metals and residue left in the sample preparation equipment. Additionally, the sample size of blank material submitted must approximate the usual sample size because most laboratories will accept a >1% carryover of elements from one sample to the next, and this 1% carryover will be proportionally larger in a small blank – making the blank test ineffective.

SRM or Pulp Blanks are effectively another type of SRM with very low values of the element in question. These are inserted into the work order and used to assess contamination in the laboratory during sample analysis.

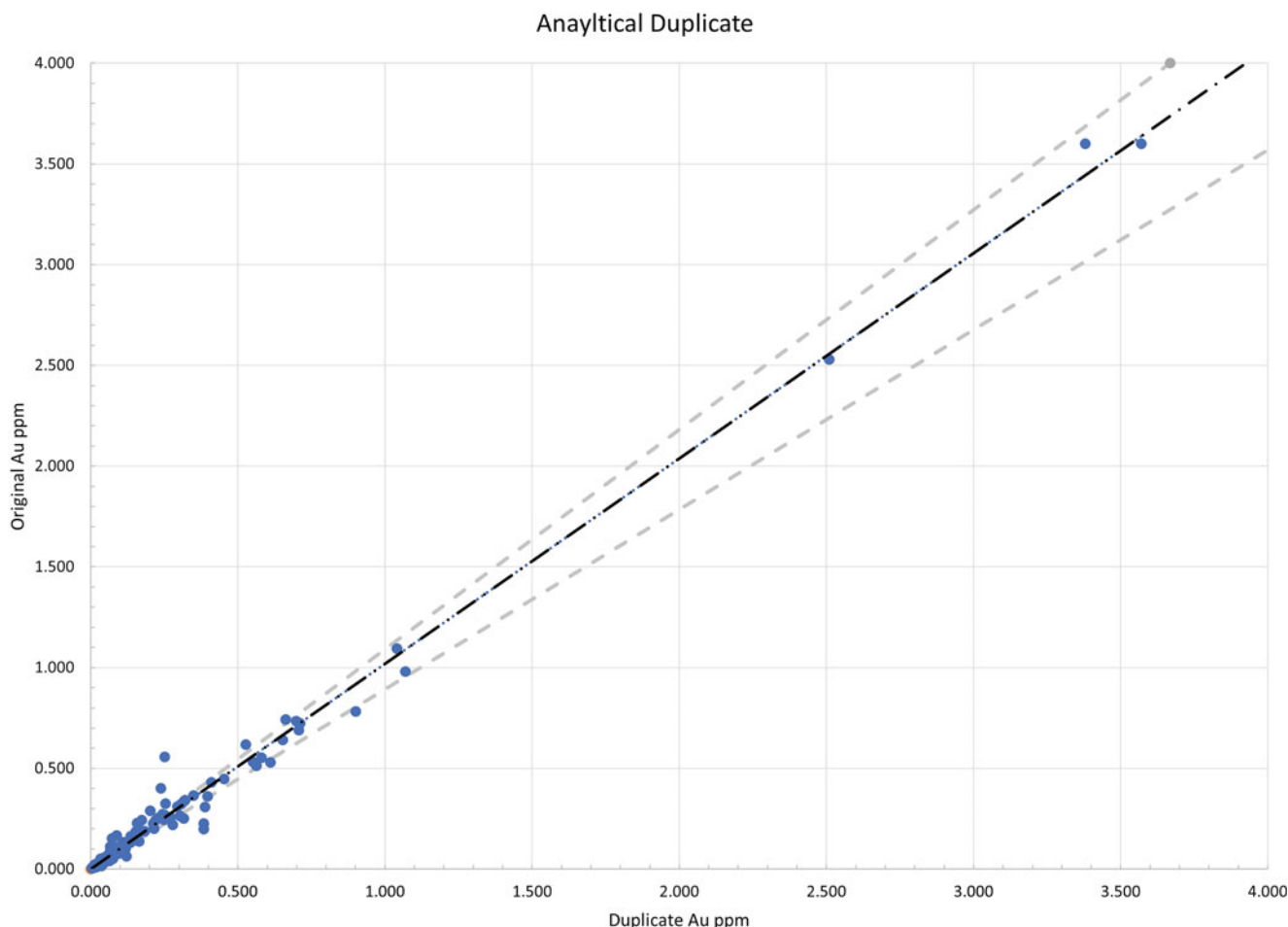
Blanks can be monitored on a chart like a Shewhart control chart, although there is no lower control limit and the upper control limit is defined as a multiple of the analytical detection limit, usually 5 times the analytical limit of detection for the method used.

Check or Umpire Assays

Check assays or umpire samples are pulverized samples (pulp) sent to a secondary lab and used to define inter-laboratory precision and bias. These should be selected to cover a range of concentrations representing the range of values encountered, with particular focus on values representing decision points (e.g., mining cut offs, ore routing, or dispatching triggers). It is expected that the primary lab and the secondary or umpire lab have less than 5% bias.

Inter-laboratory bias is identified by plotting the primary lab and secondary lab pairs on an XY-scatter plot. However, the data from the primary versus the secondary laboratory are independent of each other, and an ordinarily least squares (OLS) regression will correctly describe the relationship between the data because it assumes the X-measurement is without error.

The Reduced Major Axis regression line is the regression line that usually represents the most useful relationship between the X and Y axes. It assumes that both data sets are equally error prone and calculates the regression slope as the ratio of two standard deviations (Sokal and Rohlf 1995).



Quality Control, Fig. 2 A paired duplicate chart with control lines demonstrating the required precision ($\pm 10\%$) for the stage of duplicates

QC Elements to Measure and Monitor Precision

Sample duplicates are created at different stages of the workflow to monitor variability and detect errors related to geology, sampling, sample preparation and subsampling, and analysis. By inserting duplicates at these various points, it is possible to evaluate how much variability is contributed by these different processes.

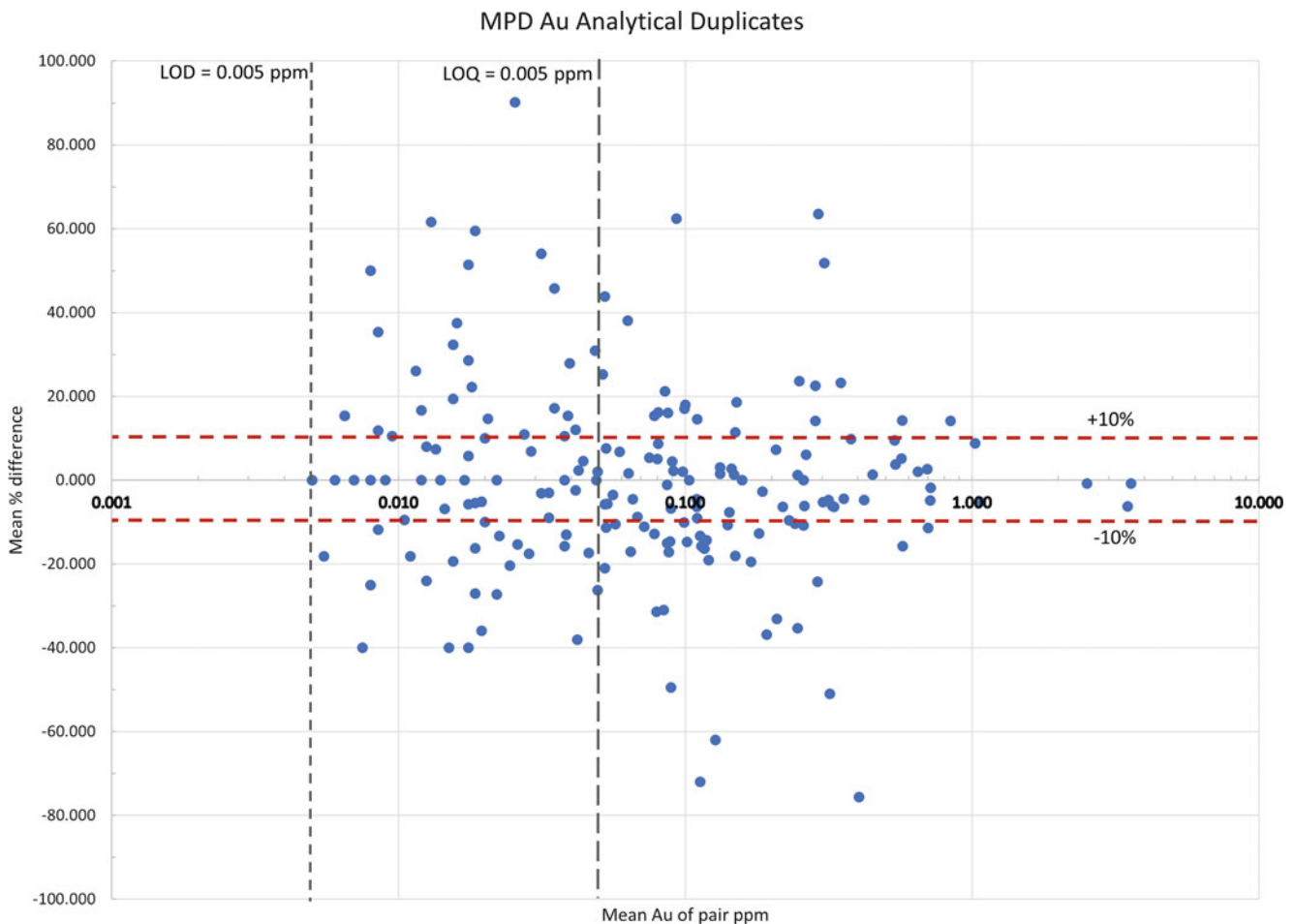
Field Duplicates or Check Samples

Check samples or field duplicates are used to provide information on the reproducibility (or confidence) of the sampling, sample prep, and assaying program. They are another sample taken at the sampling site or in the case of diamond drill core a split (one half or one quarter) of the core. Field duplicate samples reflect the total variability or precision of the material being sampled, the sampling methodology, sample preparation, and analysis. The largest source of variability in field duplicates is related to geological heterogeneity and consistency in sampling and as such does not have pass/fail criteria,

rather they serve to monitor the adequacy of the sampling protocols.

Coarse Duplicates or Prep Duplicates

Coarse duplicates or prep duplicates reflect the variability introduced from the sample preparation procedures as well as the analytical precision. These duplicates are splits of a sample by the lab after the first comminution stage, typically the primary crushing. The largest source of error and variability in these duplicates is related to the subsampling process, that is, the crushing, the splitting of the crushed material, the pulverizing of that split, and the sampling of the resulting pulps. At each step, biases can be introduced by poor methodology, or the comminution and sampling stages may be insufficient to produce a representative subsample of the original sample, and as such they usually do not have pass/fail criteria, rather prep duplicates serve to monitor the adequacy of the sampling comminution and subsampling protocols.



Quality Control, Fig. 3 MPD plot with control lines demonstrating the required precision ($\pm 10\%$) for the stage of duplicates. The limit of detection is shown (LOD) as is the limit of quantification (LOQ). Any samples that are less than the LOQ are not reliable because the signal to

noise ratio of the analysis is too low. The samples greater than the LOQ suggest that the sample prep is not adequate (sample is not sufficiently homogenized) or there are problems with the analysis. In the case, the issue was determined to be the former

Analytical Duplicates or Pulp Duplicates

Pulp duplicates are used to provide information regarding analytical variation which exists in the analysis. These duplicates are splits or second aliquots of the final product of the sample preparation process, usually pulverization for rock and soil, that is analyzed side by side with the first. These duplicates will have pass/fail criteria, typically $\pm 10\%$, as they reflect the precision of the analysis and are used to monitor the quality of the analytical data and processes that are within the lab's control.

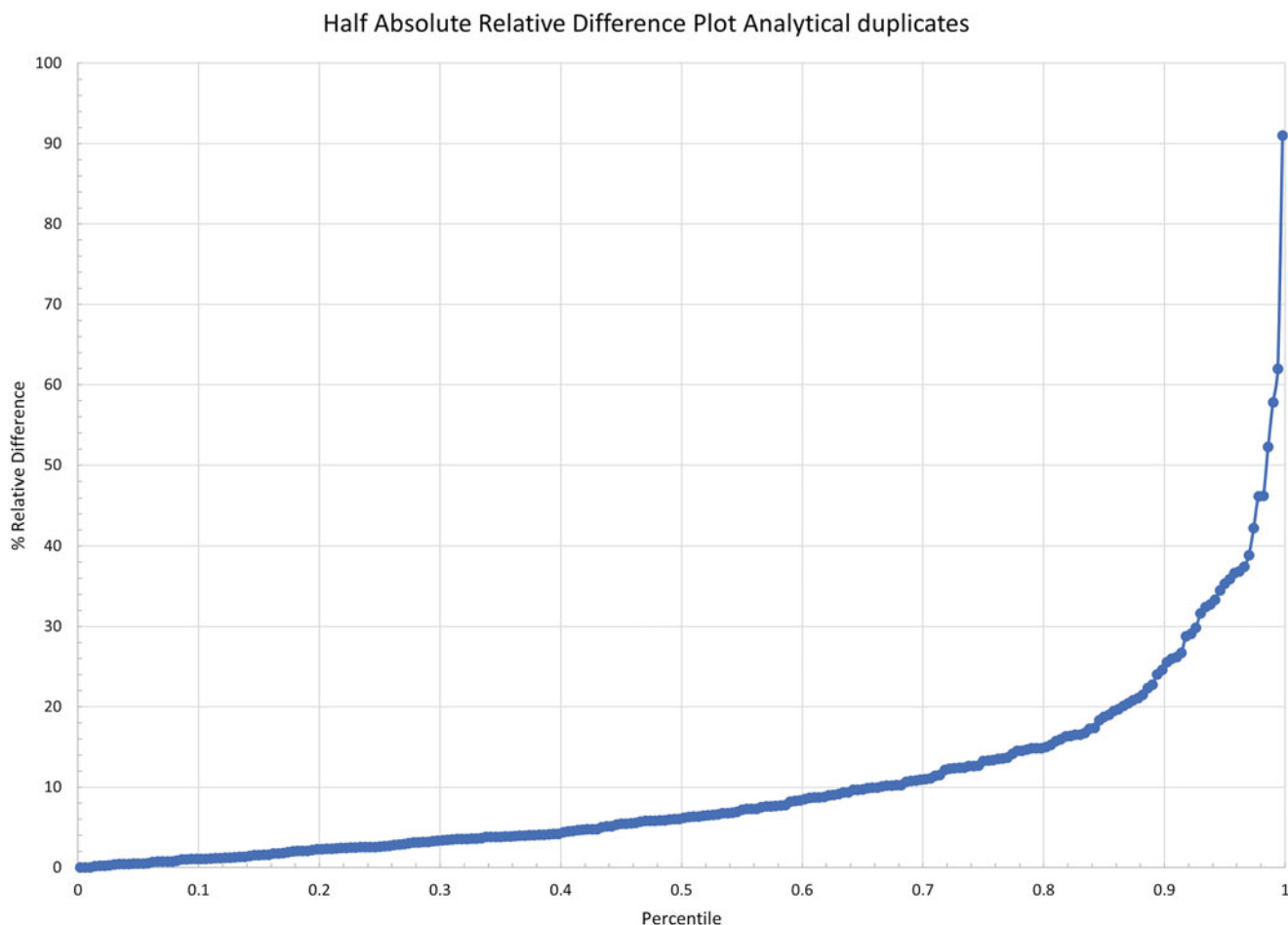
Analysis of QC Duplicates

Various graphical and statistical tools can be utilized to visualize and determine the quality of the data. A duplicate chart plots paired samples on an XY chart (Fig. 2). Control limits reflecting the required precision are also plotted. Typically,

the acceptable tolerance increases from $\pm 10\%$ for analytical duplicates, to $\pm 20\%$ for coarse duplicate, and $\pm 30\%$ for field duplicates, reflecting the increasing contribution of the sub-sampling (coarse duplicates) and geological variability (field duplicates) to the paired samples.

A mean percent difference (MPD) chart plots the mean concentration of the duplicate pair relative to the derived percent difference between the duplicate pair (Bland and Altman 1986). Control limits reflecting the required precision are also plotted (Fig. 3). These plots are used to provide a way of comparing the agreement between pairs and determine whether the degree of agreement is within required precision.

A relative difference plot or half absolute relative difference plot (HARD) ranks relative difference against percentile and is an effective tool only when there are a large number of pairs existing. Data will be deemed acceptable if a certain percentage of the data falls within specific tolerances; depending on the commodity or the mineralization system



Quality Control, Fig. 4 HARD plot for gold assay data from pulp duplicates. From this plot, it can be determined that 90% of the samples have greater than 20% variability, which suggests an unsatisfactory analysis or sample prep for this project

being investigated, this may be something like 90% of the data must have +10% precision (Fig. 4).

All the duplicate pairs can also be used in a Thompson-Howarth analysis (Thompson and Howarth 1973) to determine the effective precision as a function of concentration (see Quality Assurance Fig. 2). Separate curves expressing precision as a function of concentration can be constructed from each duplicate set to visualize precision at each stage from sampling (field duplicates) to sample prep (coarse or prep duplicates) to analysis (pulp duplicates). This will show where and at what concentration the largest variability in the sampling and analytical protocol is.

Challenges in QC Programs

Working with natural systems that have inherent variability always brings challenges. When dealing with trace element concentrations, the “nugget effect” or inability to start with a

representative sample can swamp out any variability in the analytical process making it difficult to monitor whether the analysis is in control or not.

To be effective, QC data must be reviewed on not only a “batch” basis, but also on longer timescales, such as monthly, quarterly, or at project milestones. Data trends such as drifts, shifts, bias, regular failures, and chronic sample prep issues (e.g., contamination) can only be fully realized when compiling data and reviewing longer time intervals.

Summary

Quality control refers to the procedures and tools used to monitor these requirements for quality by testing a sample of the output against the required tolerances or specifications as determined by the quality assurance requisites. Quality control can be viewed as the tools put in place for the measuring and detection of quality.

Cross-References

- [Lognormal Distribution](#)
- [Howarth, Richard J.](#)
- [Standard Deviation](#)
- [Statistical Outliers](#)
- [Statistical Quality Control](#)

Bibliography

- ASQ/ANSI (2008) A3534-2, Statistics, “Vocabulary and Symbols” Statistical Quality Control, American Society of Quality Control. Available at: <https://asq.org/quality-resources/quality-glossary>. Accessed 20 Apr 2021
- Bland J, Altman D (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 327(8476): 307–310
- CIM (2018) CIM mineral exploration best practice guidelines, Canadian Institute of Mining, Metallurgy and Petroleum
- Sokal RR, Rohlf FJ (1995) Section 14.3 “Model II regression”. In: *Biometry*, 3rd edn. W. H. Freeman, New York, pp 541–549
- Thompson M, Howarth RJ (1973) The rapid estimation and control of precision by duplicate determinations. *Analyst* 98:153–160
- Wilton S (2017) Bre-X: the real story and scandal that inspired the movie *Gold*. *Calgary Herald*. <https://calgaryherald.com/news/local-news/bre-x-the-real-story-and-scandal-that-inspired-the-movie-gold>. Accessed 27 Nov 2020

Quantitative Geomorphology

Vikrant Jain¹, Shantamoy Guha¹ and B. S. Daya Sagar²

¹Discipline of Earth Sciences, IIT Gandhinagar, Gandhinagar, India

²Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, Karnataka, India

Definition

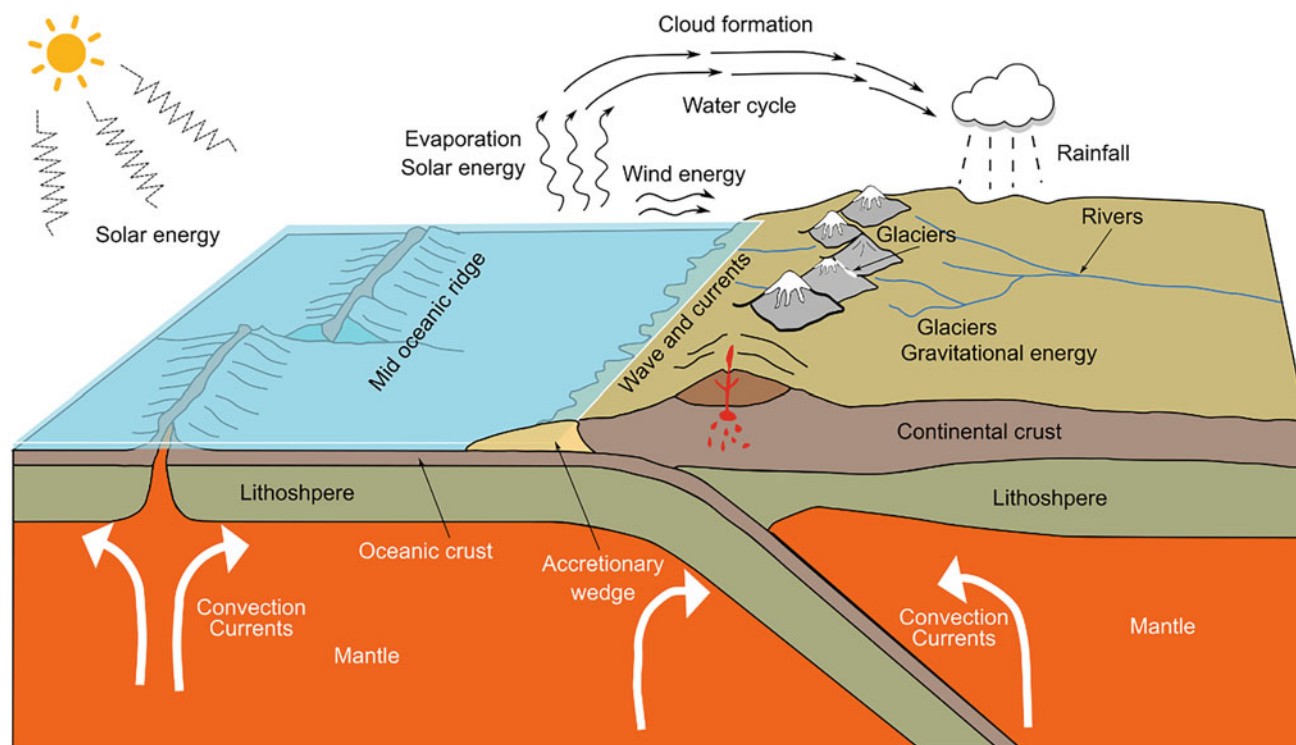
Quantitative geomorphology is a part of geomorphology that explains landscape patterns, geomorphologic phenomena, processes, and cause–effect relationships for landforms on the Earth and Earth-like planetary surfaces in a firm quantitative manner. It demands frameworks to acquire spatiotemporal geomorphologic data, retrieve geomorphologic information, geomorphologic reasoning, and model simulation of geomorphologic phenomena and processes. Quantitative geomorphology is progressing in two dimensions. The structural dimension aims to characterize and explain landscape patterns, while the functional dimension focuses on processes, rates, and analyses of cause–effect relationships using fundamental concepts from physics. The approaches of

geometric, topological, and morphological relevance that show promise in the structural dimension of quantitative geomorphology include classical morphometry, hypsometry, and modern approaches such as fractal geometry, mathematical morphology, and geostatistics. Classical approaches employ source data such as river networks and elevation contours, especially digital elevation models (DEMs), to derive morphometric quantities. Development in the functional dimension of quantitative geomorphology helped to integrate (1) processes at different scales and (2) feedback mechanisms between different landforms (geomorphic systems). Further, the cause–effect relationship in a quantitative framework also leads to developing prediction capability, and hence, providing an important tool to discuss the future of geomorphic systems.

Scope of Quantitative Geomorphology

Geomorphologists have long ventured to identify landscape patterns and understand the events that cause changes in the landscape. The first set of queries related to landscape patterns started in 1950–1960, while the second dimension initiated in the late twentieth century was focused on process modeling and quantitative analysis of cause–effect relationship (Piégay 2019). Quantitative analysis of landscape patterns can be termed as the structural dimension of quantitative geomorphology that aims to quantitatively characterize landform or landscape features. However, current progress in quantitative geomorphology is in the area of process modeling and quantification of a cause–effect relationship, which defines the functional dimension of quantitative geomorphology. The growth in the functional dimension has seen the application of various mathematical tools to study various earth surface processes. Quantitative data on processes such as tectonics, climate, sea-level change, or human disturbances are essential to acquire an in-depth understanding of the cause–effect relationship and identify the causes. The geomorphic processes can be broadly classified into three main categories, e.g., erosion, transportation, and deposition. The primary agents that carry out these processes to modify the Earth’s surface are water, air, and ice. The prominent sources of energy to carry out the geomorphic processes are solar energy and gravitational forces. Solar energy drives other secondary forces like wind energy and generates water cycle. This wind drives the drag force to move sediments in the arid region and further generates waves that drive the geomorphic work in the coastal region. Gravitational force acts on the slope of the terrain in fluvial, glacial, and hillslope processes (Fig. 1).

The geomorphic systems are analyzed through a functional approach (through quantification of driving and resisting forces in a geomorphic system) or through an evolutionary approach to assess the landform or landscape trajectory with



Quantitative Geomorphology, Fig. 1 Generalized diagram of different forms of energy and resultant geomorphic processes

time. Functional approaches were built on the observation and measurement of landforms and fluxes, while evolutionary approaches were based on qualitative analysis or quantitative modeling approaches. Quantitative geomorphology has grown significantly in the last two decades with the advent of advanced techniques of remote sensing, especially digital elevation data, chronological methods (e.g., ^{14}C and OSL for the depositional environment; Cosmogenic nuclide e.g. ^3He , ^{10}Be , ^{26}Al for erosion rates and burial dating), and measurement of isotopes in water and sediments. Further, instrumentation procedures have led to in situ measurements of landform features and fluxes, while the experimental setup in the lab has generated huge quantitative datasets, which has led to several modeling approaches. Chronological tools provide a platform to analyze the process interaction, (dis)connectivity of different landscape drivers by various agents, integration of scales, and emergence of new geomorphic concepts like dis- or nonequilibrium landforms, nonlinearity, and complexity. Key aspects of quantitative geomorphology are discussed further.

Physical and Chemical Processes on the Earth's Surface

Rock Weathering

Weathering is an in situ process that mechanically fractures a rock body or chemically alters it with the help of water. A rock

mass undergoes weathering process and disintegrates into fractured, weathered, and subsequently mobilized by various geomorphic agents. The weathering processes can be divided into two main categories, i.e., physical weathering and chemical weathering.

Physical Weathering

Physical weathering is prominently a mechanical process where a whole rock body is extensively fractured and disseminate into smaller chunks of rocks. The fractured rock mass is further favorable for chemical disintegration and formation of saprolite. A rock body expands due to the heating process that inherently depends on the mineral content. If the stress due to expansion (or sometimes contraction) exceeds the strength of the rock, then it usually fractures and subsequently disintegrates. This expansion and contraction process due to temperature change is expressed by the volumetric thermal expansion coefficient, which is defined as $\alpha = 1/V(\frac{\partial V}{\partial T})_{CP}$, where V is the volume, T is temperature, and CP indicates constant pressure. The stress σ generated due to the strain is

$$\sigma = \frac{\alpha E \Delta T}{1 - \vartheta} \quad (1)$$

where E is the Young's modulus, ΔT is the temperature difference, and ϑ is the Poisson's ratio.

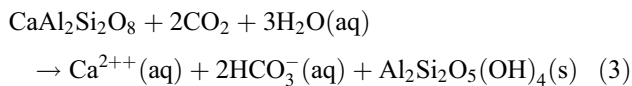
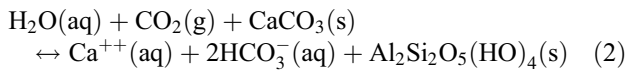
The rate of physical weathering is significantly important in arid or glacial regions due to the importance of expansion–

contraction of rock mass or frost cracking. The degree of weathering and generation of the loose sediment dictates the rate of sediment transport. Frost weathering is important as it enhances the efficiency of the rockfall in the alpine landscape (Draebing and Krautblatter 2019), whereas diurnal heating and salt weathering generate a significant amount of weathering material in the arid or semi-arid region (McFadden et al. 2005).

Chemical Weathering

The top surface of the Earth is constantly exposed to the atmosphere where it interacts primarily with air and water. This interaction leads to a change in the composition of the rock mass as chemicals present in the water interact with thermodynamically unstable minerals. This process is termed chemical weathering in which the minerals come to an equilibrium with the prevailing condition on the Earth's surface.

Apart from the alteration of the chemical composition of the rock, chemical weathering alters the physical property of the rock mass by lowering the strength of the material. The rate at which this alteration takes place is termed the chemical weathering rate. The most common process of chemical weathering is the conversion of the mineral state to form ion or dispersed colloidal molecular units. The simplest process of chemical alteration is the dissolution of the carbonate rock by water and atmospheric CO₂:



Equations 2 and 3 show the carbonate and silicate weathering reactions, respectively. Atmospheric CO₂ uptake during rock weathering is an important process that leads to global change in temperature. Previous research has shown a significant correlation between the silicate weathering rates and the CO₂ consumption rates (Gaillardet et al. 1999; Hartmann et al. 2009). Therefore, the quantification of silicate weathering rates is essential to understanding the carbon sequestration processes on a global scale. Carbon dioxide consumption rate corresponds to the flux of cations from the silicate weathering:

$$\text{CCR} = Q/A \cdot \sum (\text{Na}^+ + \text{K}^+ + \text{Mg}^{++} + \text{Ca}^{2+}) \quad (4)$$

where CCR is the carbon dioxide consumption rate, Q is the discharge in m^3s^{-1} , and A is the surface area of the same watershed in km^2 . The silicate weathering rate and CCR are

consistently higher in tropical watersheds (Gaillardet et al. 1999).

Chemical weathering index or indices of alteration are the quantitative representation of bulk major element oxide chemistry into a single value. Examples of common weathering indices are CIA (Nesbitt and Young 1982), R (Ruxton 1968), WIP (Parker 1970), PIA (Fedó et al. 1996), and CIW (Harnois 1988).

Hillslope Processes

Hillslopes constitute the basic element of any landscape where it generates and transport a huge portion of sediment to either fluvial or glacial system. Hillslopes are either bare bedrock or soil-mantled, and the form of the hillslope depends on the characteristics of the hillslope material. Most soil-mantled hillslopes are convex upward in nature, and diffusive hillslope processes are the most dominant mode of erosion. Erosion of a soil-mantled hillslope can be modeled with a simple two-dimensional diffusion equation (Fernandes and Dietrich 1997):

$$\frac{\partial z}{\partial t} = D \left(\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} \right) \quad (5)$$

where z is the elevation, D is the diffusion coefficient, x and y are the spatial coordinates, and t is the time. However, the soil-mantled hillslopes are mostly dominated by creep, which is a slow transport process. A generalized equation for creep is

$$\dot{\epsilon} = A\sigma^n \quad (6)$$

where $\dot{\epsilon}$ is the strain rate, and σ is the tensile stress. Further,

$$A = A' \frac{D_0 \mu b}{kT} \exp \left(-\frac{Q}{kT} \right) \quad (7)$$

where A' is a dimensionless, experimental constant, μ is the shear modulus, b is the Burgers' vector, k is the Boltzmann's constant, T is the absolute temperature, D_0 is the diffusion coefficient, and Q is the diffusion activation energy.

Apart from the slow diffusive hillslope processes, landslides are a very common case in the geomorphic system where the rapid downslope movement of materials takes place due to a sudden breach in the threshold condition. In the case of a planar segment of the hillslope, the weight of the overburdened material acts vertically and tries to shear the material away from the hillslope. The driving force on a section with width L and height h is

$$F_d = \rho(hdydx \cdot \cos \theta)g \sin \theta \quad (8)$$

where ρ is the bulk density of the material, g is the gravitational constant, and θ is the slope of the planar segment.

The resisting forces are essentially frictional forces due to contact with the overburdened material on the surface. The Frictional forces depends on the characteristics of grains of the material as well as of the surface. The resisting force is

$$F_r = [\rho(hdydx \cdot \cos \theta)g \cos \theta] \tan \phi \quad (9)$$

where $\tan \phi$ is the friction coefficient.

Therefore, the stress balance for a planar segment at the failure can be represented as a “factor of safety,” which is

$$F_s = \frac{[\rho_b g h \cos \theta \cos \theta - \rho_w g d] \tan \phi + C}{\rho_b g h \sin \theta \cos \theta} \quad (10)$$

where d is water table height above the failure plane, ρ_w is the density of water, and C is the cohesion of the material that restricts the failure. This simplistic equation of factor of safety provides quantitative information regarding the propensity of a hillslope for failure.

Geomorphic Agents and Their Actions

Fluvial Systems

Rivers are the most efficient conduit of water, sediment, and nutrients that shape the Earth's surface and sustain aquatic life (Fig. 2). A river generally occupies a valley, which may be a bedrock valley (bedrock rivers) with less amount of sediments on the channel or the river may flow over its alluvium (alluvial channels). A major shift in river science studies took place by

incorporating understanding from various disciplines, including hydrology, hydraulic engineering, physics, mathematics, water and sediment chemistry, ecology, sedimentary geology, and different approaches of numerical modeling. These have provided many datasets to further understand the driving mechanisms that cause changes in the processes and patterns. Water flows under the gravitational force and exerts driving forces for river processes. Furthermore, sediment supply in the channel and bedrock or alluvial characteristics (channel roughness) governs resisting forces for the erosion process.

Erosion in the fluvial system requires power exerted by the flowing water. This power is termed stream power, which is defined as the rate of energy conversion from potential to kinetic energy as water moves downstream (Bagnold 1966). It is expressed as

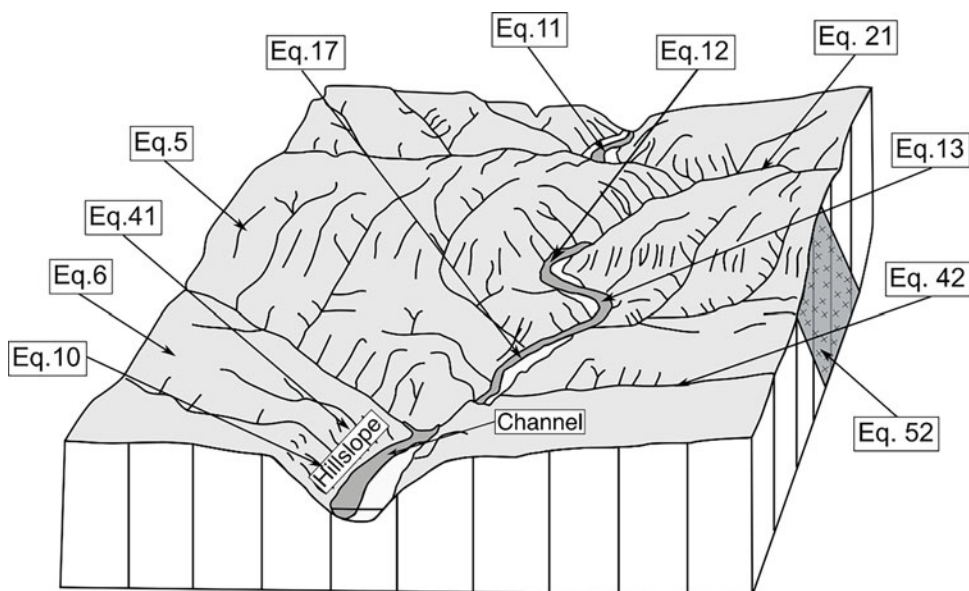
$$\Omega = \rho g Q s \quad (11)$$

where Ω is the stream power, ρ is the density of water (1000 kg/m^3), g is the acceleration due to gravity (9.8 m/s^2), Q is discharge (m^3/s), and s is the channel slope. However, the rate of energy expenditure on the river bed inherently depends on the channel width. Therefore, unit stream power is a more prominent measure of geomorphic work in a channel reach. Channel reach generally defines river characteristics for a length of around 8–10 times of channel width, which represents uniform characteristics of river channel. Unit stream power is expressed as

$$\omega = \Omega/w \quad (12)$$

Quantitative Geomorphology,

Fig. 2 Fluvial landscape and associated processes and landforms. The equation numbers correspond to the different processes. (Modified after Tucker and Slingerland 1994)



where w is the channel width. Empirical studies have shown that unit stream power is an effective metric to understand the sediment transport and erosion processes in a river channel. It also elucidates the significance of the channel width on channel processes (Yanites et al. 2010).

The water column exerts shear stress on the channel bed. Bed shear stress is calculated as

$$\tau = \rho g d s \quad (13)$$

where d is the average depth of the channel.

Furthermore, the shear stress and unit stream power are related as follows:

$$\omega = \tau V \quad (14)$$

where V is the average velocity of the water.

Hence, expressions of driving force vary with scale. Stream power, unit stream power, and shear stress represent the average value of driving force for a given channel length, cross section, or a given point in a cross section, respectively.

Water Flow and Sediment Scale Processes

Transportation of water is a classical fluid mechanics problem that is the basis of deriving the equations of vertical velocity profiles. These approaches can also be compared with traditional empirical equations (e.g., Manning's, Chezy, and Darcy–Weisbach equations). Navier–Stokes equation is used to solve the flow condition in open-channel flow. However, with a few assumptions (i.e., flow is steady, horizontally uniform, and flow is driven by gravity) the equation can be simplified as

$$g_x + \vartheta \frac{\partial^2 u}{\partial z^2} = 0 \quad (15)$$

where $g_x = g \sin \theta$, u is the horizontal velocity of water, z is the water depth, and ϑ is the viscosity of water (Anderson and Anderson 2010). Apart from this methodology, mean velocity is usually estimated with an empirical formula where Manning's equation is the most accepted.

$$v = \frac{k}{n} R_h^{2/3} S^{1/2} \quad (16)$$

where v is the mean cross-sectional velocity, R_h is the hydraulic radius, S is the channel slope, n is the Manning's roughness coefficient, and k is the unit factor (1 for SI units and 1.49 for English units).

Sediment is an important element in the fluvial system that shapes landforms, helps in enhancing nutrients, and sometimes causes havoc to human society. Sediment is usually transported either as bedload or suspended load. The most prominent method to empirically estimate bedload transport

rate is presented by Meyer-Peter and Müller (1948), which is as follows:

$$S_b \propto 8 \left(\left(\frac{k_r}{k'_r} \right)^{3/2} \tau^* - 0.047 \right)^{3/2} \quad (17)$$

where k_r is the roughness coefficient, k'_r is the roughness coefficient based on the grains, and τ^* is the dimensionless mobility parameter.

Suspended sediment load is primarily modeled by empirical power equations, e.g., sediment rating curves (Asselman 2000). However, the continuity equation is also utilized to estimate suspended sediment load and can be expressed as

$$\frac{\partial(AS)}{\partial t} + \frac{\partial(QS)}{\partial x} - \alpha w \omega (S - S_*) = 0 \quad (18)$$

where Q is the flow discharge, A is the flow area, S is the sediment concentration, S_* is the averaged sediment-carrying capacity, α is the adjustment coefficient for a cross section, and w is the water surface width (Zhang et al. 2014).

Reach Scale

River morphology is important to understand the connection between the channel and watershed. It also defines habitat for riverine ecology. Therefore, it is important to understand and quantify the causative factors of river morphology. Reach-scale studies of river morphology primarily focus on the effect of physical factors on the water and sediment transport in a river system. Unit stream power is the most effective measure of the driving force to understanding morphological processes (Eq. 12). A river reach can be represented in distinct categories. The most dominant patterns are braided, meandering, and anabranching. A braided river consists of a multiple channel network usually separated by a high amount of bedload deposited in a channel. A meandering channel on the contrary is a single-thread channel with multiple bends along its path. Anabranching channels are also multithread channels but fundamentally stable. Van den Berg (1995) suggested an empirical equation to separate single-thread and multithread channels based on a substantial dataset from natural channels:

$$\omega = 900 D_{50}^{0.42} \quad (19)$$

where ω is the unit's stream power at the transition between single-thread and multithread channels, and D_{50} is the median grain size. The sinuosity values also suggested that the higher gradient with higher unit stream power values results in a transition from single-thread to multithread channel (especially braided channels). Sediment characteristics govern resisting force. Chang (1985) expressed the grain size as

an important factor that leads to morphological changes in the natural river. The regression between $S/\sqrt{D}$ versus Q was used to understand the morphological characteristics of alluvial rivers.

The erosion process in an alluvial channel is usually modeled with Exner's equation:

$$\frac{\partial z}{\partial t} = \frac{1}{\rho_b} \frac{\partial Q}{\partial x} + \frac{\partial S}{\partial t} \quad (20)$$

where Q is the sediment transport rate as bedload, S is the suspended sediment, and ρ_b is the porosity factor (Anderson and Anderson 2010).

In the case of the bedrock rivers, the erosion processes at reach scale are modeling using an excess shear stress model:

$$E = K(\tau - \tau_c)^a \quad (21)$$

where τ is the bed shear stress, τ_c is the critical shear stress, and K and a are parameters. An implication of several studies has shown that Eq. 21 is a nonlinear equation that is dependent on the local slope and water discharge (Tucker and Hancock 2010).

Basin Scale

Apart from the reach-scale analysis, understanding of topographic and network characteristics at basin scale is important to quantify the topography. Morphometric analysis is perhaps the most established method to quantify, describe, and analyze landforms at different scales of studies. For the basin-scale studies, researchers have used morphometric analysis for several decades. These morphometric parameters are primarily linear or areal (Table 1).

Hypsometric curve and integral are also important to study drainage basin characteristics. Hypsometric analysis corresponds to the distribution of surface area to elevation. Hypsometric curves have been largely utilized to describe the flatness of a drainage basin surface. Hypsometric integral is the integration value of the hypsometric curve and estimated using the following expression:

$$\frac{V}{HA} = \int_0^1 x \, dy \quad (22)$$

where V is the total volume, HA is the entire volume of the reference area, $x = \frac{a}{A}$ (a is the area within a band of elevation; A is the total area of the basin), and $y = \frac{h}{H}$ (h is the relative elevation; H is the absolute elevation). This integral primarily describes the nature of the dissection of a landmass within a drainage basin (Strahler 1952).

River long profile is the fundamental feature to assess the spatial variability of fluvial processes. The shape of the long profile reflects the evolutionary trajectory of a fluvial system (Fryirs and Brierley 2012), and therefore, modeling the long profile has remained a fundamental question. Various functions like logarithmic, power, and exponential functions have been used to express the shape of long profiles. Jain et al. (2006) observed that natural river long profile shape could be better represented by the sum of two exponential functions:

$$Z = \alpha_1 e^{-\beta_f L} + \alpha_2 e^{-\beta_s L} \quad (23)$$

where Z is the elevation of the channel long profile at any point, L is the length from channel head up to any point along the river long profile, and α and β are constants and

Quantitative Geomorphology, Table 1 A list of morphometric parameters to assess the drainage basin characteristics

Category	Morphometric parameters	Definition/equation
Linear	Stream order	Method to assign a numeric order to links in a stream network
	Stream number (N_u)	Number of stream segments of order u
	Stream length (L_u) (km)	Length of stream of particular order
	Stream length ratio (L_{ur})	$\overline{L_u}/\overline{L_{u+1}}$
	Bifurcation ratio (R_b)	N_u/N_{u+1}
Areal	Basin length (km)	The linear distance between source to mouth of a river in a basin
	Basin perimeter (km)	Length of the basin boundary
	Form factor ratio (R_f)	A/L_b^2
	Elongation ratio (R_e)	$1.128\sqrt{A}/L_b$
	Circularity ration (R_c)	$4\pi A/P^2$
	Drainage density (D_d)	$\sum L_s/A$
	Drainage texture (T)	$\sum N_u/P$
	Stream frequency (F_s)	N_u/A
	Compactness coefficient (C_c)	$2(\sqrt{A/\pi})/L_b$

Total length of stream for (L_s), length of particular stream order (L_u), number of streams for particular stream order (N_u), drainage basin area (A), length of drainage basin (L_b), and drainage basin perimeter (P)

coefficients, respectively, of the exponential function (Sonam and Jain 2018).

The morphometric analysis has been effectively used to evaluate the geological or tectonic controls in a drainage basin (Różycka et al. 2021). The differential geometry has been effectively utilized to understand the tectonics and landslide processes (Jordan 2003). Recently, the application of fractals on morphometric data also highlighted the role of lithology on landscape characteristics (Sahoo et al. 2020).

Morphometric characteristics define the physical characteristics of a river basin and can also be utilized to understand the hydrologic response utilizing geomorphologic instantaneous unit hydrograph (GIUH). GIUH is the probability density function for the time of arrival of a raindrop, randomly placed at any point in the watershed (Rodríguez-Iturbe and Valdés 1979; Jain and Sinha 2003). The general equation of GIUH for an N th-order stream can be derived by application of semi-Markov process on the drainage network and is written as

$$\frac{d\theta_{N+2}(t)}{dt} = \sum_{i=1}^N \theta_i(0) \frac{d\varphi_{i(N+2)}(t)}{dt} \quad (24)$$

where $\theta_{N+2}(t)$ is defined as the probability that the drop is found in the state $(N+2)$ at the interval “ t ,” and “ $N+2$ ” represents the final state at the basin outlet; $\theta_i(0)$ is the probability that the transport of the drop starts at state “ i ”; and $\varphi_{i(N+2)}(t)$ is the interval transition probability from state “ i ” to state $N+2$.

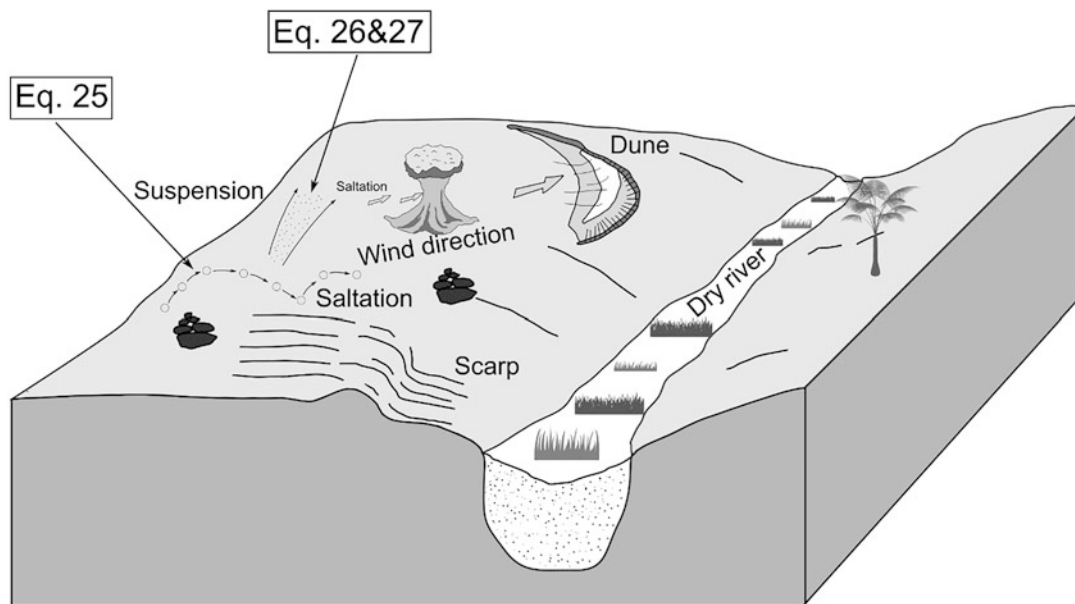
GIUH is efficient model for predicting direct runoff based on Horton’s stream-order ratio and kinematic wave theory. Additionally, the model can also depict the role of geomorphic characteristics in the flood hazard. Length ratio and length of the channel of maximum orders are proved to be the most effective control on the hydrologic response of a drainage basin in the Himalayan setting (Jain and Sinha 2006). The knowledge of the hydrologic response of a drainage basin subsequently helps to better model the response of an ungauged drainage basin.

Aeolian Systems

Aeolian processes in the arid region are primarily controlled by the action of wind, which is essentially a form of incoming solar energy in terms of a pressure gradient (Fig. 3). Apart from solar energy, the wind is also generated due to the Earth’s rotation and surface irregularities. Erosion and sediment transport processes in the arid region are primarily quantified by wind velocity and shear stress. Sediment is first entrained by the shear stress of the wind by overcoming the particle weight and cohesion. Bagnold (1941) first presented the simplified model representing a critical wind shear for a particular grain size:

$$u_{*ct} = A \sqrt{\frac{(\sigma - \rho)}{\rho}} g.d \quad (25)$$

where u_{*ct} is the critical shear stress, A the constant dependent upon the grain Reynolds number, σ is the particle density, ρ is



Quantitative Geomorphology, Fig. 3 Aeolian landscape consisting of major landforms associated with processes. The equation numbers correspond to the different processes

the wind density, g is the acceleration due to gravity, and d is the grain diameter.

After the selective entrainment, the sand grains are transported either via saltation or suspension. A general model for the transportation of sediment is therefore the function of the grain size, though a general model of sediment transport should incorporate a range of grain sizes and variable wind speed. Therefore, in a given shear velocity of the wind, larger grains are transported via saltating trajectories and finer grains follow suspension trajectories. Four main forces act on the sand particle: (1) body force due to gravity, (2) aerodynamic drag force, (3) aerodynamic lift force, and (4) Magnus lift due to the spin of the particle (Anderson and Hallet 1986).

Empirical models are also used to predict wind-driven transportation processes. One such model is the Rodok model, which is a fully empirical model that is particularly used to model the movement of fine particles. Whenever the wind velocity exceeds the threshold of sediment entrainment, this model is applicable (Leenders et al. 2011). The model is given as

$$F_e = a_1 e^{b_1 u_*} \quad (26)$$

where F_e is the flux of sediment, u_* is the shear velocity of wind, and a_1 and b_1 are the empirical constants. Further, physics-based equations have also been used to understand the sediment flux in the arid and semi-arid regions. A dynamic mass balance approach is applied to the sediment transport problem. The model suggests an increase in the flux when the horizontal wind velocity exceeds the threshold of the entrainment of the sand grain. Experimental studies have shown that the dynamic mass balance approach provides a better means to quantify the sediment flux (Mayaud et al. 2017). Furthermore, this model provides temporal-scale variation of sediment transport mechanisms. Dynamic mass balance model is given as

$$\frac{dQ_{ip}}{dt} = H \left[a_1 (u - u_t)^2 \right] - \left(\frac{Q_{ip}}{\left(\frac{b_1 d}{D} + \frac{c_1 u}{u_t} \right)} \right) \quad (27)$$

where Q_{ip} is the predicted mass flux for the given time interval, u is the mean horizontal wind velocity over the given time interval, u_t is the horizontal wind velocity threshold, d is the grain diameter (mm), D is a reference grain diameter, a_1 is the constant with dimension, b_1 and c_1 are dimensionless constants. Integrated wind erosion modeling system (IWEMS), on the other hand, provides a quantitative prediction of wind erosion from local to global scale (Lu and Shao 2001). This model incorporates physics-based as well as empirical equations that takes care of starting from particle entrainment to deposition. Researchers have observed that

these integrated models provide a better means to simulate the effect of dust storms over a large area (Lu and Shao 2001).

The grains that are entrained and transported eventually get deposited when the gravitational force acting on the grains overcomes the shear stress force. However, the rate of settling of particles in fluid flow (water and air) also depends on the density and viscosity of the transportation agents following Stoke's law. Furthermore, vegetation also plays a major role in the sediment deposition processes by disturbing the wind velocity profiles by acting as a roughness on the surface.

Glacial Systems

Glaciers are the accumulated body of ice that moves downstream due to the weight of the ice and the slope of the surface under the gravitational forces (Fig. 4). Glacial systems have been used as markers of climate change because the change in the mass of ice in the glacier dictates temperature change (Vuille et al. 2008). Depending on the thermal regime, glaciers can be categorized into two types: (1) cold-based glaciers and (2) warm-based glaciers. Cold-based glaciers have the basal part entirely below the pressure melting point and are therefore also termed "dry-based glaciers." The temperature at the base is characterized by subzero temperature. Warm-based glaciers or temperate glaciers are characterized by the temperature at 0° or above the melting point. Warm-based glaciers slide over a thin sheet of water and under the gravitational force through internal deformation. Whereas the cold-based glaciers move through internal deformation only. Basal sliding velocity is modeled as a function of basal shear stress (τ_b) (driving force) and effective normal stress (N) (resisting force):

$$U_b = k \tau_b^m N^{-p} \quad (28)$$

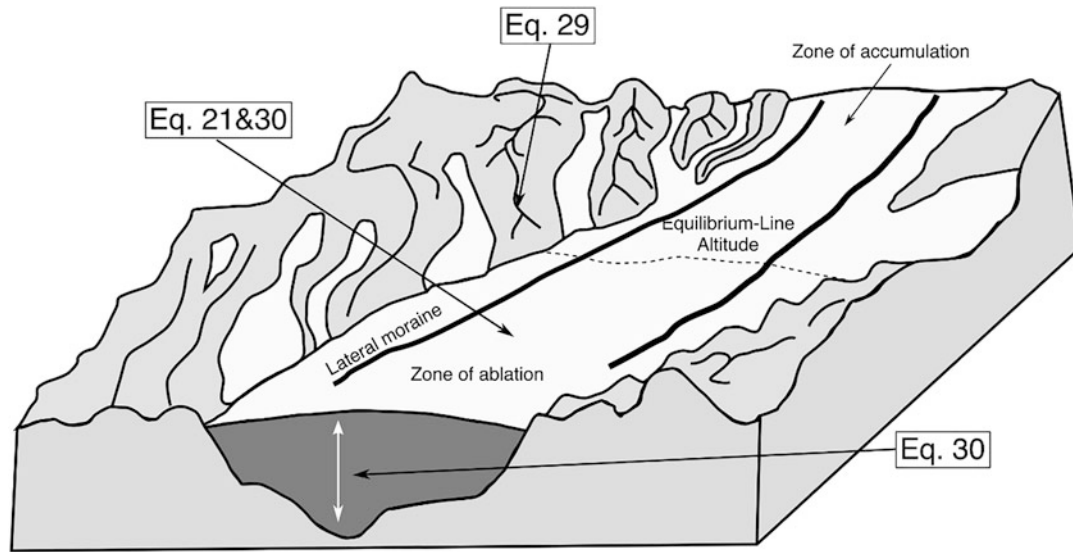
where m and p are positive constants, k is a sliding parameter, and N is the effective normal stress that is the difference between ice overburden and the basal water pressure (Raymond and Harrison 1987). Empirical data also support that the erosion rates are positively correlated to the basal sliding velocity across the globe (Cook et al. 2020).

The glacial erosion rate is generally assumed to be proportional to the sliding velocity of the ice in the glacial valley. Bedrock surface erosion rate (E) is estimated as

$$E = e_c T U_b^{e_v} \quad (29)$$

where e_c is erosion rate constant, T is the time-step size, U_b is the sliding velocity, and e_v is an erosion exponent (Harbor 1992).

The flow of ice behaves as non-Newtonian fluid where strain rate and shear stress are nonlinearly related:



Quantitative Geomorphology, Fig. 4 Glacial landscape and processes. The equation numbers correspond to the different processes

$$\frac{du}{dz} = a \left(\frac{\tau}{\tau_0} \right)^n \quad (30)$$

where u is the fluid velocity, z is the distance along the profile, τ is the shear stress, τ_0 is a reference or yield stress, and a and n are rheological parameters. The values of a and n may be constant in some cases but practically depend on the temperature or density of the flow (Pelletier 2008).

Sediment transport in the glacial system happens via (1) debris-rich ice transport or (2) transport by meltwater. However, most sediment transport occurs by meltwater, which is characterized as glacio-fluvial sediment transportation (Delaney et al. 2018). The study utilized an empirical power-law function between the bedload and discharge from the melt. Additionally, a substantial mass of the sediment is also transported supraglacially or within the ice mass. These piles of sediment are only visible below the equilibrium line. One of the primary sources of sediment in glacial streams is the debris transported by the ice and subglacial rivers.

Depositions in the glaciated region consist of poorly sorted grains, primarily called “tills.” A distinctive depositional feature in the glacial valley is eskers. Eskers represents the landform where subglacial tunnels swiftly transport water and coarse sediments. The equation for the potential field that is the total head is described as follows (Anderson and Anderson 2010):

$$\Phi = \rho_w g z + \rho_i g (H - (z - z_b)) \quad (31)$$

where z is elevation, z_b is the elevation of the bed, and H is the thickness of the ice. Water is transported along a line perpendicular to the line of equipotential:

$$\frac{d\Phi}{dx} = \rho_w g \frac{dz}{dx} + \rho_i g \frac{d(H - (z - z_b))}{dx} = 0 \quad (32)$$

The expression for the potential gradient along the interface of ice and rock is

$$\left. \frac{d\Phi}{dx} \right|_{bed} = (\rho_w - \rho_i) \frac{dz_b}{dx} + \rho_i g \frac{dz_s}{dx} \quad (33)$$

where z_s is the elevation of ice surface. The term bed slope can be written as a function of the slope of the ice surface as

$$\frac{dz_b}{dx} < -11 \frac{dz_s}{dx} \quad (34)$$

Overall, the slope of the glacier, temperature dynamics, and glacio-fluvial interaction governs the depositional characteristics.

Coastal Processes

Coastal regions are one of the most dynamic landscapes. The coastal landscapes are prominently shaped by the oceanic surface waves (driven by wind energy) and the tides (driven by the gravitational attraction of the Moon and the Sun). The change in climatic conditions marks the imprint on the coastal geomorphology through the change in the geomorphic processes driven by the aforementioned agents. Contemporary as well as ancient depositional features record the period of sea-level changes, which is a marker of climate change. Wave energy is the most important geomorphic agent in the coastal system. The surface waves originate when the wind energy is transferred to the ocean surface. Therefore, strong wind-storms are a causative factor of higher-amplitude waves.

The wave energy transported toward the coastal region in the form of oscillatory waves on the beach area leads to substantial modification of the physical features. The energy per unit length of wave or the energy density may be written as

$$E = \frac{\rho g H^2}{8} \quad (35)$$

where H is the wave height, ρ is the density of ocean water, and g is the gravitational constant.

The mean potential energy is

$$\overline{PE} = \frac{\rho g h^2}{2} + \frac{\rho g H^2}{16} \quad (36)$$

Here, the first term is the mean potential energy associated with the water column height above the ocean surface and the second term represents potential energy associated with the wave (Holthuijsen 2010).

The flux of energy or the power of the wave is the product of the energy density with the group wave speed:

$$\omega = \frac{\rho g^{3/2}}{8} H^2 h^{1/2} \quad (37)$$

Wave energy is further transformed into currents, which aid in transporting and depositing sediments in the beach area. Coastal barriers are generally formed by shore-parallel transportation of sand and gravels. The most used empirical equation for the longshore drift is (Komar 1998)

$$Q_{\text{long}} = 1.1 \rho g^{3/2} H_{\text{br}}^{5/2} \sin \alpha_{\text{br}} \cos \alpha_{\text{br}} \quad (38)$$

where H_{br} is the height of the breaker, g is the acceleration due to gravity, ρ is the density of water and the sediment transport rate Q_{long} , and α_{br} is the angle between the breaker line and local shoreline. Although this mathematical model proved to be the proper justification of the physics of the movement of grains, real-time estimation of sediment transport is extremely difficult to comprehend. Therefore, these are usually achieved by measuring the actual amount of loose sediments trapped behind man-made structures.

Deltas are the other important landform that forms when the river enters the sea. Deltas are primarily important because of the diverse flora and fauna. The formation of the delta can be examined by using the continuity equation of sediment (Eq. 20).

Although this simplified formulation provides meaningful results, the shape and sedimentology largely depend on the water density difference, shape of the continental shelf, and amount of sediment in the channel. Therefore, modeling of deltas remains an important problem to date (Takagi et al. 2019).

Landscape Patterns and Process Interactions

Mathematical Morphology in Quantitative Geomorphology

Terrestrial surfaces of Earth and Earth-like planets exhibit variations across spatiotemporal scales. Understanding the organizational complexity of such terrestrial surfaces and associated features both across spatial and temporal scales leads to a study of interest to quantitative geomorphologists. Quantitative geomorphology, which gained wide attention from the classical geomorphologists, was popularized through Horton–Strahler’s contributions to morphometric and hypsometric analyses, respectively, carried out on two fundamental topological quantities, namely, river networks and elevation contours. The main source of these two quantities was topographic maps. However, the data, in particular, DEMs, available at multiple spatial scales, offer numerous advantages over the topographic maps but also pose challenges to the quantitative geomorphologists. The DEM data have rich geometric, morphological, and topological (GMT) relevance immensely useful in quantitative geomorphology. To develop a cogent spatiotemporal model, well-analyzed and well-reasoned information retrieved from spatiotemporal data are important ingredients (Sagar and Serra 2010). Such models are essential to understanding the dynamical behavior of terrestrial surficial processes in a firm quantitative manner. Three ways of understanding complexity involved in spatiotemporal behavior of terrestrial surfaces and associated phenomena include considering (1) topographic depressions and their relationships with the rest of the surface, (2) unique topological networks, and (3) terrestrial surfaces. The three features of geomorphologic relevance represented in mathematical terms as functions, sets, and skeletons are respectively surfaces, planes, and networks. From the context of geomorphology, some examples of such functions, sets, and networks include DEMs, water bodies, and river networks (Sagar 2020). Mathematical morphology (Serra 1982; Matheron 1975) offers an approach to deal with all the intertwined topics.

With the advent of powerful computers with high-resolution graphics facilities and the availability of DEMs, a new set of mathematical approaches offered quantitative geomorphologists insights. One of such mathematical approaches is mathematical morphology (Serra 1982) originally popular in shape and image analysis studies. Mathematical morphology offers numerous operators and transformations – named after terms stemming from the literature of the Earth Sciences (Beucher 1999) – to deal with retrieval of information from DEMs, quantitative characterization of DEMs, quantitative reasoning of the information retrieved from DEMs, modeling, and simulation of various surficial processes that involve DEMs, and spatiotemporal visualization of various surficial phenomena and processes.

Initial impetus for quantitative geomorphology through the applications of mathematical morphology was given by Sagar (1996, 2013). For the retrieval of GMT quantities from the DEMs, quantitative analysis of DEMs and GMT quantities thus retrieved from the DEMs, quantitative spatiotemporal reasoning of those retrieved quantities, modeling and simulation, as well as visualization of spatiotemporal behavior of geomorphologically relevant phenomena and processes, mathematical morphology offers numerous operators, transformations, algorithms, and frameworks. Some of the transformations, to name a few, include

- Skeletonization to extract topological quantities such as the ridge, the valley connectivity networks (Sagar et al. 2000, 2003), mountain objects (Dinesh et al. 2007), and the hierarchical watersheds (Chockalingam and Sagar 2003) from the DEMs that would be eventually used in the terrestrial surface characterization.
- Morphological distances in classification and clustering of geomorphologic units such as basins, sub-basins, and watersheds represented in planar form (Vardhan et al. 2013) and terrestrial surficial data represented in DEM form (Sagar and Lim 2015).
- Multiscale morphological operations, skeletonization by zones of influence (SKIZ), morphological pruning, thinning, thickening, and Hit-Or-Miss Transformation were employed in deriving a host of allometric power-law relationships in channel networks, water bodies, watersheds, and several other geomorphologic phenomena (Tay et al. 2005a, b, c; Sagar and Tien 2004; Sagar and Chockalingam 2004; Nagajothi et al. 2021). The characterization via a set of derived power-law relationships highlighted the evidence of self-organization via scaling laws – in networks, hierarchically decomposed sub-watersheds, and water bodies and their zones of influence, which evidently belong to different universality classes – which possess excellent agreement with geomorphologic laws such as Horton's laws, Hurst exponents, Hack's exponent, and other power-laws given in non-geoscientific context (Sagar 1996, 1999, 2000a; Sagar et al. 1998a, b, 1999). This aspect is further extended based on intuitive arguments that these universal scaling laws possess limited utility in exploring possibilities to relate them with geomorphologic processes. These arguments formed the basis to provide alternative methods that yield scale-invariant but shape-dependent power laws.
- Morphological shape decompositions in the characterization of hillslope region by deriving scale-invariant but shape-dependent measures (Sagar and Chockalingam 2004; Chockalingam and Sagar 2005) to quantitatively characterize the spatiotemporal terrestrial complexity that explains the commonly sharing physical mechanisms involved in terrestrial phenomena and processes.
- Granulometries in surficial roughness characterization of the basins and watersheds partitioned from the DEMs (Tay et al. 2005c, 2007; Nagajothi and Sagar 2019).
- Multiscale openings and closings in the modeling of the geomorphologic processes, in discrete space, under the perturbations caused due to cascade forces (flood–drought, expansion–contraction, uplift–erosion, protruding–flattening, and shortening–amplification) in a nonlinear fashion mimicking the realistic situations (Sagar et al. 1998b). Morphological modeling of the geomorphologic phenomena and processes provides unique contributions to simulations of geomorphologic networks (Sagar et al. 1998a, 2001), terrestrial surfaces (Sagar and Murthy 2000), water bodies and sand dunes (Sagar 2000b, 2001) laws of geomorphic structures under the perturbations created through an interplay between numerical simulations and graphic analysis (Sagar et al. 1998b), and understanding spatial and/or temporal behaviors of certain evolving and dynamic geomorphic phenomena (Sagar 2005).
- Morphological interpolations, which provide an effective way to transform sparse data and/or information into dense data and/or information (Sagar 2010; Sagar and Lim 2015; Challa et al. 2018), in spatiotemporal visualizations of (geomorphologic) phenomena. While developing spatiotemporal geomorphologic models that provide a better understanding of the behavioral patterns of the geo(morpho)logic phenomena and processes, the availability of the relevant data at the time intervals as close as possible is important.

Mathematical morphology – A robust mathematical theory offers a plethora of operations and transformations that provide the opportunity to transform the way quantitative geomorphologists examine terrestrial surfaces. Applications of numerous mathematical morphological transformations, which are new to the geomorphologic community, provide insights for quantitative geomorphologists. For the illustrations appropriate for section “[Mathematical Morphology in Quantitative Geomorphology](#)”, the reader may refer to the relevant cross-referenced chapters.

Landscape Evolution Models (LEMs)

The geomorphic studies explain the origin and dynamics of the topography over the Earth's surface. Various physical and chemical processes, along with actions of different geomorphic agents at long time scales, lead to variability across the landscape. Tools in mathematical morphology are usually utilized to quantitatively characterize a landscape and its trajectory with time (Fig. 2). Further, to explain the time-dependent evolution of the topography, there is a need to model the evolution processes mathematically. LEMs integrate the dynamics of the different landforms to develop the evolutionary trajectory of landscape with the help of various

equations corresponding to different processes at the landform scale. LEMs consider spatial locations as grids or TINs, and output from the equations flows across the grid to develop a feedback mechanism in a landscape. LEM not only explains the temporal variation of a landscape but more generally indicates a process–response system due to the interaction of landform and formative processes. LEM aims to describe the changes in the aggregate of landform shape, size, and relief over the smallest to the largest spatial scale. Integration of scales is a major challenge in the geomorphic system. The mathematical expressions of landscape and its constituent landforms help to integrate the processes at different scales in a given landscape and provide an evolutionary trajectory of landscape or explain major landscape pattern. The landform-scale equations are mostly derived from the functional approach, while its integration leads to the evolutionary trajectory of the landscape. Hence, LEMs not only integrate processes at different scales, but also help to integrate two approaches of geomorphic studies, i.e., functional and evolutionary approaches.

Various governing equations for landscape evolution are too complex to obtain coupled closed-form solutions, and therefore, numerical solutions methods are necessary to obtain the solutions. LEMs generally require the following steps: (1) continuity of mass, (2) geomorphic transport equations for movement of water and sediment, and (3) equations for erosion, transport, and deposition by geomorphic agents (Tucker and Hancock 2010). Continuity of mass is expressed as

$$\frac{\partial \bar{\rho} \eta}{\partial t} = \rho_{sc} B - \nabla \cdot (\rho_s \mathbf{q}_s) \quad (39)$$

where η is the height of land surface relative to a datum, t is time, $\mathbf{q}_s$ is a vector depicting transport rate, B is a source term (maybe uplift or subsidence), ∇ is the divergence operator, $\bar{\rho}$ is the average density of rock/sediment, and ρ_{sc} is the density of source materials.

Geomorphic transport functions are the main component of any numerical LEM. The theory of landscape evolution, therefore, requires complete knowledge of weathering, biological activities, and action of other geomorphic agents. Anderson (2002) presented a simple approach:

$$\frac{dz_b}{dt} = \min[P_{s0} + bH, P_{s1} \exp(H/H^*)] \quad (40)$$

where $\frac{dz_b}{dt}$ is the rate of weathering, “min” is “take the minimum of,” P_{s0} is the production of bare bedrock, b scales the production rate increase due to increase in thickness of regolith, P_{s1} defines the intercept of the exponential decay, and H^* scales the exponential decline of weathering rate for large

regolith thicknesses. This equation shows that if the thickness of regolith decreases below a critical thickness, the rate of production also decreases. Although this exponential decay rule successfully predicts the weathering processes, there is a further need to include the physiochemical processes of weathering to establish mathematical models in landscape evolution.

The movement of water and sediment is another problem in landscape evolution modeling. In the low-slope region, soil creep is a primary factor that transports the sediment down the slope. A simple linear or nonlinear equation (Eq. 6) for creep is utilized to model the creep. Although this equation can explain the transport of materials, this does not explain the fast transport on the higher slope. This further needs a nonlinear transport function as follows:

$$\mathbf{q}_s = \frac{K_D \eta}{1 - (|\nabla \eta| S_c)^a} \quad (41)$$

where S_c is the threshold gradient that represents the point of failure (Howard 1994; Roering et al. 1999).

Rivers are another important feature of a landscape at a larger spatial scale. The erosion driven by the water in a river depends on the ability to transport large materials and erode its bed. The detachment limited model offers to model the river processes where the sediment is efficiently transported downstream. Expression of the 1D detachment limited erosion model is

$$\frac{\partial h}{\partial t} = U - K A^m \frac{\partial h^n}{\partial x} \quad (42)$$

where h is elevation, and U is the uplift rate. The term $K A^m \frac{\partial h^n}{\partial x}$ represents channel incision as a function of driving force (Eqs. 11–13), where basin area is used as a proxy for discharge, and K is the erodibility parameter. The rate of erosion per unit time E depends on the excess bed shear stress (Eq. 21).

Bed shear stress is also responsible for sediment transport processes (Eq. 17). The overall 1D model for landscape evolution with detachment limited stream power model and linear hillslope diffusion model can be expressed as

$$\frac{\partial h}{\partial t} = U - K A^m \left| \frac{\partial h}{\partial x} \right|^n + D \frac{\partial^2 h}{\partial x^2} \quad (43)$$

where D is the diffusion coefficient.

Isostatic rebound is another important component in the LEM. Isostatic rebound can be associated with surface erosion where the topography is relaxed due to the removal of surface load:

$$D \left(\frac{\partial^4 w}{\partial x^2} + 2 \frac{\partial^4 w}{\partial x^2 \partial y^2} + \frac{\partial^4 w}{\partial y^4} \right) = \Delta \rho g w + \rho_s g \Delta h \quad (44)$$

where w is the surface displacement due to isostatic adjustment owing to increment of erosion Δh , ρ_s is the density of rocks at the surface, $\Delta \rho$ is the density contrast between the surface and asthenosphere, and D is the flexural rigidity. Studies have shown that without tectonic components the landscape can be elevated by several hundreds of meters with isostatic adjustment due to denudational unloading (Gilchrist and Summerfield 1990).

These equations are usually solved with the help of different numerical methods. These equations are always coupled equations, and therefore, they can be simply solved by an explicit finite difference scheme. However, research has shown that implicit methods are also very efficient and do not require several checks for the stability of the solution (Braun and Willett 2013; Fagherazzi et al. 2002). Currently, Landlab is one of the most widely used Python-based numerical LEMs that solves various problems of surface, subsurface, and biological processes (Barnhart et al. 2020). Landlab numerical model contains different separate process components that can be integrated for high spatial and temporal resolution problems. Large spatiotemporal-scale models aided by the availability of big data are the current trend in this field of research. Although highly complex models are necessary for obtaining significant results, they are not sufficient to model highly complex natural processes. Therefore, all models have uncertainty. The important step to overcome the issue of uncertainty is to focus on the correctness of building a model as close as possible to reality. However, an increase in the complexity does not necessarily provide better or desired results; rather, this can attenuate important model outcomes.

Geomorphic Connectivity

In recent times, the concept of connectivity is an emergent property of geomorphic systems (Wohl 2017). Any geomorphic system is composed of different landforms, and there exists a hierarchical link between the operative geomorphic processes (Brierley et al. 2006). The term connectivity can be defined as the efficiency of the water, sediment, and nutrient transport from one component to another. Therefore, the study of geomorphic connectivity provides an opportunity to understand the interrelationship and interdependencies of the components (landforms) of a geomorphic system (landscape). Wohl et al. (2019) have explained three types of geomorphic connectivity: (1) sediment connectivity, (2) landscape connectivity, and (3) hydrologic connectivity. The nature of connectivity could be structure-based (structural connectivity) or process-based (functional connectivity) (Jain and Tandon 2010; Wainwright et al. 2011).

Quantification of connectivity and prediction remains a major research question for the last couple of decades. A major focus has been given to quantifying the functional connectivity of sediment through a different element of a landscape (e.g., hillslope to rivers). Connectivity index (IC) based on an empirical approach is a common tool to quantitatively express the nature of connectivity in a landscape. A commonly used IC for a river basin is expressed as (after Borselli et al. (2008))

$$IC = \log_{10} \left(\frac{D_{up}}{D_{dn}} \right) \quad (45)$$

where D_{up} and D_{dn} are upslope and downslope components of connectivity, respectively. The upslope component of sediment connectivity is the potential for the downslope movement of the available sediment and is estimated as

$$D_{up} = \overline{W} \overline{S} \sqrt{A} \quad (46)$$

where $\overline{W}$ is the average weighting factor of the contributing area in the upslope, $\overline{S}$ is the average slope of the upslope area, and A is the contributing area.

D_{dn} is the component that considers the flow path length for a particle to reach the immediate sink and is estimated as

$$D_{dn} = \sum_i \frac{d_i}{W_i S_i} \quad (47)$$

where d_i is the length of the flow path along the i th cell following the steepest downslope direction, and W_i and S_i are the weighting factor and gradient of the i -th cell, respectively. Cavalli et al. (2013) suggested further modification to the slope factor, upstream contributing area, and weighting factors and suggested that these simple methods can help to understand the sediment dynamics in a very complex morphodynamic system. They also suggested that the sediment delivery pathways, amount of vegetation, and channel bed morphology are other important factors for understanding the connectivity.

Recently, the application of graph theory to quantify connectivity resulted in new insights into the sediment mobilization process. Heckmann and Schwanghart (2013) pioneered an approach to describe the geomorphic system as a network consisting of nodes (sediment sources) and edges (links between geomorphic processes). A graph with n nodes can be represented as $n \times n$ adjacency matrix A . Depending on the topological and spatial characteristics, the fluxes are associated with the edges that transport the sediment from upstream to downstream. The betweenness centrality index (B) is the measure of the extent to the presence of a node (i) between other nodes:

$$B_i = \frac{\sum n_{ijk}}{n_{jk}} \quad (48)$$

where n_{ijk} is the number of edges or path that is present between node j and node k and that pass through i , and n_{jk} is the total number of edges or paths between node j and node k .

The Shimbel index (Shi) is the factor that considers the distance between nodes and checks if the location of the nodes changes the total possible paths within the network:

$$Shi_i = \frac{\sum d_{ij}}{\sum d_{jk}} \quad (49)$$

If the Shimbel index is high, then the node contributes to creating long paths within the network; if the Shimbel index is low, then the compactness of the network is maximized by the nodes (Cossart and Fressard 2017). For the local scale, the network structural connectivity index is developed based on the potential fluxes and the accessibility of the transportation of flux (Cossart and Fressard 2017).

$$NSC_i = \frac{F_i}{Shi_i} \quad \text{and} \quad F_i = \frac{\sum F_{ijo}}{\sum F_{jo}} \quad (50)$$

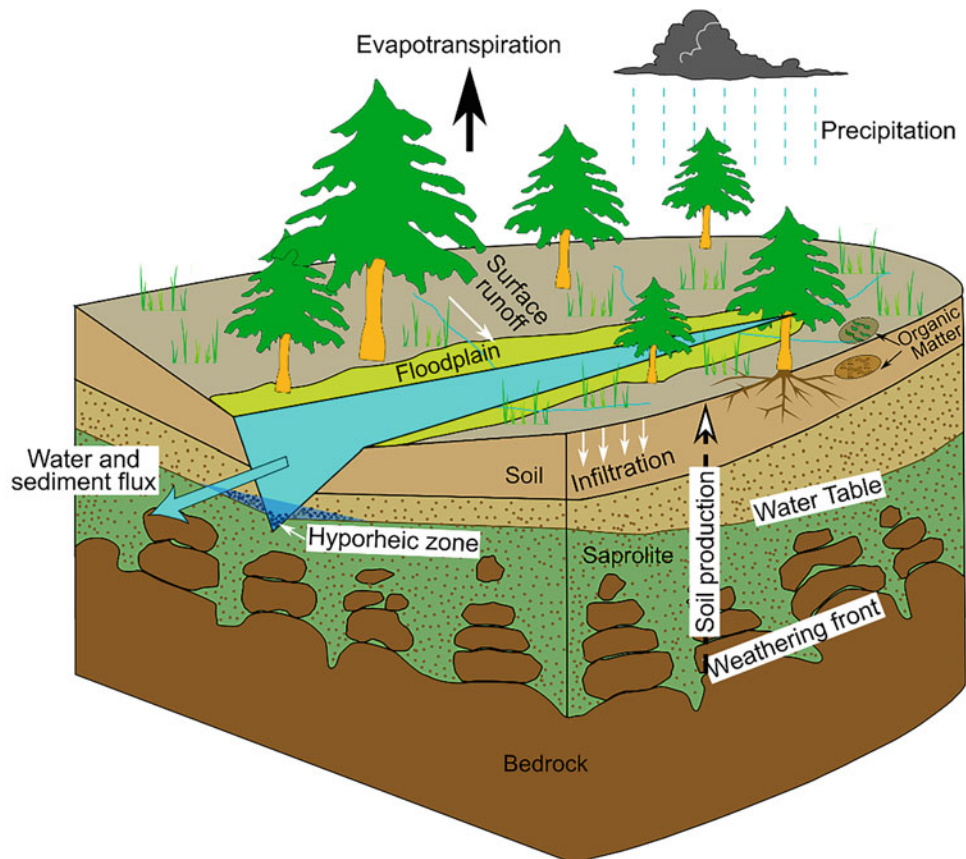
F_{ijo} and F_{jo} are calculated from the reconstruction of sediment pathways throughout the cascade. These methodologies have successfully described the scale-dependent sediment mobilization and delivery behavior in a catchment. Additionally, as the graph theory also considers the structural connectivity between the nodes (or sediment sources), hotspots of geomorphic works can be identified.

Critical Zones

The topmost layer of the Earth's surface, i.e., from the bedrock to the lower atmosphere, is termed "critical zone" (Fig. 5). This term was introduced by the National Science Foundation (NRC 2001), defined as "... the heterogeneous, near-surface environment in which complex interactions involving rock, soil, water, air and living organisms regulate the natural habitat and determine availability of life-sustaining resources" (NRC 2001). Therefore, this zone connects the atmosphere, lithosphere, hydrosphere, and biosphere for various physiochemical processes. The experimental setup to study this thin layer is known worldwide as critical zone observatories (CZO). Recently, several CZOs were set up around the world to understand the

Quantitative Geomorphology,

Fig. 5 Schematic diagram showing major components of critical zone observatory. (Modified after Riebe et al. 2017)



coupled processes and their outcome on the natural world (Godd  ris and Brantley 2013).

Mathematical models for the critical zone are essentially needed to quantify the rate of weathering and soil production in order to address soil sustainability. Therefore, an integrated model has been established to link the soil formation, nutrients' transport, and transport of water (Giannakis et al. 2017). The water flow is expressed by Richard's equation:

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial z} \left(k \frac{\partial h}{\partial z} \right) + S \quad (51)$$

where θ is the volumetric water content, t is time, z is the vertical coordinate, k is the unsaturated hydraulic conductivity, h is the pressure head, and S is the source of water. A knowledge-based reactive transport model is generally used to understand the time-varying water saturation (Aguilera et al. 2005). Representation of 1D continuum coupled mass transport and chemical reactions can be expressed mathematically as

$$\xi \frac{\partial C_j}{\partial t} = \left[\frac{\partial}{\partial x} \left(D \cdot \xi \cdot \frac{\partial C_j}{\partial x} \right) - \frac{\partial}{\partial x} (v \cdot \xi \cdot C_j) \right]_j + \sum_{k=1}^{N_r} \lambda_{k,j} \cdot \sigma_k; \quad j = 1, \dots, m \quad (52)$$

where t is time, and x is the position along the 1D spatial domain. The two terms in the square bracket are part of the transport operator (T) whereas the last term is the sum of all transformation processes. j is a particular species and C_j is the concentration of that species. ξ , v and D are variables that depends on the environmental system. σ_k represents the rate of the k -th kinetic reaction and $\lambda_{k,j}$ is the stoichiometric coefficient of species.

Although the critical zone is the fragile skin of the earth, the role of groundwater is an important factor that sustains the flora and fauna. If the groundwater table is raised, then it causes an increase in soil moisture and evapotranspiration, and further affects regional climate in the long term. Climate models without surface water–groundwater interactions tried to assess the interaction between climate–soil vegetation dynamics. The results show that vegetation properties can alter groundwater table dynamics (Leung et al. 2011). Numerical models are also utilized to understand surface water–groundwater interaction. MODFLOW's streamflow equations are utilized to assess the connectivity between surface and groundwater (Brunner et al. 2010). The exchange of volumetric flux between the surface water and groundwater is estimated by

$$Q_{SG} = \frac{K_c L w}{h_c} (h_{riv} - h) = c_{riv} (h_{riv} - h) \quad (53)$$

where K_c is the hydraulic conductivity of the clogging layer, L is the length of the river within a cell, w is the width of the river, h_c [L] is the thickness of the clogging layer, h_{riv} is the hydraulic head of the river, h is the groundwater head, and c_{riv} is the conductance of the clogging layer (McDonald and Harbaugh 1988).

Critical zone science has brought new opportunities and challenges for quantitative geomorphologists. Understanding the feedback mechanism between different processes using large datasets requires mathematical models. Handling multiple datasets from various sources and assimilation in a single generic modeling framework is complex. It is expected that the assimilation of large datasets with mathematical modeling will yield new insight into process interactions.

Conclusions

Individual geomorphic agent-based processes responsible for landscape evolution have been studied for several decades in the early twentieth century. Although these studies developed process-based understanding, the recent research queries are advancing to integrate the processes at different scales across different landforms or components. Quantitative geomorphology deals with this integration of different physiochemical processes to elucidate the landscape dynamics in a more holistic way. With the advent of large field data from different processes, comparison of model output with the real-world scenario is now easier to achieve. Furthermore, an interdisciplinary approach has brought several physics-based and mathematical tools to understand the geomorphic system. The use of coupled differential equations (ODE and PDE), statistical techniques, and many mathematical tools has helped to better explain the geomorphologic processes at different spatial and temporal scales. Furthermore, the use of long-term LEMs helps to assess the relative contribution of different geomorphic, tectonic, and climatic agents. Such quantitative approaches also help to integrate functional and evolutionary branches of geomorphology.

Quantitative geomorphology offers several approaches to derive indexes of relevance to geometry, topology, and morphology of terrestrial surfaces. Such indices would be of immense use in developing parameter-specific models through which one can make predictions of the dynamical behavior of the terrestrial surfaces. What is more interesting is to develop approaches to construct "geomorphologic attractors" that eventually provide short- and long-term behavioral predictions.

Cross-References

- [Digital Elevation Model](#)
- [Earth Surface Processes](#)
- [Mathematical Morphology](#)
- [Morphometry](#)
- [Scaling and Scale Invariance](#)

Bibliography

- Aguilera DR, Jourabchi P, Spiteri C, Regnier P (2005) A knowledge-based reactive transport approach for the simulation of biogeochemical dynamics in earth systems. *Geochem Geophys Geosyst* 6(7)
- Anderson RS (2002) Modeling the tor-dotted crests, bedrock edges, and parabolic profiles of high alpine surfaces of the Wind River Range. *Wyoming Geomorphol* 46(1–2):35–58
- Anderson RS, Anderson SP (2010) *Geomorphology: the mechanics and chemistry of landscapes*. Cambridge University Press
- Anderson RS, Hallet B (1986) Sediment transport by wind: toward a general model. *Geol Soc Am Bull* 97(5):523–535
- Asselman NEM (2000) Fitting and interpretation of sediment rating curves. *J Hydrol* 234:228–248
- Bagnold RA (1941) *The physics of blown sand and desert dunes*. Methuen, London
- Bagnold RA (1966) *An approach to the sediment transport problem from general physics*. US Government Printing Office
- Barnhart KR, Hutton EW, Tucker GE, Gasparini NM, Istanbuloglu E, Hobbey DE, Lyons NJ, Mouchene M, Nudurupati SS, Adams JM, Bandaragoda C (2020) Landlab v2. 0: a software package for earth surface dynamics. *Earth surface. Dynamics* 8(2):379–397
- Beucher S (1999) Mathematical morphology and geology: when image analysis uses the vocabulary of earth science: a review of some applications. *Proceedings of geovision—99*. University of Liege, Belgium, pp 6–7
- Borselli L, Cassi P, Torri D (2008) Prolegomena to sediment and flow connectivity in the landscape: a GIS and field numerical assessment. *Catena* 75(3):268–277
- Braun J, Willett SD (2013) A very efficient O (n), implicit and parallel method to solve the stream power equation governing fluvial incision and landscape evolution. *Geomorphology* 180:170–179
- Brierley G, Fryirs K, Jain V (2006) Landscape connectivity: the geographic basis of geomorphic applications. *Area* 38(2):165–174
- Brunner P, Simmons CT, Cook PG, Therrien R (2010) Modeling surface water-groundwater interaction with MODFLOW: some considerations. *Groundwater* 48(2):174–180
- Cavalli M, Trevisani S, Comiti F, Marchi L (2013) Geomorphometric assessment of spatial sediment connectivity in small Alpine catchments. *Geomorphology* 188:31–41
- Challa A, Danda S, Sagar BSD, Najman L (2018) Some properties of interpolations using mathematical morphology. *IEEE Trans Image Process* 27(4):2038–2048
- Chang HH (1985) River morphology and thresholds. *J Hydraul Eng* 111(3):503–519
- Chockalingam L, Sagar BSD (2003) Mapping of sub-watersheds from digital elevation model: a morphological approach. *Int J Pattern Recognit Artif Intell* 17(02):269–274
- Chockalingam L, Sagar BSD (2005) Morphometry of network and nonnetwork space of basins. *J Geophys Res Solid Earth* 110(B8)
- Cook SJ, Swift DA, Kirkbride MP, Knight PG, Waller RI (2020) The empirical basis for modelling glacial erosion rates. *Nat Commun* 11(1):1–7
- Cossart É, Fressard M (2017) Assessment of structural sediment connectivity within catchments: insights from graph theory. *Earth Surf Dyn* 5(2):253–268
- Delaney I, Bauder A, Werder MA, Farinotti D (2018) Regional and annual variability in subglacial sediment transport by water for two glaciers in the Swiss Alps. *Front Earth Sci* 6:175
- Dinesh S, Radhakrishnan P, Sagar BSD (2007) Morphological segmentation of physiographic features from DEM. *Int J Remote Sens* 28(15):3379–3394
- Draebing D, Krautblatter M (2019) The efficacy of frost weathering processes in alpine rockwalls. *Geophys Res Lett* 46(12):6516–6524
- Fagherazzi S, Howard AD, Wiberg PL (2002) An implicit finite difference method for drainage basin evolution. *Water Resour Res* 38(7):21–21
- Fedo CM, Eriksson KA, Krogstad EJ (1996) Geochemistry of shales from the Archean (3.0 Ga) Buhwa Greenstone Belt, Zimbabwe: implications for provenance and source-area weathering. *Geochim Cosmochim Acta* 60:1751–1763
- Fernandes NF, Dietrich WE (1997) Hillslope evolution by diffusive processes: the timescale for equilibrium adjustments. *Water Resour Res* 33(6):1307–1318
- Fryirs KA, Brierley GJ (2012) *Geomorphic analysis of river systems: an approach to reading the landscape*. Wiley, Chichester, p 362
- Gaillardet J, Dupré B, Louvat P, Allegre CJ (1999) Global silicate weathering and CO₂ consumption rates deduced from the chemistry of large rivers. *Chem Geol* 159(1–4):3–30
- Giannakis GV, Nikolaidis NP, Valstar J, Rowe EC, Moirgiorgiou K, Kotronakis M, Paranychiakis NV, Rousseva S, Stamati FE, Banwart SA (2017) Integrated critical zone model (1D-ICZ): a tool for dynamic simulation of soil functions and soil structure. *Adv Agron* 142:277–314
- Gilchrist AR, Summerfield MA (1990) Differential denudation and flexural isostasy in formation of rifted-margin upwarps. *Nature* 346(6286):739–742
- Goddéris Y, Brantley SL (2013) *Earthcasting the future Critical Zone*. Elementa: Science of the Anthropocene, 1
- Harbor JM (1992) Numerical modeling of the development of U-shaped valleys by glacial erosion. *Geol Soc Am Bull* 104(10):1364–1375
- Harnois L (1988) The CIW index: a new chemical index of weathering. *Sediment Geol* 55:319–322
- Hartmann J, Jansen N, Dürr HH, Kempe S, Köhler P (2009) Global CO₂-consumption by chemical weathering: what is the contribution of highly active weathering regions? *Glob Planet Chang* 69(4):185–194
- Heckmann T, Schwanghart W (2013) Geomorphic coupling and sediment connectivity in an alpine catchment—exploring sediment cascades using graph theory. *Geomorphology* 182:89–103
- Holthuijsen LH (2010) *Waves in oceanic and coastal waters*. Cambridge University Press
- Howard AD (1994) A detachment-limited model of drainage basin evolution. *Water Resour Res* 30(7):2261–2285
- Jain V, Preston N, Fryirs K, Brierley G (2006) Comparative assessment of three approaches for deriving stream power plots along long profiles in the upper Hunter River catchment, New South Wales, Australia. *Geomorphol* 74(1–4):297–317
- Jain V, Sinha R (2003) Derivation of unit hydrograph from GIUH analysis for a Himalayan river. *Water Resour Manag* 17(5):355–376
- Jain V, Sinha R (2006) Evaluation of geomorphic control on flood hazard through geomorphic instantaneous unit hydrograph. *Curr Sci* 85(11):1596–1600
- Jain V, Tandon SK (2010) Conceptual assessment of (dis) connectivity and its application to the Ganga River dispersal system. *Geomorphology* 118(3–4):349–358

- Jordan G (2003) Morphometric analysis and tectonic interpretation of digital terrain data: a case study. *Earth Surface Proc Landforms: J Br Geomorphol Res Group* 28(8):807–822
- Komar PD (1998) *Beach processes and sedimentation*, 2nd edn. Prentice-Hall, Englewood Cliffs
- Leenders JK, Sterk G, Van Boxel JH (2011) Modelling wind-blown sediment transport around single vegetation elements. *Earth Surface Proc Landform* 36(9):1218–1229
- Leung LR, Huang M, Qian Y, Liang X (2011) Climate–soil–vegetation control on groundwater table dynamics and its feedbacks in a climate model. *Clim Dyn* 36(1):57–81
- Lu H, Shao Y (2001) Toward quantitative prediction of dust storms: an integrated wind erosion modelling system and its applications. *Environ Model Softw* 16(3):233–249
- Matheron G (1975) *Random sets and integral geometry*. Wiley, New York, p 468
- Mayaud JR, Bailey RM, Wiggs GF, Weaver CM (2017) Modelling aeolian sand transport using a dynamic mass balancing approach. *Geomorphology* 280:108–121
- McDonald MG, Harbaugh AW (1988) A modular, three-dimensional finite-difference ground-water flow model. USGS, Reston
- McFadden LD, Eppes MC, Gillespie AR, Hallet B (2005) Physical weathering in arid landscapes due to diurnal variation in the direction of solar heating. *Geol Soc Am Bull* 117(1–2):161–173
- Meyer-Peter E, Müller R (1948) Formulas for bed-load transport. *Proceedings of the second Meeting of the International Association for Hydraulic Structures Research*, pp 39–64
- Nagajothi K, Sagar BSD (2019) Classification of geophysical basins derived from SRTM and Cartosat DEMs via directional granulometries. *IEEE J Select Topics Appl Earth Observ Remote Sensing* 12(12):5259–5267
- Nagajothi K, Rajasekhara HM, Sagar BSD (2021) Universal fractal scaling laws for surface water bodies and their zones of influence. *IEEE Geosci Remote Sens Lett* 18(5):781–785
- Nesbitt HW, Young GM (1982) Early Proterozoic climates and plate motions inferred from major element chemistry of lutites. *Nature* 299:715–717
- NRC (2001) *Basic research opportunities in earth science*. National Academies Press, Washington, DC
- Parker A (1970) An index of weathering for silicate rocks. *Geol Mag* 107:501–504
- Pelletier JD (2008) *Quantitative modeling of earth surface processes*. Cambridge University Press
- Piégay H (2019) Quantitative geomorphology. In: *International encyclopedia of geography: people, the earth, environment and technology: people, the earth, environment and technology*, pp 1–3. <https://doi.org/10.1002/9781118786352.wbieg0417.pub2>
- Raymond CF, Harrison WD (1987) Fit of ice motion models to observations from Variegated Glacier, Alaska. In: Waddington ED, Walder JS (eds) *The physical basis of ice sheet modelling*, vol 170. International Association of Hydrologic Sciences Publication, pp 153–166
- Riebe CS, Hahn WJ, Brantley SL (2017) Controls on deep critical zone architecture: a historical review and four testable hypotheses. *Earth Surf Process Landf* 42(1):128–156
- Rodríguez-Iturbe I, Valdés JB (1979) The geomorphologic structure of hydrologic response. *Water Resour Res* 15(6):1409–1420
- Roering JJ, Kirchner JW, Dietrich WE (1999) Evidence for nonlinear, diffusive sediment transport on hillslopes and implications for landscape morphology. *Water Resour Res* 35(3):853–870
- Różycka M, Jancewicz K, Migoń P, Szymanowski M (2021) Tectonic versus rock-controlled mountain fronts—Geomorphometric and geostatistical approach (Sowie Mts., Central Europe). *Geomorphology* 373:107485
- Ruxton BP (1968) Measures of the degree of chemical weathering of rocks. *J Geol* 76:518–527
- Sagar BSD (1996) Fractal relations of a morphological skeleton. *Chaos, Solitons Fractals* 7(11):1871–1879
- Sagar BSD (1999) Estimation of number-area-frequency dimensions of surface water bodies. *Int J Remote Sens* 20(13):2491–2496
- Sagar BSD (2000a) Letter to editor's fractal relation of medial axis length to the water body area. *Discret Dyn Nat Soc* 4(2):97–97
- Sagar BSD (2000b) Multi-fractal-inter-slipface angle curves of a morphologically simulated sand dune. *Discret Dyn Nat Soc* 5(1):71–74
- Sagar BSD (2001) Generation of self organized critical connectivity network map (SOCCNM) of randomly situated surface water bodies. *Letters to Editor, Discrete Dynamics in Nature and Society* 6(3): 225–228
- Sagar BSD (2005) Discrete simulations of spatio-temporal dynamics of small water bodies under varied streamflow discharges. *Nonlinear Process Geophys* 12(1):31–40
- Sagar BSD (2010) Visualization of spatiotemporal behavior of discrete maps via generation of recursive median elements. *IEEE Trans Pattern Anal Mach Intell* 32(2):378–384
- Sagar BSD (2013) *Mathematical morphology in geomorphology and GISci*. Chapman & Hall, Taylor & Francis Group, p 546
- Sagar BSD (2020) Digital elevation models: an important source of data for geoscientists [education]. *IEEE Geosci Remote Sensing Magazine* 8(4):138–142
- Sagar BSD, Chockalingam L (2004) Fractal dimension of non-network space of a catchment basin. *Geophys Res Lett* 31(12):L12502
- Sagar BSD, Lim SL (2015) Ranks for pairs of spatial fields via metric based on grayscale morphological distances. *IEEE Trans Image Process* 24(3):908–918
- Sagar BSD, Murthy KSR (2000) Generation of fractal landscape using nonlinear mathematical morphological transformations. *Fractals* 8(3):267–272
- Sagar BSD, Serra J (2010) Spatial information retrieval, analysis, reasoning and modelling. *Int J Remote Sensing* 31(22):5747–5750
- Sagar BSD, Tien TL (2004) Allometric power-law relationships in a Hortonian fractal digital elevation model. *Geophys Res Lett* 31(6): L06501. <https://doi.org/10.1029/2003GL019093>
- Sagar BSD, Omeregic C, Rao BSP (1998a) Morphometric relations of fractal-skeletal based channel network model. *Discret Dyn Nat Soc* 2(2):77–92
- Sagar BSD, Venu M, Gandhi G, Srinivas D (1998b) Morphological description and interrelationship between force and structure: a scope to geomorphic evolution process modelling. *Int J Remote Sens* 19(7):1341–1358
- Sagar BSD, Venu M, Murthy KSR (1999) Do skeletal network derived from water bodies follow Horton's laws? *J Math Geol* 31(2):143–154
- Sagar BSD, Venu M, Srinivas D (2000) Morphological operators to extract channel networks from digital elevation models. *Int J Remote Sens* 21(1):21–30
- Sagar BSD, Srinivas D, Rao BSP (2001) Fractal skeletal based channel networks in a triangular initiator basin. *Fractals* 9(4):429–437
- Sagar BSD, Murthy MBR, Rao CB, Raj B (2003) Morphological approach to extract ridge-valley connectivity networks from Digital Elevation Models (DEMs). *Int J Remote Sens* 24(3):573–581
- Sahoo R, Singh RN, Jain V (2020) Process inference from topographic fractal characteristics in the tectonically active Northwest Himalaya, India. *Earth Surf Proc Landforms* 45(14):3572–3591
- Serra J (1982) *Image analysis and mathematical morphology*. Academic Press, London, p 610
- Sonam, Jain V (2018) Geomorphic effectiveness of a long profile shape and the role of inherent geological controls in the Himalayan hinterland area of the Ganga River basin, India. *Geomorphology* 304: 15–29
- Strahler AN (1952) Hypsometric (area-altitude) analysis of erosional topography. *Geol Soc Am Bull* 63(11):1117–1142

- Takagi H, Quan NH, Anh LT, Thao ND, Tri VPD, Anh TT (2019) Practical modelling of tidal propagation under fluvial interaction in the Mekong Delta. *Int J River Basin Manag* 17(3):377–387
- Tay LT, Sagar BSD, Chuah HT (2005a) Analysis of geophysical networks derived from multiscale digital elevation models: a morphological approach. *IEEE Geosci Remote Sens Lett* 2(4):399–403
- Tay LT, Sagar BSD, Chuah HT (2005b) Allometric relationships between travel-time channel networks, convex hulls, and convexity measures. *Water Resour Res* 46(2)
- Tay LT, Sagar BSD, Chuah HT (2005c) Derivation of terrain roughness indicators via Granulometries. *Int J Remote Sens* 26(18):3901–3910
- Tay LT, Sagar BSD, Chuah HT (2007) Granulometric analysis of basin-wise DEMs: a comparative study. *Int J Remote Sens* 28(15):3363–3378
- Tucker GE, Hancock GR (2010) Modelling landscape evolution. *Earth Surf Process Landf* 35(1):28–50
- Tucker GE, Slingerland RL (1994) Erosional dynamics, flexural isostasy, and long-lived escarpments: a numerical modeling study. *J Geophys Res: Solid Earth* 99(B6):12229–12243
- Van den Berg JH (1995) Prediction of alluvial channel pattern of perennial rivers. *Geomorphology* 12(4):259–279
- Vardhan SA, Sagar BSD, Rajesh N, Rajashekara HM (2013) Automatic detection of orientation of mapped units via directional granulometric analysis. *IEEE Geosci Remote Sens Lett* 10(6):1449–1453
- Vuille M, Francou B, Wagnon P, Juen I, Kaser G, Mark BG, Bradley RS (2008) Climate change and tropical Andean glaciers: past, present and future. *Earth Sci Rev* 89(3–4):79–96
- Wainwright J, Turnbull L, Ibrahim TG, Lexartza-Artza I, Thornton SF, Brazier RE (2011) Linking environmental regimes, space and time: interpretations of structural and functional connectivity. *Geomorphology* 126(3–4):387–404
- Wohl E (2017) Connectivity in rivers. *Prog Phys Geogr* 41(3):345–362
- Wohl E, Brierley G, Cadol D, Coulthard TJ, Covino T, Fryirs KA, Grant G, Hilton RG, Lane SN, Magilligan FJ, Meitzen KM (2019) Connectivity as an emergent property of geomorphic systems. *Earth Surf Process Landf* 44(1):4–26
- Yanites BJ, Tucker GE, Mueller KJ, Chen YG, Wilcox T, Huang SY, Shi KW (2010) Incision and channel morphology across active structures along the Peikang River, central Taiwan: implications for the importance of channel width. *Bulletin* 122(7–8):1192–1208
- Zhang W, Jia Q, Chen X (2014) Numerical simulation of flow and suspended sediment transport in the distributary channel networks. *J Appl Math*:2014

Quantitative Stratigraphy

Felix Gradstein¹ and Frits Agterberg²

¹University of Oslo, Oslo, Norway

²Central Canada Division, Geomathematics Section, Geological Survey of Canada, Ottawa, ON, Canada

Quantitative stratigraphy uses relatively simple or complex mathematical-statistical methods to calculate stratigraphic models that with a minimum of data provide a maximum of predictive potency, and include formulation of confidence limits. (F. P. Agterberg 1990)

Definition

The RASC method for RAnking and SCaling of biostratigraphic events was developed by the authors and associates to calculate large zonations with estimates of uncertainty. The purpose of ranking is to create an optimum sequence of fossil events observed in many different wells or sections subject to stratigraphic inconsistencies in the direction of the arrow of time. These inconsistencies, which result in crossovers of lines of correlation between sections, are due to various sampling errors and other sources of uncertainty including reworking and misclassification. They can be resolved by statistical averaging combined with stratigraphic reasoning. Subsequent scaling of the events may be carried out by estimating intervals between successive events along a relative timescale. This results in the scaled optimum sequence. Either the ranked optimum sequence or the scaled optimum sequence may be used for biozonation. The observed positions in the sections of different biostratigraphic events can have different degrees of precision. These differences may be evaluated through analysis of variance. Other types of stratigraphic events including logmarkers can be integrated and incorporated with the biostratigraphic events. Stratigraphically, the (scaled) optimum sequence represents the average order in which events occur in a sedimentary basin. The latter is realistic for practical correlations of strata.

Introduction

The purpose of this contribution is to provide a succinct description of RASC using the geomathematical method outlined in Agterberg and Gradstein (1999) and the Cretaceous microfossil data set of Gradstein et al. (1999) for illustration. The Cretaceous data set contains 1753 records on occurrences of 517 stratigraphic microfossil events in 31 exploration wells, offshore Norway.

Quantitative stratigraphy is also concerned with the correlation of lithologies or logs in different wells or outcrop sections. The emphasis in this entry is on biostratigraphic correlation although lithologies with lithostratigraphic signals can be incorporated into quantitative methods such as RASC.

Modern biostratigraphy must cope with occurrence data (events) from hundreds of fossil taxa, in thousands of samples derived from many wells or sections in a sedimentary basin. New tools in stratigraphy, using quantitative methods, make it easier to objectively build zonations integrating different microfossil groups. In addition, individual sections may be tested for “stratigraphic normality.” A prime method is called RASC for RAnking and SCaling of fossil events, and subject of this study.

RAnking and SCaling is a probabilistic method, available with Version 20. The method uses fossil event order in wells

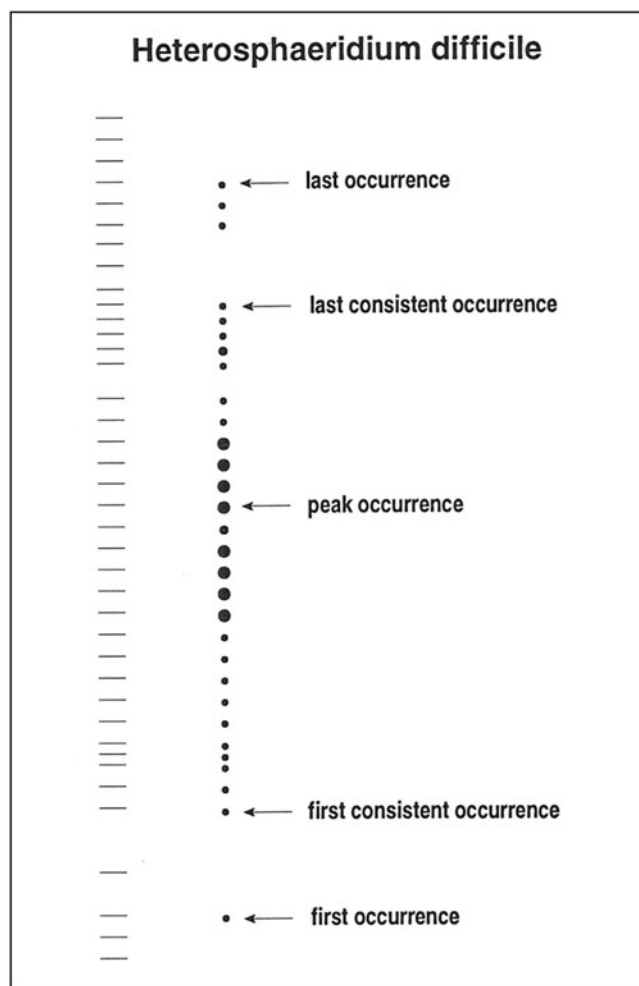
or outcrop sections to construct a most likely and average sequence of events. This optimal order is scaled in relative time using crossover frequencies of all event pairs. The method provides detailed stratigraphic error analysis, and several correlation options, the latter executed in the program called CASC. RASC with CASC operates on all wells simultaneously, is very fast, handles large and complex data sets, and is relatively insensitive to noise. Literature on the method is extensive, and there are many applications, particularly in petroleum basins (see later). Before we outline the method itself, we briefly deal with the properties of fossil data and the data input.

A paleontological record is the position of a fossil taxon in a rock sequence. The stratigraphic range of a fossil is a composite of all its records from oldest to youngest in a stratigraphic sense. The endpoints of this range are biostratigraphic events, which include the first occurrence and appearance in time and last occurrence or disappearance from the geologic record. *A biostratigraphic event is the presence of a taxon in its time context, derived from its position in a rock sequence.* The fossil events are the result of the continuing evolutionary trends of Life on Earth; they differ from physical events in that they are unique, nonrecurrent, and that their order is irreversible. Often, first and last occurrences of fossil taxa are relatively poorly defined records, based on few specimens in scattered samples. Particularly with time-wise scattered last occurrences, one may be suspicious that reworking has locally extended the record, a reason why it is useful to distinguish between last occurrence (LO), and last common or last consistent occurrence (LCO), where consistency refers to occurrence in consecutive samples in a well or outcrop section. The shortest spacing in relative time between successive fossil events is called resolution. The greater the probability that such events follow each other in time, the greater the likelihood that correlation of the event record models isochrons. Most industrial data sets make use of sets of LO and LCO events. In an attempt to increase resolution in basin stratigraphy, particularly when many sidewall cores in wells are available, up to five events may be recognized along the stratigraphic range of a fossil taxon. Such events include last stratigraphic occurrence “top” or LO event, last common or last consistent occurrence LCO event, last abundant occurrence LAO event, first common or consistent occurrence FCO event, and first occurrence FO event (Fig. 1). Unfortunately, such practice may not yield the desired increase in biostratigraphic resolution sought after, for reason of poor event traceability. Event traceability is illustrated in Fig. 2, where cumulative event distribution is plotted against the number of wells for seven exploration and Deep Sea Drilling Project data sets with different microfossil records studied by the senior author. The events stem from dinoflagellates, foraminifers, and relatively few miscellaneous microfossils. The curves are asymptotic, showing an inverse relation between event

distribution and the number of wells. None of the events occur in all wells; clearly, far fewer events occur in five or six wells than in one or two wells, and hence the cumulative frequency drops quite dramatically with a small increase in number of wells. Obviously, the majority of fossil events have poor traceability, which is true for most data sets, either from wells or from outcrops (Gradstein and Agterberg 1998). Microfossil groups with higher local species diversity, on average, have less event traceability. Data sets with above average traceability of events are those where one or more dedicated observers have spent above average time examining the fossil record, verifying taxonomic consistency between wells or outcrop sections, and searching for “missing” data. In general, routine examination of wells by consultants yields only half or much less of the taxa and events than may be detected with a slightly more dedicated approach. There are other reasons than lack of detail from analysis why event traceability is relatively low. For example, lateral variations in sedimentation rate change the diversity and relative abundance of taxa in coeval samples between wells, particularly if sampling is not exhaustive, as with well cuttings or sidewall cores. It is difficult to understand why fossil events might be locally missing. Since chances of detection depend on many factors, stratigraphical, mechanical, and statistical in nature (Fig. 3), increasing sampling and studying more than one microfossil group in detail is beneficial. Although not admitted or clarified, biostratigraphy relies almost as much on the absence, as on the presence of certain markers. This remark is particularly tailored to microfossils that generally are widespread and relatively abundant, and harbor stratigraphically useful events. Only if nonexistence of events is recognized in many well-sampled sections, may absences be construed as affirmative for stratigraphic interpretations. If few samples are available over long stratigraphic intervals, the chance to find long-ranging taxa considerably exceeds the chance to find short-ranging forms, unless the latter are abundant. In actual practice, so-called index fossils have a short stratigraphic range and generally are less common, hence easily escape detection, a reason why their absence should be used with caution (Fig. 4).

Data Selection, Run Parameters, and Unique Events

It will be obvious to the user of RASC that a set of wells or outcrop sections to be zoned should have maximal stratigraphic overlap between sections and be devoid of major breaks or disconformities, etc. Since RASC is a statistical method, it is desirable to have sufficient sections (usually 10 or more) and sufficient events for calculation. Hence, events to be zoned should occur in at least 4, 5, 6, 7, or more sections (threshold K_c), with event pairs obviously set



Quantitative Stratigraphy, Fig. 1 Events recognized along the stratigraphic range of a fossil taxon. Such events include last stratigraphic occurrence “top” or LO event, last common or last consistent occurrence LCO event, last abundant occurrence LAO event, first common or

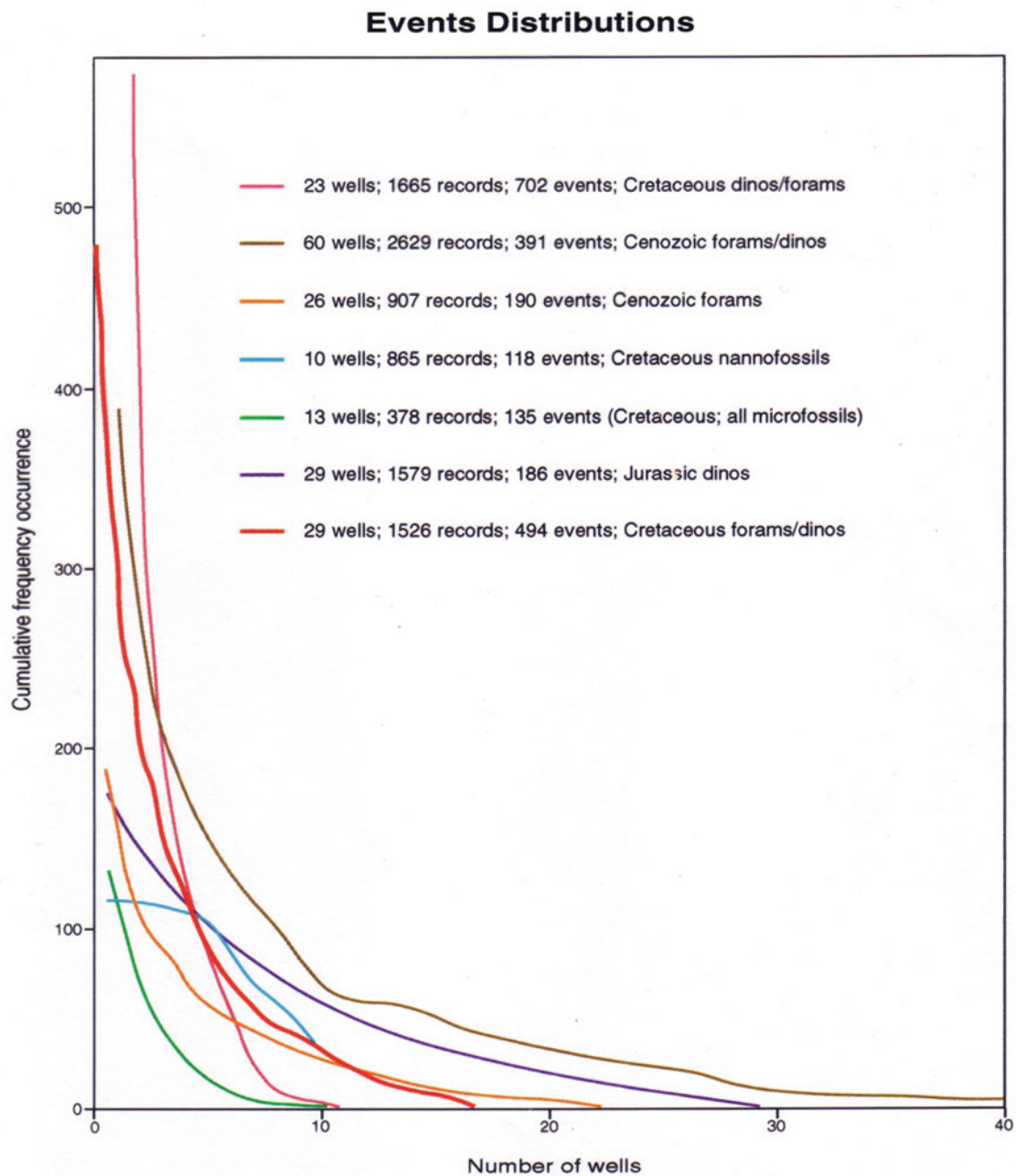
consistent occurrence FCO event, and first occurrence FO event. Unfortunately, such practice may not yield the desired increase in biostratigraphic resolution sought after, for reason of poor event traceability. Event traceability is illustrated in Fig. 3

to occur in fewer sections, e.g., 3 (threshold Mc). Index fossils, occurring in fewer than threshold Kc sections, may be introduced in the Optimum Sequence by the Unique Event routine. After calculation of the Optimum Sequence, assigned unique events are looked up in the single or few wells/sections where they occur. Their Optimum Sequence position is determined from the position of their nearest (section) neighbors above and below, known in the calculated optimum sequence. In the Optimum Sequence, these Unique Events have two stars.

QSCreator

Bundled with the RASC Programme Version 20 comes a simple routine called QSCreator, where QS stands for Quantitative Stratigraphy. QSCreator makes it fast and easy to

generate the input data, for RASC, and also provides a printable file for input data curation. The basic input for RASC consists of a list of event names dictionary (max. 999 events per dictionary, with fossil names of max. 45 characters, incl. spacing) and a well data file consisting of multiple records of the events in all wells considered (max. 150 wells, but <50 often desirable for reason of “noise” control). Many RASC applications are based on data derived from exploratory wells, as in the example of this study. In each well, the events observed to occur in the samples are characterized by their depths. However, it should be kept in mind that the method can be applied equally well to stratigraphic events observed in land-based outcrop sections. A reminder: When generating the RASC input files with QSCreator, the master dictionary should be in numeric sort mode (not alphabetic sort mode).



Quantitative Stratigraphy, Fig. 2 Cumulative frequency of micro-fossil event occurrences versus number of wells for seven subsurface data sets in the North Sea, offshore Canada and Atlantic and Indian Oceans (simplified after Gradstein and Agterberg 1999). Note the poor

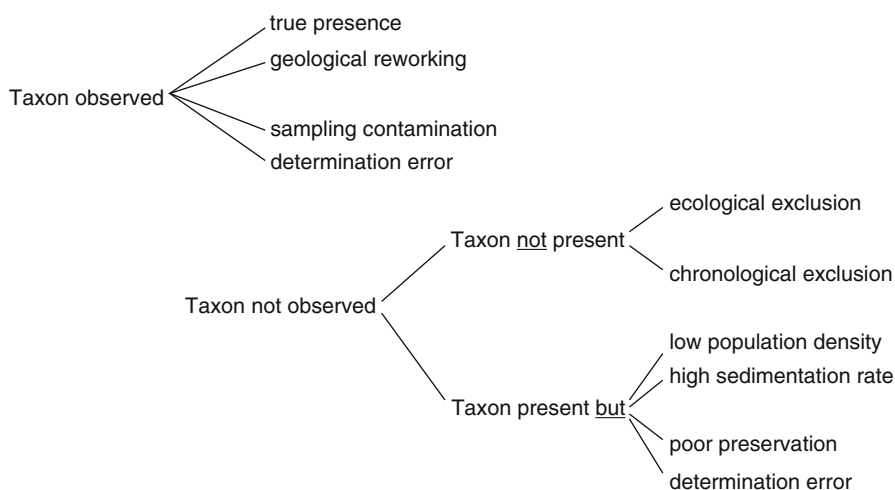
traceability, which is a property of many biological-paleontological events in nature, with all events occurring in at least one well or section, but few events occurring in many sections. This feature is important to arrive at suitable RASC runs

Ranking and Scaling

RASC is an acronym for ranking and scaling of biostratigraphic events. There have been three distinct RASC method development phases:

1. 1978–1985: Starting from an algorithm for ordering biostratigraphic events proposed by Hay (1972), the basic

concepts of RASC were introduced and applied initially to Cenozoic Foraminifera in offshore wells drilled along the northwestern Atlantic Margin (Gradstein and Agterberg 1982). Other applications and evaluations of the method include Doeven et al. (1982) and Harper (1984). Implementation of the method was in the form of a mainframe computer program (Agterberg and Nel 1982a, b). Methodology and applications are detailed in

Quantitative Stratigraphy,**Fig. 3** Reasons why taxa might, or might not, be observed, leading to incomplete sampling

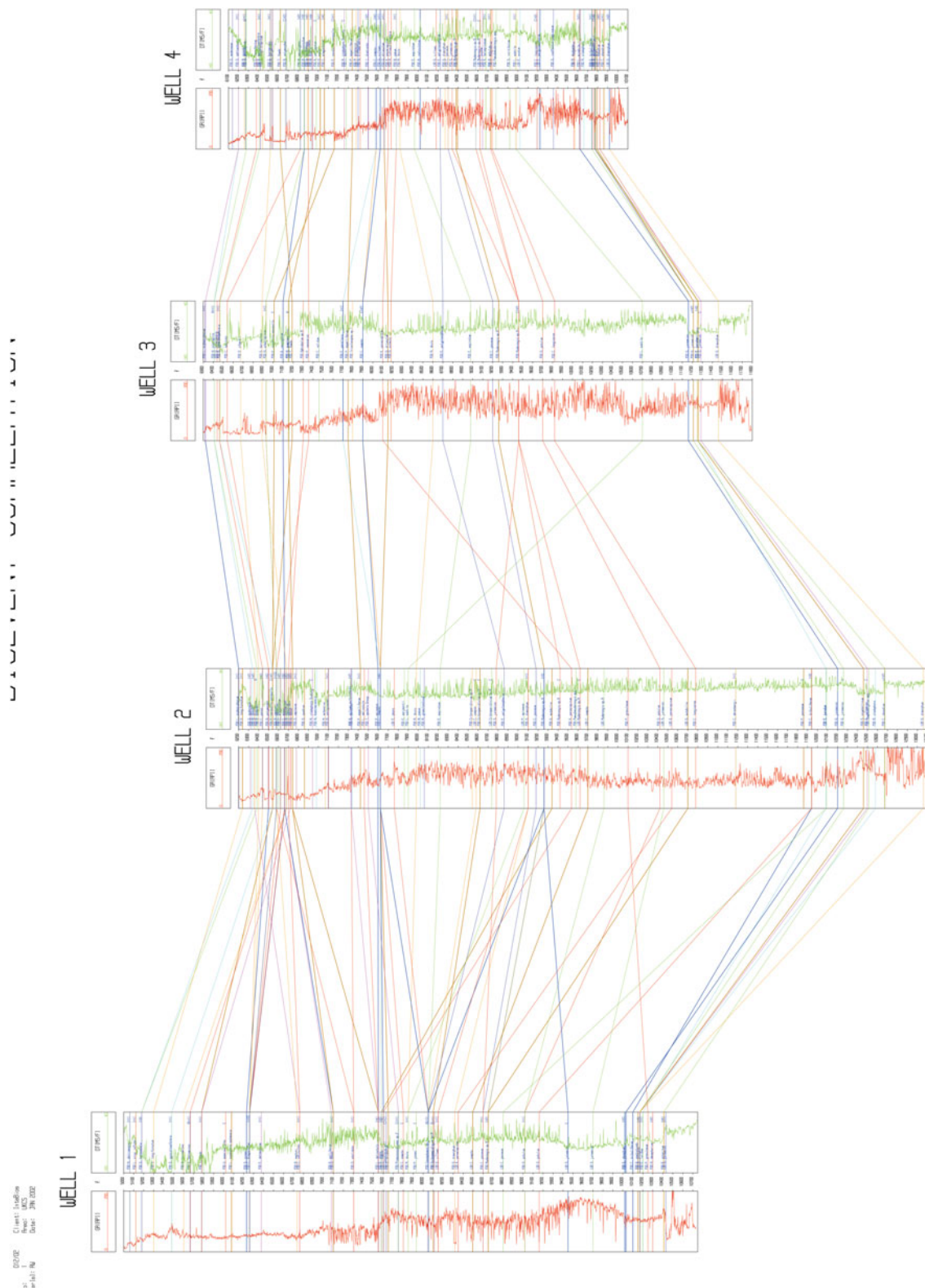
Gradstein et al. (1985). Development of RASC was carried out at the Geological Survey of Canada in Ottawa and Dartmouth, Nova Scotia, as part of Project 148 of the International Geological Correlation Program – Evaluation and Development of Quantitative Stratigraphical Correlation Methods, 1976–1985.

2. 1986–1994: Minor modifications of the method were introduced (D'Iorio and Agterberg 1989), and implementation was in the form of programs for personal computers. Large new applications are in Williamson (1987) and Gradstein et al. (1994).
3. 1995–1998: Modifications of RASC with addition of variance analysis were introduced in Agterberg and Gradstein (1997a, b, 1999), Agterberg et al. (1985), and Gradstein and Agterberg (1998). Novel applications included Gradstein and Bäckström (1996), Whang and Zhou Di (1998), and Corbett et al. (2014). Microcomputer application with attractive color graphics output was being facilitated by the development of a user-friendly Windows version of RASC. A 2-D Graphic Chart Control module in RASC allows editing of the final results for publication. This stage of computer development was sponsored by Saga Petroleum, Norway. RASC has seen large-scale professional applications by oil companies including Unocal, Shell, Saga Petroleum, Arco, Chevron, British Petroleum, Norsk Hydro, British Gas, and consulting companies like Millennia, UK.
4. 1999–2013: Emphasis in later versions of RASC and CASC was on graphic representation of the results (with original code made available on the website of the Editor, *Computers & Geosciences*). Additional features of RASC included the modeling frequency distributions of depth differences between successive events in the scaled optimum sequence as bilateral gamma distributions (Agterberg et al. 2007).

Ranked Optimum Sequence

The concept of an optimum sequence based on ranking was originally introduced by Hay (1972). All relative frequencies by which events are observed to occur stratigraphically above or below other events are considered. In the original Hay-type optimum sequence, the events are ranked in such a way that those above others in the optimum sequence have higher frequencies of occurring above these others, and lower frequencies of occurring below them. This ranking is based on superpositional relationships between events. In general, there are three possible types of superpositional relationships for a pair of events co-occurring in the same section. An event can be observed to occur above or below another event, or the two events coexist in the same sample. This third type of relationship, which implies that the two events are observed to be coeval, is ignored when an original Hay-type optimum sequence is constructed. In the ranked optimum sequence, which is based on a larger number of sections using nannofossils, two events can be coeval on the average when one of them occurs exactly as many times above the other one as it occurs below it. If an event is observed above another event in some sections but below it in others, a stratigraphic inconsistency involving these two events is indicated. The purpose of constructing the “modified Hay” method of ranking is to resolve such inconsistencies. Lines of correlation connecting observed positions of events in sections show crossovers when there are inconsistencies.

The so-called crossover frequency is defined as follows. If two events (A and B) occur in n sections with A above B in $n(AB)$ sections, A below B in $n(BA)$ sections, and A with B in the same sample in $n(A-B)$ sections, then the crossover frequency of A with respect to B is $f(AB) = \{n(AB) + 0.5n(A-B)\}/n$. Two events that are coeval on average have a crossover frequency equal to 0.5. When an event occurs



Quantitative Stratigraphy, Fig. 4 Crossover of dinoflagellate event correlations in four wells that sampled Upper Jurassic strata in the Troll field, offshore Norway (Gradstein, unpublished)

above another event in all sections, the crossover frequency is one, indicating absence of inconsistency in the observed superpositional relationships.

In most applications of ranking to large data sets, it is not possible to construct an optimum sequence according to the criteria outlined above. Pairwise inconsistencies, in which one event occurs above another in some sections but below it in other sections, are not the only possible type of inconsistency in superpositional relationships between events. Consider the following possibility: An event labeled A occurs more frequently above event B than the other way around, B occurs more frequently above C, and C occurs more frequently above A. This is an inconsistency involving three events. When a three-event cycle of this type occurs in the data set, it is not possible to construct the original Hay-type optimum sequence. However, by means of searching with the help of a computer, it is possible to identify those groups of three events which are involved in three-event cycles, and to study their superpositional relationships in detail. In RASC, such three-event cycles, and cycles involving more than three events (see next paragraph), are identified one after another. When a three-event cycle is found, the superpositional relationship of the two events that have the smallest difference between frequencies of occurring above and below one other is ignored.

Successive application of this rule to all cycles involving three or more events eventually results in an optimum sequence. In reality, the situation may be more complex if more than three events may be involved in a single cycle. For example, if event A occurs more frequently above B, B above C, C above D, and D above A, the result is a four-event cycle, provided that A and C as well as B and D are coeval on the average. If the latter condition is not satisfied, the four-event cycle is not real, because then the four events considered contain at least one three-event cycle which would be identified as such. Similar considerations apply to cycles involving more than four events. In general, the frequency of occurrence of four-event cycles is much less than that of three-event cycles, and cycles with more than four events are even rarer. This method of ranking is called the modified Hay method. Stratigraphically, the optimum sequence represents the average order in which events occur in a basin; a special graph in RASC shows the maximum extent of a last occurrence (LO event) or the very first occurrence of FO events, together representing the average range of a fossil in relative time. The latter often is more useful and more realistic for correlation than the extreme end points of a fossil in relative time (see below under Correlation).

The modified Hay method is not the only method that can be used for ranking stratigraphic events. Before the modified Hay method is utilized in RASC, the following ranking method called Presorting is applied first. Every event is compared with all other events. If an event occurs more frequently above another event, it receives an additional score of 1; if it

occurs more frequently below the other event, its score is not increased; and if it is coeval on the average with another event, both event scores are increased by 0.5. After adjusting the score with the ratio between linked events and the total number of events, the events can be ranked according to the magnitudes of their total scores in order to obtain an optimum sequence. Ranking methods based on scores have drawbacks because scores for many pairs may be missing (but corrected with above ratio), and events near the top or base of an optimum sequence automatically receive a full score. The advantage of presorting is that, with little computational effort, it results in a preliminary ranking which is likely to provide a good, preliminary solution. In RASC, the presorting result is refined by means of the modified Hay method.

Scaled Optimum Sequence

Biostratigraphic zonation not only determines stratigraphic order of events, but also provides a measure of separation in relative or linear time. The purpose of scaling in RASC is to compute interevent distances which are intervals between successive events in the optimum sequence. It is logical that two events that are coeval on the average will be assigned zero interevent distance along this scale. Scaling of the optimum sequence makes use of the stratigraphic reasoning that pairs of events in the optimum sequence with higher crossover frequency are closer together in relative time. Lines of correlation connecting observed positions of events in sections show crossovers when there are inconsistencies in the relative order of the events from section to section. Interevent distances are computed from z values which, in turn, are calculated from the crossover frequencies by the "probit" transformation $z_{ij} = \Phi^{-1}(f_{ij})$ where i and j are indexes that represent two stratigraphic events, and F represents the cumulative frequency of the normal Gaussian distribution in standard form. The probit transformation is widely applied in biometrics although not commonly for the purpose of scaling. It can be approximated by the linear transformation $z_{ij}^* = 2.93(f_{ij} - 0.5)$. For example, $f_{ij} = 0.5$ yields $z_{ij} = z_{ij}^* = 0$; $f_{ij} = 0.9$ gives $z_{ij} = 1.28$ and $z_{ij}^* = 1.17$.

Every estimated interevent distance is the weighted average of a single direct distance estimate and $(n^* - 1)$ indirect distance estimates. The direct distance estimate is simply the z_{ij} -value of the two events considered. Usually, there are many more indirect distance estimates which are based on differences between z -values. Each difference $(z_{ik} - z_{jk})$ involves a third event with index k that occurs in the vicinity of i and j in the optimum sequence. An advantage of using a single linear scale is that, on the average, any difference involving a third event is equal to the direct estimate. The expected value mathematical expectation of a direct estimate and all indirect

estimates of an interevent distance is the same. Discrepancies observed in practice between direct and indirect estimates can be ascribed to small-sample fluctuations and their effects reduced by averaging. The final estimated interevent distance is a function of the distance between each of the two events and all other events in the sequence which are in the vicinity of the two events. Weighting is applied during the computation of the average of the n^* individual estimates in order to account for the fact that crossover frequencies based on many superpositional relationships are more precise than those based on few; during this weighting, it also is considered that, taken separately, indirect estimates of distance are less precise than direct estimates. Standard deviations of the final interevent distances can be computed as well (see Agterberg 1990 for mathematical equations). The ranked optimum sequence is the starting point for all calculations outlined in the preceding paragraphs. Usually, several of the computed interevent distances turn out to be negative. However, by adding successive interevent distances, it is possible to calculate a cumulative RASC distance for every event. This cumulative distance is measured from the first event in the optimum sequence which can be defined as origin. Both origin and scale of the axis along which the events are plotted can be changed arbitrarily, because this scale is relative. Using cumulative distances, every event obtains a fixed position along the scale, and consequently, all interevent distances between successive events become positive, even if, initially, one or more negative values had been obtained. This procedure is called reordering. If there are no negative values, the order of the events remains unchanged. If negative interevent distances are computed, the preceding scaling calculations are repeated using the order of events in the initial scaled optimum sequence as the input sequence instead of the original ranking solution. The entire scaling process is repeated several times and results in elimination or significant reduction of the appearance of negative interevent distances during the calculations before final reordering is applied on the basis of the cumulative RASC distances.

A recent addition to RASC involves calculation of the scaled optimum sequence using thickness scaling of the events, as observed in each well or outcrop section (Agterberg et al. 2007). Depth differences between successive events in the optimum sequence satisfy frequency distributions as modeled for three large Cenozoic microfossil data sets consisting of 30 wells in the North Sea Basin, 27 wells on the Labrador Shelf and Grand Banks, and 11 wells in the western Barents Sea. The shapes of the three frequency distributions satisfy bilateral gamma distributions with similar parameters. These distributions are fitted by the construction of straight lines on normal Q–Q plots of square root-transformed average-corrected depth differences. The gamma distribution model is approximately satisfied except for small negative and positive depth differences, which have

anomalous frequencies because of the discrete sampling method used in exploratory well-drilling to collect microfossils. It implies not only comparable average stratigraphic order of events, but also comparable average sedimentation rates in the three Cenozoic basins selected for study.

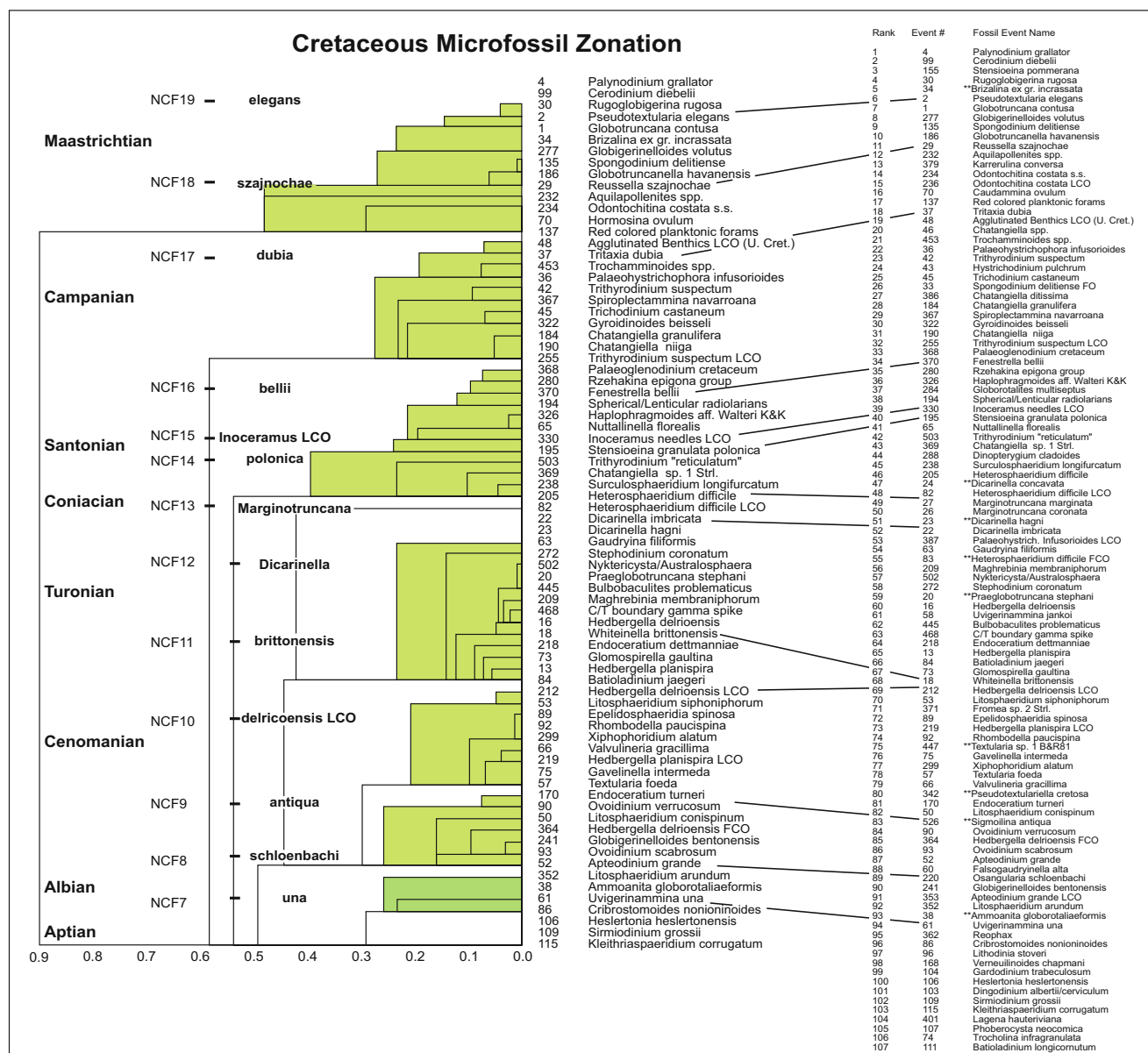
Dendrogram

Graphic representation of scaling the optimum sequence using calculated interfossil distances is in the format of a dendrogram (see Fig. 5 left below). In this graph, the interfossil/interevent distances are plotted along a horizontal scale. At the end of each line segment, a vertical line is drawn in the stratigraphically downward direction until it hits another line or the base line. Groups of events with short interfossil distances form clusters which can be assigned zonal names, being representative of interval zones. Although the unique events are shown in the dendrogram, interevent distances involving them are not plotted, because they are less precise than the interevent distances calculated by scaling. Although trivial, it might be mentioned that the scaled clusters are not the result of cluster analysis. Pairs of events in the optimum sequence with higher crossover frequency are closer together in relative time. Large interfossil distances likely represent break in the local sedimentary sequence.

Testing Stratigraphic Fidelity

An important task in RASC is to calculate how well the order of events in the individual wells or sections compares to that of the computed optimum sequence zonation, either ranked or scaled (Gradstein 1984).

The step model is the one of three methods in RASC to compare individual sections to the (scaled) optimum sequence which can be regarded as an average or “standard” based on all sections. The order of the stratigraphic events in a section that belong to those included in the optimum sequence is compared to their order in the optimum sequence in the following way. Each event is compared to all others. One penalty point is assigned when the order of two events in the section is not the same as in the order of these two events in the optimum sequence. Half a penalty point is assigned when two events are observed to be coeval. Consequently, an event that, in comparison with its neighbors, occurs much higher in a section than in the optimum sequence, e.g., due to reworking, will receive a relatively large number of penalty points, one for every event with which it is out of step in comparison to the optimum sequence. Similar considerations apply to an event that occurs stratigraphically too low in a section.



Quantitative Stratigraphy, Fig. 5 (right) Optimum sequence of Cretaceous events foraminifera, dinoflagellates, miscellaneous microfossils, and one physical log marker, offshore mid-Norway, using the Ranking and Scaling RASC method on 1753 records in 31 wells. The arrow of time is upward. The majority of events are last occurrences LO in relative time; LCO stands for Last Common or Last Consistent Occurrence. Each event occurs in at least 6 out of 31 wells, leaving 97 events. The display also shows the relative stratigraphic position of nine unique events. The optimum sequence is graphically tied to the scaled optimum sequence RASC zonation to the left, largely using the nominate zone markers.

(left) Scaling in relative time of the optimum sequence of Cretaceous events shown in Fig. 5 (right). The arrow of time is upward; the intervall distances are plotted on the relative scale to the left in dendrogram display format. Each event occurs in at least 7 of the 31 wells, leaving 72 events. Sixteen stratigraphically successive interval zones are recognized, middle Albian through late Maastrichtian in age. Large breaks at events 52, 84, 205, 255, and 137 indicate transitions between natural microfossil sequences; such breaks relate to hiatuses or facies changes, some of which are known from European sequence stratigraphy (Gradstein et al. 1999)

The penalty events scored for all events in a section can be added. The total number of penalty points of a section not only reflects how well the order of the events in the section resembles the optimum sequence but it also depends on the number of optimum sequence events occurring in the section.

This total can be normalized to make it independent of total numbers of events in section and optimum sequence.

Two more stratigraphic fidelity tests in RASC are (1) the bivariate scattergram display of (scaled) optimum sequence versus well with a notation if events are outside the expected

cluster and (2) normality testing. Details are in Agterberg and Gradstein (1999).

Cretaceous Biozonation, Offshore Mid-Norway

In this section, we outline a Cretaceous biozonation in the Norwegian Sea, offshore mid-Norway, using foraminifera, dinoflagellates, and some miscellaneous microfossil taxa. Rather than trying to create a zonation that maximizes stratigraphic resolution, and often is difficult to apply over a broader region, we prefer to outline zonal units that are easy to correlate regionally. Hence, use was made of the RASC method to erect a zonation for the Cretaceous, which also has helped to integrate dinoflagellate events in the shelly microfossil biozonation.

Initially, the data set for RASC comprised the multiple microfossil event record in 37 wells, mostly LO or LCO events of 550 foraminifers, some siliceous microfossils, and dinoflagellates, for a total of 1873 records. After several intermediate zonations, using RASC and its normality-testing function of the event record, five erratic events and seven wells with very low sampling quality (partly due to turbo-drilling) were deleted from the data. For example, the following fossil event records were deleted, as being far too high, or too low stratigraphically in wells, as determined from RASC stepmodel penalty points: *Globorotalites multiseptus* in 34/7-24s, *Endoceratium dettmaniae* in 35/3-4, *Gaudryina filiformis* in 6507/6-2, *Gyroidinoides beisseli* in 34/7-22, and *Hedbergella* sp. in 6607/5-1; several LO events for dinoflagellate cyst taxa were changed into LCO events. In addition, *Rhaphidodinium fucatum*, *Heterohelix globulosa*, *Caudammina ovuloides*, *Dorothia trochoides*, *Hedbergella* sp., *Dorocysta litotes*, *Xenascus ceratoides*, *Florentinia mantelli*, *F. deanei*, and *F. ferox* showed highly erratic occurrences through the wells, leading to deletion from the run data. From our own observations, it is clear that *Hedbergella*

sp. consistently occurs in the nominate upper Cenomanian-lower Turonian zone, but in the original well record that is not obvious. In most cases, reasons for the erratic stratigraphic positions are not clear.

Despite filtering out (removal) of “bad” data from the RASC runs, 21 cycles needed to be broken prior to zonation, where groups of 3 or more events had an unresolvable stratigraphic order. In our experience with datasets from a variety of regions over the world, with different microfossils, 21 (event cycles) is a low number. Inclusion of the 10 events that were deleted from the data set would have increased the number of cycles from 24 to 43, and the number of events with 6 or more stepmodel penalty points from 14 to 36; out of place (AAA events) events in the scattergrams also were higher, whereas the number of out of place events using the RASC normality test would have declined from 78 to 64, not a significant change, but possibly the result of a “more complete,” although not more precise, optimum sequence.

As mentioned, the few wells omitted from the RASC run suffer from poor data coverage. The remaining 31 wells harbor 519 events, with 1755 records. Table 1 provides a summary of data properties and RASC results. The number of events in the optimum sequence, with SD < ave SD, deals with the number of events that have a standard deviation below the average for the optimum sequence; the fact that 44 out of 72 optimum sequence events have a lower than average standard deviation is good. Stratigraphic coverage is relatively good, even when, as usual, only 98 out of 519 events occur in 6 or more wells, and only 72 events occur in 7 or more wells (Fig. 2). Cretaceous dinoflagellate cysts in particular show large taxonomic diversity, but limited traceability for many taxa.

The Cretaceous optimum sequence (Fig. 5, right) includes 98 events (“tops”), spanning Hauterivian through Maastrichtian strata. In addition, 9 events noted with ** occur in fewer than 7 wells but are inserted for the purpose of zonal definition and chronostratigraphic calibration. The

Quantitative Stratigraphy, Table 1 Summary of data properties and RASC results

NUMBER OF NAMES (TAXA) IN THE DICTIONARY	589
NUMBER OF WELLS	31
NUMBER OF DICTIONARY TAXA IN THE WELLS	519
NUMBER OF EVENT RECORDS IN THE WELLS	1755
NUMBER OF CYCLES PRIOR TO RANKING	21
NUMBER OF EVENTS IN THE OPTIMUM SEQUENCE	72
NUMBER OF EVENTS IN OPTIMUM SEQUENCE WITH SD < ave SD	44
NUMBER OF EVENTS IN THE FINAL SCALED OPTIMUM SEQUENCE (INCLUDING UNIQUE EVENTS SHOWN WITH **)	81
NUMBER OF STEPMODEL EVENTS WITH MORE THAN SIX PENALTY POINTS AFTER SCALING	14
NUMBER OF NORMALITY TEST EVENTS SHOWN WITH * OR **	78
NUMBER OF AAAA EVENTS IN SCALING SCATTERGRAMS	20

events listed are average last occurrences (average “top”), unless otherwise indicated; LCO stands for last common or last consistent occurrence, and FCO means first common or first consistent occurrence (in an uphole sense).

The so-called “Range” of taxa in the optimum sequence, which is a measure of the number of sequence positions an event spans (= rank positions in Fig. 5, right), is always low, except for *Fromea* sp. 2 (5 rank positions), *Dinopteridium cladoideus* (6 rank positions), *Trithyrodinium reticulatum* (5 rank positions), *Gyroidina beisseli* (5 rank positions), *Chatangiella niiga* (5 rank positions), and *Chatangiella* sp. (10 rank positions). The last occurrence record for these taxa is not reliable for detailed correlations in the wells examined. The optimum sequence serves as a guide to the stratigraphic order in which events in wells are expected to occur; hence, it is both predictive and serves in high-resolution correlation, as outlined below.

Figure 5, left is the scaled optimum sequence for the Cretaceous in the 31 wells, with interfossil distance from crossover pairs of events; this preferred scaled optimum sequence contains 72 events. Over a dozen stratigraphically successive Norwegian Sea Foraminifera (NCF) interval zones stand out, middle Albian through Maastrichtian in age. Resolution is poor in the lowermost and uppermost parts of the scaled sequence and reflects lack of superpositional data, and condensation of “tops” in the uppermost and lowermost Cretaceous part of wells.

The zones are named after prominent foraminiferal taxa that are easily recognized in the wells and are relatively widespread. The zones are informal, and their description is bound to change with addition of more data, and further assessment of taxonomy. Some of the foraminiferal index taxa (*Trocholina*, *Blefusciwana*, *Globigerinelloides gyroidinaeformis*, *Hedbergella brittonensis*, *Marginotruncana coronata*, and *Dicarinella concavata*) are rare due to the high-latitude setting of the study region, or poor preservation of carbonate tests in the siliceous mudstones. As mentioned earlier, integration, using the RASC method, with dinoflagellate events, improves the applicability of both types of microfossils for regional biostratigraphy.

Chronostratigraphic assignments are tentative, hampered among others by the absence, or scarcity or keeled planktonic foraminifers, like *Rotalipora*, and the majority of *Dicarinella*, *Marginotruncana*, and *Globotruncana*. Zones of short stratigraphic duration, like the *Sigmoilina antiqua* Zone, assigned to lower Cenomanian may be difficult to observe in ditch cuttings, particularly when the lithology is sandy. There is ample room for improving the resolution of the lower and upper parts of the zonation, for example, in the condensed Hauterivian-Barremian, and hiatus riddled, Aptian-middle Albian sections, and in the thick, carbonate devoid strata of the inner Voering Basin, assigned to Campanian (Gradstein

et al. 1999). The latter represents a poorly sampled section, probably laid down on a deep marine, middle bathyal basin floor.

The scaled optimum sequence shows major breaks between several successive zones, including between the *una* and *schloenbachi* zones, *delrioensis* LCO and *brittonensis* zones, *Marginotruncana* and *polonica* zones, *bellii* and *dubia* zones, and particularly between *dubia* and *szajnochae* zones. The question, of what the meaning is of these large breaks, is best answered when we consider how RASC scaling operates. Scaling calculates the relative stratigraphic distance between successive events in the RASC optimum sequence. This relative distance is calculated from the frequency of crossover of all possible pairs of events in the optimum sequence over all wells. Hence, large distances between successive events, corresponding to the larger RASC zonal breaks indicated above, signify strongly diminished crossover between events that occur below and above the breaks. Reasons for these breaks thus must be thought in stratigraphic gaps, and paleoenvironmental changes that likely are connected to sequence stratigraphic changes.

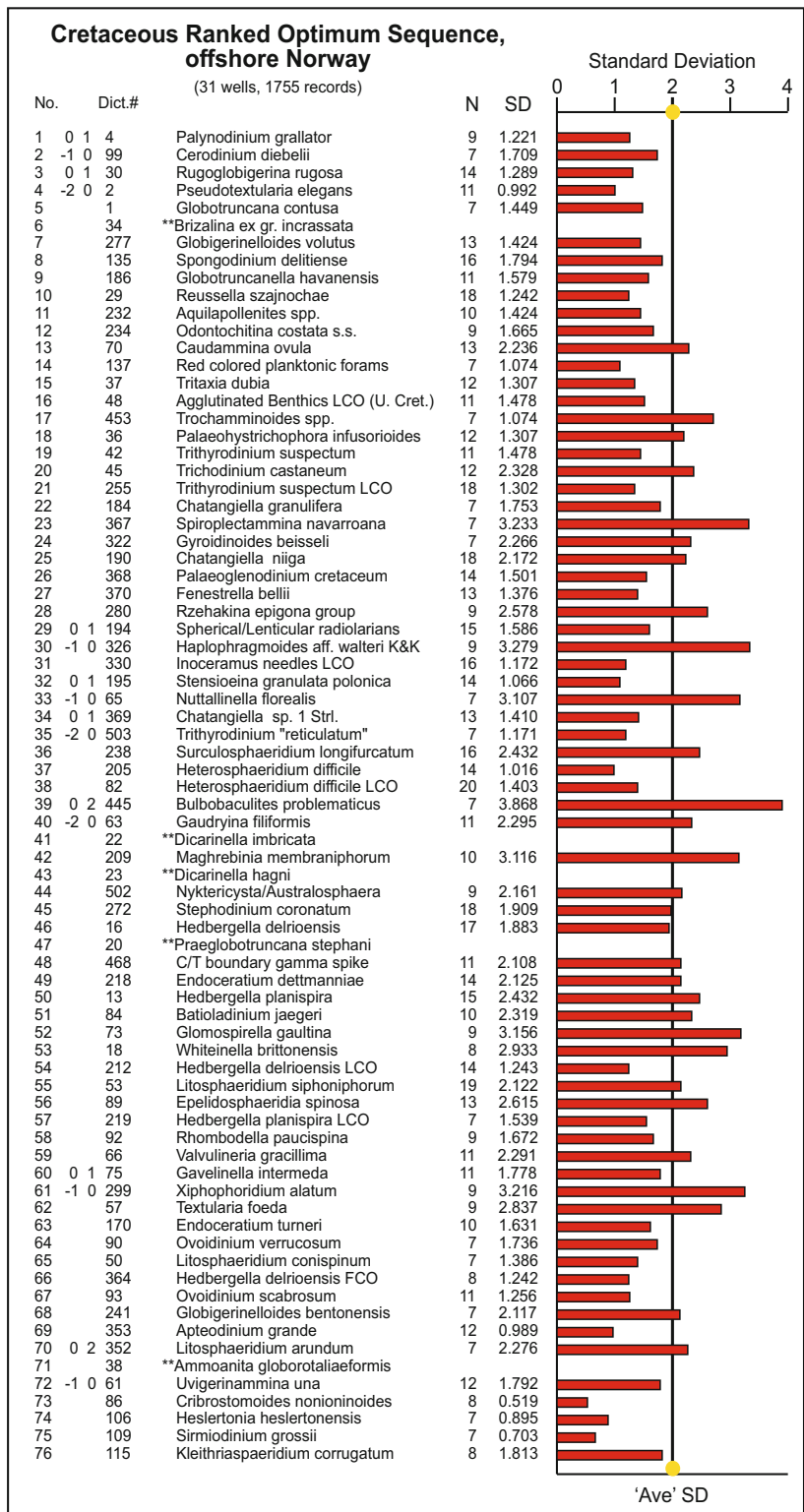
The reasons for this may be summarized as follows:

1. The *una-schloenbachi* break reflects a latest Albian lithofacies change and hiatus, connected to the Octeville hiatus in NW Europe.
2. The *delrioensis* LCO-*brittonensis* break reflects the mid-Cenomanian lithofacies change and hiatus, connected to the mid-Cenomanian nonsequence, and Rouen hardground.
3. The *Marginotruncana-polonica* break may be the turnaround in the tectono-eustatic phase of middle Coniacian, near the end of the Lysing sand deposition. The break occurs above the level of *Heterosphaeridium difficile* LCO, which represents a maximum flooding surface.
4. The *bellii-dubia* break occurs above the change from marly sediments to siliciclastic sediments that takes place in lowermost Campanian. The break, as the previous one, is near a maximum flooding event, i.e., the LCO of *T. suspectum* in the early middle Campanian.
5. The *dubia-szajnochae* break reflects the abrupt change from siliciclastic to marly sediment at the Campanian-Maastrichtian boundary, noted in the southern part of the region.

Within the major units, smaller zonal subdivisions may be recognized, using a combination of subjective judgment from individual well study, and objective subdivision from ranking and scaling analysis. Hence, the final zonation takes advantage of both subjective and more objective analysis and integrates the two into one framework.

Quantitative Stratigraphy,

Fig. 6 Summary of event variance analysis results for the Cretaceous optimum sequence of Fig. 5. To the left, the optimum sequence of Cretaceous microfossil events in 31 wells, offshore mid-Norway, where each event occurs in at least 7 out of 31 wells; N is the number of wells sampled to calculate the S.D. per event N, and S.D. is the standardized deviation from the line of correlation in each well. The average standard deviation of 1.8903 is the sum of all S.D.'s, divided by the total number of events in the optimum sequence. The asterisk behind events indicates an event with an S.D. that is smaller than the average S.D. From event S.D. theory, it follows that events with a below average S.D. correlate the same relative stratigraphic level more faithfully from well to well than events with a higher S.D.

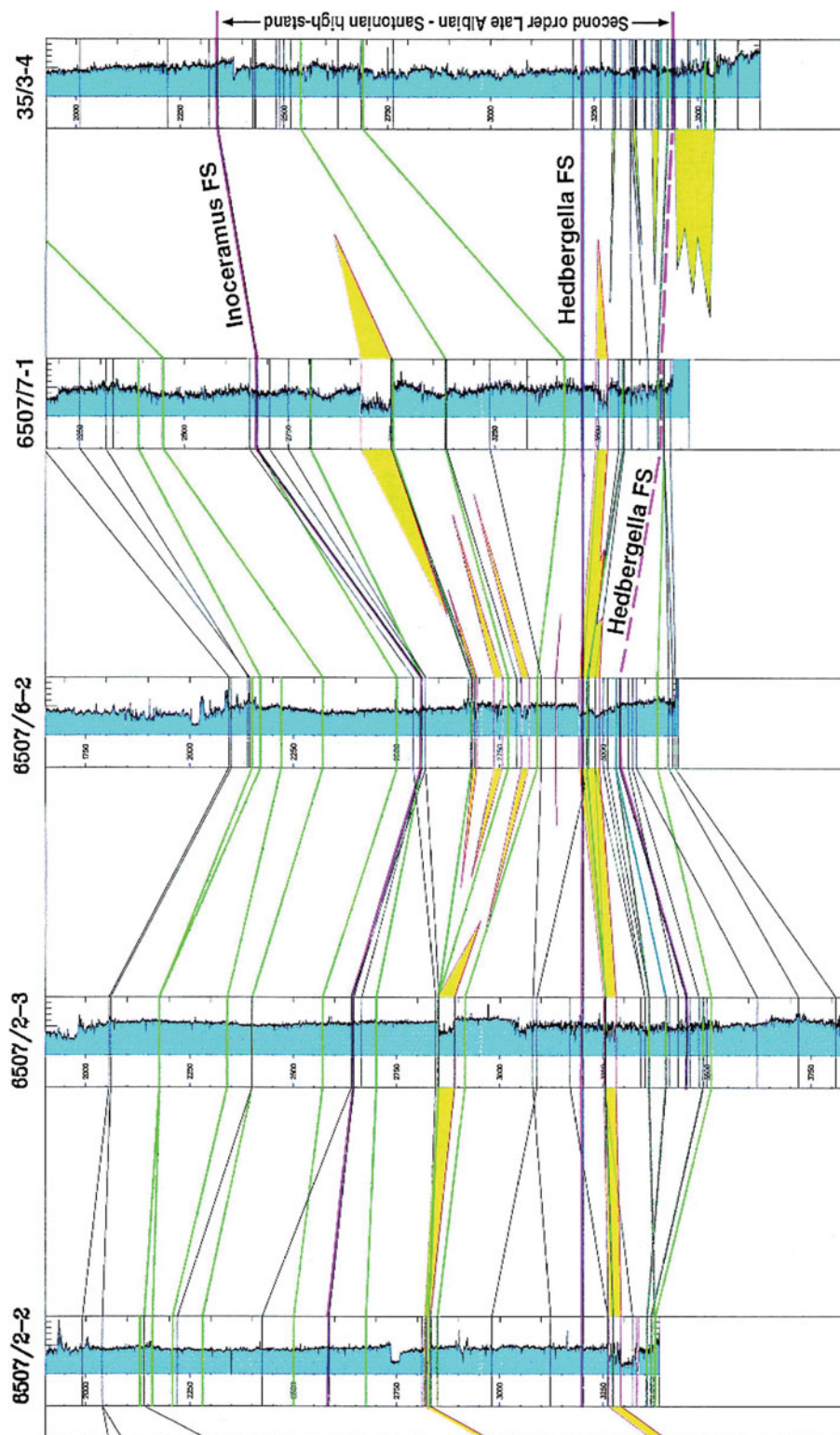
**Correlation**

The complete RASC and CASC module executes automated stratigraphic correlation of wells or sections, using the (scaled) optimum sequence as zonal standard. However,

geologic correlation generally is a deterministic, and not a probabilistic process, using discrete tools like lithologic and/or log events and seismic markers with finite position in metric or travel time units.

Quantitative Stratigraphy,

Fig. 7 Correlation of events in the optimum sequence and zonation of Figs. 5 and 6 in Albian through Santonian strata in five wells, offshore mid-Norway. Green and red colored correlation lines are from events that have below average stratigraphic standard deviation, and are best track the same stratigraphic levels. Black lines are from events that have above average stratigraphic standard deviation and hence are less certain in correlation, and show some crossovers. Gravity flow sands are in yellow. The *Hedbergella delrioensis* FCO and LCO events (lower Cenomanian) and the *Inoceramus* LCO event (Santonian) are colored red and are considered flooding surfaces (FS). These events represent reliable regional correlation levels, with below average S.D. that reflect major marine transgressions, together with slow sedimentation. This graph clarified the localized and gravity flow nature of the sands, which could not be easily correlated using seismic and physical well logs stratigraphy



Automated correlation would not correct for events ending up on the wrong side of unconformities; it does not emphasize breaks between scaling clusters; it does not prevent events

from inadvertently ending up in gravity flow intervals; it does not provide weight to events with low variance; and it does not emphasize index fossil correlations, nor does it project

CASC correlations lines in an easy way on seismic or log correlation diagrams.

Having said so, there is an easy manner to operate reliable stratigraphic correlation through use of events with lower and higher stratigraphic fidelity from the optimum sequence with variance, like in Fig. 6 that shows correlation of events in Albian through Santonian strata in five wells, offshore mid-Norway (Fig. 7). Green and red colored correlation lines are from events that have below average stratigraphic standard deviation, and are best track the same stratigraphic levels. Black lines are from events that have above average stratigraphic standard deviation and hence are less certain in correlation, and show some crossovers. Gravity flow sands are in yellow. The *Hedbergella delrioensis* FCO and LCO events (lower Cenomanian) and the *Inoceramus* LCO event (Santonian) are colored red and are considered flooding surfaces (FS). These events represent reliable regional correlation levels, with below average S.D. that reflect major marine transgressions, together with slow sedimentation. This graph clarifies the local and gravity flow nature of the sands, which could not be easily correlated using seismic stratigraphy.

Conclusion

Although operation and execution of quantitative methods seemingly might appear to have a steep threshold and be difficult, the use of QSCreator for data input, rapid execution, and attractive graphic results, in reality, makes Ranking and Scaling (RASC) easy to use. Advantages may be summarized as follows:

1. Standardization of the fossil record and stratigraphic method gives easy access to data and results.
2. Data sets and results are easy to communicate and rapidly updated with new information.
3. Integration of all fossil and also physical (e.g., isotope, well-log) events in one stratigraphic solution increases resolution and practical use.
4. Methods and results (zonation + correlation) are more objective than handmade solutions.
5. Analysis of uncertainty in the event record and in the quantitative zonation greatly improves insight in true stratigraphic resolution and reliability of event correlation.
6. Interpolation of most likely (including missing) event positions in sections increases detail in correlation between sections.
7. Unlike subjective stratigraphy, the quantitative methods generally provide more than one possible solution, depending on run conditions, and provide biostratigraphy with multiple working hypotheses.

8. Quantitative stratigraphy methods can handle large and complex data sets, and calculate stratigraphic solutions fast and easy.

References

- Agterberg FP (1990) Automated stratigraphic correlation. Elsevier Publ. Co., Amsterdam, 424 p
- Agterberg FP, Gradstein FM (1999) The RASC method for ranking and scaling of biostratigraphic events. In: Proceedings conference 75th birthday C.W. Drooger, Utrecht, November 1997. Earth Sci Rev 46(1–4):1–25
- Agterberg FP, Nel LD (1982a) Algorithms for the ranking of stratigraphic events. Comput Geosci 8(1):69–90
- Agterberg FP, Nel LD (1982b) Algorithms for the scaling of stratigraphic events. Comput Geosci 8(2):163–189
- Agterberg FP, Oliver J, Lew SN, Gradstein FM, Williamson MA (1985) CASC Fortran IV interactive computer program for correlation and scaling in time of biostratigraphic events. Geological Survey of Canada, Open file report 1179
- Agterberg F, Gradstein F, Liu G (2007) Frequency distribution of thickness of sediments bounded by Cenozoic biostratigraphic events in wells drilled offshore Norway and along the Northwestern Atlantic Margin. Earth Sci Resour 16(3):220–233
- Corbett MJ, Watkins DK, Pospichal JJ (2014) A quantitative analysis of calcareous nannofossil bioevents of the Late Cretaceous (Late Cenomanian-Coniacian) Western Interior Seaway and their reliability in established zonation schemes. Mar Micropaleontol 109:30–45
- D'Iorio MA, Agterberg FP (1989) Marker event identification technique and correlation of Cenozoic biozones on the Labrador Shelf and Grand banks. Bull Can Petrol Geol 37:346–357
- Doeven PH, Gradstein FM, Jackson A, Agterberg FP, Nel LD (1982) A quantitative nannofossil range chart. Micropaleontology 28:85–92
- Gradstein FM (1984) On stratigraphic normality. Comput Geosci 19:43–57
- Gradstein FM, Agterberg FP (1982) Models of Cenozoic foraminiferal stratigraphy, northwestern Atlantic Margin. In: Cubitt JM, Reymont RA (eds) Quantitative stratigraphic correlation. Wiley, Chichester, pp 119–173
- Gradstein FM, Agterberg FP (1998) Uncertainty in stratigraphic correlation. In: Sequence stratigraphy, concepts and applications. Norwegian Petroleum Society, Elsevier, Amsterdam, pp 9–29
- Gradstein FM, Agterberg FP (1999) RASC and CASC – Tools for the biostratigraphic workstation- users guide. Authors Edition, Version 17
- Gradstein FM, Bäckström SA (1996) Cainozoic biostratigraphy and palaeobathymetry, northern North Sea and Haltenbanken. Norsk Geologisk Tidsskrift 76:3–32
- Gradstein FM, Agterberg FP, Brower JC, Schwarzacher W (1985) Quantitative stratigraphy. Reidel Publ. Co. & UNESCO, 598 p
- Gradstein FM, Kaminski MA, Berggren WA, Kristiansen IL, D'Iorio M (1994) Cainozoic biostratigraphy of the North Sea and Labrador Shelf. Micropaleontology 40(supplement 1994):152 pp
- Gradstein FM, Kaminski MA, Agterberg FP (1999) Biostratigraphy and paleoceanography of the Cretaceous Seaway between Norway and Greenland. In: Proceedings conference 75th birthday C.W. Drooger, Utrecht, November 1997. Earth Sci Rev 46(1–4):27–98
- Harper CW Jr (1984) A Fortran IV program for comparing ranking algorithms in quantitative biostratigraphy. Comput Geosci 10:3–29
- Hay WW (1972) Probabilistic stratigraphy. Eclogae Geol Helv 65(2):255–266

- Whang P, Zhou Di (1998) The application of RASC/CASC methods to quantitative biostratigraphic correlation of Neogene in northern South China Sea. In: Proceedings 30th international geological congress, Beijing
- Williamson MA (1987) A quantitative foraminiferal biozonation of the Late Jurassic and Early Cretaceous of the East Newfoundland Basin. *Micropaleontology* 33(1):37–65

Quantitative Target Selection

Brandon Wilson¹, Emmanuel John M. Carranza² and Jeff B. Boisvert¹

¹University of Alberta, Edmonton, AB, Canada

²University of KwaZulu-Natal, Durban, South Africa

Definition

Determining an area of interest for mineral exploration involves risk and often has a low probability of success because economic ore deposits are rare. The main goal of early prospecting is mineral target selection – the identification of areas where targeted exploration and drilling should be pursued. Targets are defined as regions with favorable geology, anomalous geochemical concentrations, or other deposit-specific intrinsic characteristics that increase the probability of discovering a recoverable mineral resource.

Quantitative target selection is a data-driven process that identifies targets utilizing mathematical models fit to data from known mineral deposits using measured geological, geochemical, geophysical, and remotely sensed anomalies. Typical quantitative outputs include indicators of prospectivity or relative probability of mineralization. Conversely, qualitative target selection methods focus on using the understanding of ore formation processes to identify locations where the necessary conditions for mineralization exist.

Quantitative target selection models rely on the detection of multivariate signatures unique to specific mineral deposits of interest. These signatures can be directly fit to measured data in known mineralization locations using a variety of models. Alternatively, cutoff values can be applied to spatial models to represent the presence of ore and the probability of exceeding the cutoff value. Both methods can be used to generate prospectivity maps, which often display the relative probability of finding economic ore at a given location (Fig. 1).

The efficacy of quantitative target selection relies on analogous mineral deposits and available exploration data, which can be lacking depending on the area of interest. A combination of qualitative and quantitative target selection is preferred in practice, with qualitative methods used to limit the spatial extents in which quantitative models are

considered. In effect, quantitative methods are used to narrow down the results of a qualitative approach.

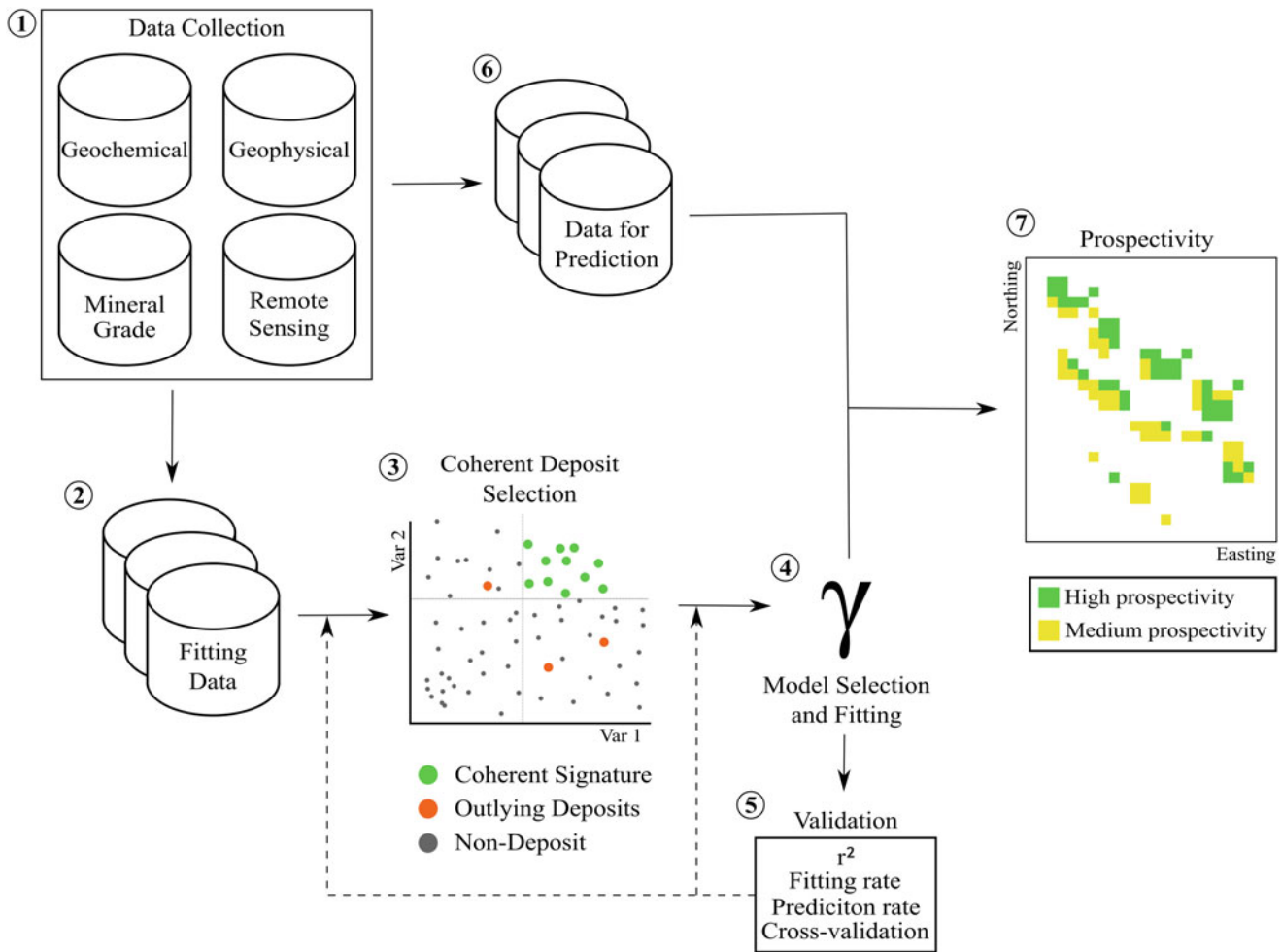
Motivation

Choosing exploration targets is often overlooked as a part of the mining cycle. Prospecting has become more difficult as a large proportion of outcropping deposits have been identified and exploited (Marjoribanks 2010). Additional sampling and predictive models are required to find new mineral deposits. Techniques for target selection vary between empirical and conceptual, while most practical implementations are a combination of both (Marjoribanks 2010). Conceptual methods focus on utilizing ore generation models and local geology to generate a relatively small number of high-probability ore targets.

In contrast, quantitative methods are used to associate measured variables with mineralized areas (Fig. 1). Measured geochemical, geophysical, mapping, and other data are related to known instances of mineralization using a variety of models. Data can be used to directly predict mineralization by assigning a binary variable to locations with known mineralization, followed by logistic regression or machine learning to predict ore likelihood in areas of interest (Carranza 2008; Xiong et al. 2018; Zuo 2020). Alternatively, data can be modeled to determine where critical genetic factors exist (Pan 2018). Critical genetic factors are geological processes that are necessary to generate concentrated ore. These factors are predicted through critical recognition criteria: combinations of variables meeting cutoffs that indicate the factor may be present (Pan 2018). Quantitative modeling predicts the probability that these criteria are met at a location of interest.

Before considering the construction of a mine, a selected target proceeds through a series of evaluation stages, including drilling, delineation, prefeasibility, and feasibility studies. Targets are eliminated from consideration at each evaluation stage, with only a small proportion proceeding because economic ore deposits are rare. Increasing the rate at which targets are converted to further exploration stages is the objective of target selection. Evaluating the success of target selection methods is difficult due to the random nature of mineralization and the sensitivity of evaluation metrics to the success or failure of single targets. A more robust but indirect measure of target selection success is the frequency of ore intersections found during early exploration drilling (Marjoribanks 2010); methods that identify locations where subsequent drilling more often pierces high grade mineralization are preferred, even if an ore deposit is not eventually found.

The goal of quantitative target selection is to generate an objective, data-driven measure of prospectivity to reduce subjectivity, increase reproducibility, and improve



Quantitative Target Selection, Fig. 1 Schematic of a typical quantitative target selection workflow and results. Concepts are labeled 1–7. 1) All available data is collected, transformed, interpolated, filtered, enhanced, etc. 2) Samples that are known to belong to deposit or non-deposit locations are split into training and testing sets and used in model development (steps 3, 4, 5). 3) Deposits are assessed for coherency, ensuring reliable detection and separation while reducing Type I errors.

4) A model type, such as logistic regression or random forest, is fit to coherent deposit data. 5) Model fit is assessed with training and testing data sets, facilitating model refinement if the fit is unacceptable (returning to steps 3, 4) or used as the final model if the fit is acceptable. 6) Data in locations with unknown mineralization are used in the fit model to predict prospectivity. 7) Outputs, such as prospectivity maps, are used to determine locations that are favorable for further exploration

consistency of target selection (Pan 2018); however, quantifying geological controls in a way that allows them to be effectively captured in a quantitative model is not trivial. Combining qualitative and quantitative methods is complementary and can improve model accuracy. Combined models include limiting quantitative model runs to areas determined to be appropriate by qualitative workflows.

Implementation

Many implementation decisions influence the quality of predictive models for target acquisition. These decisions include the size and location of the study area, defining positive and

negative examples of mineralization, and selection of data to use. For effective modeling, positive samples must share a multivariate data signature that is differentiable from background, non-mineralized locations. All deposits in a region that share such a signature are considered coherent (Fig. 1). An example that would confound most models follows: a set of known deposits are found to have elevated geochemical measurements with a measurable magnetic anomaly, but a subset of deposits does not present the anomaly – these sets of deposits are incoherent. Fitting a model to these samples would increase predicted false positives for deposits where the magnetic anomaly is necessary to indicate mineralization. To ensure deposit-type locations are coherent, Carranza et al. (2008) suggest logistic regression based on mineral

occurrence scores to classify deposits as coherent and non-coherent; a cutoff value splits the dataset, and only coherent deposit-type locations are used in modeling (Carranza et al. 2008).

Prediction scale is critical for quantitative target selection. Coarse grids limit the presence of outlying values necessary to detect a deposit due to the impacts of averaging data values. Conversely, fine grids are not reasonable where data is not correspondingly dense. It can be helpful to consider the probability of two deposit-type locations to be within a certain distance of each other; the maximum distance at which there is no probability of producing neighboring deposits is the maximum level of discretization that should be considered (Carranza 2008).

Numerous algorithms exist to fit and predict mineral occurrence scores or relative probabilities. Methods used as early as the 1980s include logistic regression, weights of evidence, and evidential belief functions. Since the 2000s, machine learning methods have increased in popularity as data availability and computing power improve, allowing for deeper models (Zuo 2020). Machine learning approaches that have been considered for quantitative target selection include neural networks, support vector machines, random forests, autoencoder networks, and convolutional neural networks (Zuo 2020).

Quantitative target selection workflows result in a variety of outputs. Typical deliverables include (1) mineral prospectivity maps (Fig. 1) showing relative potential to find deposits over a study area in the form of mineral prospectivity scores, and (2) relative target ranking, where specified target areas are prioritized by similar metrics (Pan 2018). Regardless of the specific type of output, the goal is to improve future exploration decisions.

Model validation is necessary to ensure that predictions are reliable and can be used to guide decision-making (Carranza 2008). Both model fit and predictive value must be considered and assessed with appropriate metrics. Fit metrics, such as r^2 , quantify how well the model represents the data it is modeling. Predictive metrics typically consider cross-validation techniques, such as k-fold validation, where data are left out and predicted with the proposed modeling workflow. The frequency of Type I and Type II errors and associated recall and precision values quantify: (1) how effectively ore locations are correctly identified as exploration targets, (2) how often they are missed, and (3) how often negative samples are mistakenly identified as targets. Validation can also take the form of splitting the data at locations with known levels of mineralization into training and testing sets and evaluating the efficacy of the model for predicting the test set (Carranza 2008). These metrics can be used to compare and choose between various models.

Data

Quantitative modeling is flexible in what types of data can be considered. Data can be continuous or discrete, exhaustively or sparsely sampled, measured or calculated, or even digitized qualitative data. Marjoribanks (2010) explores various types of geological data, including drilling, remote imagery, geochemical information, and geophysical variables. Many predictive models, such as logistic regression, require all data to be present at every evaluated location to produce a prediction. Many data sources, such as drillholes, are sparsely sampled and must be spatially interpolated (i.e., Kriging) or simulated (i.e., sequential Gaussian simulation) to be used as exhaustive inputs to quantitative target selection models. Transforming, filtering, and enhancing data are necessary preprocessing operations to develop a set of data appropriate for modeling and to maximize the value of information from data (Marjoribanks 2010; Pan 2018). Because geochemical anomalies are often critical to target identification, filters are applied to distinguish these extreme cases from background information. Care should be taken to ensure that issues such as smoothing and artifacts do not dilute data value (Pan 2018).

While there is an abundance of data types that can inform quantitative target selection, data availability is a significant limitation. Unexplored regions rarely have abundant, accessible, high quality data due to the expense of data collection and the remote nature of many prospective areas. However, many under-explored jurisdictions worldwide provide datasets to facilitate target selection, increase exploration investment, and encourage mining activities. Government agencies may publish regional target evaluations to incentivize exploration. For example, the Geological Survey of Canada provides mineral prospectivity maps on the Melville Peninsula in Nunavut (Harris et al. 2015). The quantitative data available in this region includes geochemical samples of lake sediment, airborne magnetic data, and gamma-ray spectrometer data. As the geochemical samples are discretely sampled, they are interpolated with Kriging to form a continuous data layer. A random-forest model is fit to predict the potential for gold mineralization based on 51 known occurrences, outperforming a knowledge-driven weighted index overlay model (Harris et al. 2015). The overall accuracy of the random-forest model is 85% based on 17 randomly selected gold occurrences. The data types used in this case study are commonly available for other regions in northern Canada.

The primary limitations to quantitative target selection are (1) a lack of available data, and (2) a lack of relevant known positive and negative examples of deposit-like locations required for model fitting. The same data types must exist in areas that are being evaluated, which often do not yet have

exploration-level drilling, leading to the prioritization of inexpensive remotely obtainable data. This can be particularly limiting for modeling methodologies that do not handle missing data during fitting, such as logistic regression and many machine learning algorithms (Carranza et al. 2008; Zuo 2020). Locations with missing data can be removed from consideration to fit the model. Alternatively, missing data can be simulated with imputation. Some machine learning methods, such as random forests, can be adapted to missing data to improve model flexibility (Zuo 2020).

Summary

Quantitative target selection is critical to managing risk associated with mineral exploration. Data is lacking during the early stages of exploration; therefore, reducing the number of targets considered while prioritizing high quality exploration targets optimizes capital expenditures and increases the conversion of targets to mineral reserves. The techniques employed to leverage existing data include mineral prospectivity analysis and machine learning approaches. Quantitative models alone may not be sufficient and combining them with qualitative methods improves results. Overall, quantitative target selection is a reasoned approach to increasing the conversion of exploration targets by leveraging all available data to decrease investment risk. The process has value for governments seeking investment and for companies seeking new targets.

Cross-References

- Machine Learning
- Mineral Prospectivity Analysis
- Remote Sensing

Bibliography

- Carranza EJM (2008) Geochemical anomaly and mineral prospectivity mapping in GIS. In: Handbook of exploration and environmental geochemistry. Elsevier, p 366. <https://app.knovel.com/hotlink/toc/id:kpHEEGVGA4/handbook-exploration-2/handbook-exploration-2>
- Carranza EJM, Hale M, Faassen C (2008) Selection of coherent deposit-type locations and their application in data-driven mineral prospectivity mapping. *Ore Geol Rev* 33(3):536–558. <https://doi.org/10.1016/j.oregeorev.2007.07.001>
- Harris JR, Grunsky E, Behnia P, Corrigan D (2015) Data- and knowledge-driven mineral prospectivity maps for Canada's North. *Ore Geol Rev* 71:788–803. <https://doi.org/10.1016/j.oregeorev.2015.01.004>
- Marjoribanks R (2010) Geological methods in mineral exploration and mining. Springer, Berlin/Heidelberg. <https://doi.org/10.1007/978-3-540-74375-0>

- Pan G (2018) General framework of quantitative target selections. In: Sagar BSD, Cheng Q, Agterberg F (eds) Handbook of mathematical geosciences. Springer International Publishing, pp 411–435. https://doi.org/10.1007/978-3-319-78999-6_21
- Xiong Y, Zuo R, Carranza EJM (2018) Mapping mineral prospectivity through big data analytics and a deep learning algorithm. *Ore Geol Rev* 102:811–817. <https://doi.org/10.1016/j.oregeorev.2018.10.006>
- Zuo R (2020) Geodata science-based mineral prospectivity mapping: a review. *Nat Resour Res* 29(6):3415–3424. <https://doi.org/10.1007/s11053-020-09700-9>

Quaternions and Rotations

Helmut Schaeben

Technische Universität Bergakademie Freiberg, Freiberg, Germany

Definition

The skew field of real quaternions provides an extension of the system of numbers beyond complex numbers. Quaternions form a four-dimensional associative algebra $\mathbb{H}$ (associated with the name of Hamilton) over the field of real numbers $\mathbb{R}$. The components of a unit real quaternion can be related to the angle-axis parametrization of rotations. Expressing rotations in this way and applying quaternion multiplication simplify the concatenation of rotations and yield Rodrigues' formula providing the angle and axis of the resulting rotation in a straightforward manner.

Historical Note

At a time, mathematicians were pondering on two apparently unrelated problems. One problem pertained to the hierarchical number system of natural, integer, rational, real, and complex numbers $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$, and the open question whether a reasonable extension exists. Another problem concerned the concatenation, i.e., the sequential composition, of two rotations in three-dimensional Euclidean space $\mathbb{R}^3$ both given in terms of their angles and axes of rotation, i.e., $R(\omega, \mathbf{n}) = R_2(\omega_2, \mathbf{n}_2)R_1(\omega_1, \mathbf{n}_1)$ (to be read from right to left), and the question what the angle ω and the axis $\mathbf{n} \in \mathbb{S}^2$ of their concatenation are. The answer to both problems is provided by quaternions, more specifically the skew field of real quaternions $\mathbb{H}$. They are generally credited to Sir William Rowan Hamilton, who introduced the quaternion algebra with $i^2 = j^2 = k^2 = ijk = -1$ (Hamilton 1844) as carved in the stone of Broome Bridge of the Royal Canal. However, the resolution of the problem of concatenated rotations had been found earlier by Rodrigues (1840), who did not receive any

credit for his work on the special orthogonal group $SO(3)$ of rotations until recently (Altmann 1989). Gauss knew about quaternions since 1819, but his results remained unpublished until 1900 (Gauss 1900; Pujol 2012).

Yet another problem with rotations in $\mathbb{R}^3$ is the characterization of all rotations mapping a given unit vector $\mathbf{u} \in \mathbb{S}^2 \subset \mathbb{R}^3$ onto another given unit vector $\mathbf{v} \in \mathbb{S}^2$, i.e., $G(\mathbf{u}, \mathbf{v}) = \{R \in SO(3) | R\mathbf{u} = \mathbf{v}\}$. In terms of quaternions, this set of rotations is a great circle of the quaternion unit sphere $\mathbb{S}^3 \subset \mathbb{H}$ (Meister and Schaeben 2004), which in turn is a geodesic of $SO(3)$ (Novelia and O'Reilly 2015).

Rotations in $\mathbb{R}^2$ or $\mathbb{R}^3$

In Euclidean geometry, a rotation is a linear transformation that preserves distances. Thus a rotation is an isometry specified by its properties that it leaves at least one point fixed and that it preserves handedness. A rotation in two-dimensional Euclidean $\mathbb{R}^2$ is an isometry characterized by one fixed point and an invariant plane of rotation containing the fixed point; a rotation in three-dimensional Euclidean $\mathbb{R}^3$ is an isometry characterized by an axis of rotation, i.e., a line of its fixed points, and an invariant plane of rotation orthogonal to that axis. Then, rotations in $\mathbb{R}^3$ are determined by their angle $\omega \in (-\pi, \pi]$ and their axis of rotation $\mathbf{n} \in \mathbb{S}^2 \subset \mathbb{R}^3$. Rotations in $\mathbb{R}^2$ are determined by an angle of rotation $\omega \in (-\pi, \pi]$ referring to a two-dimensional plane spanned by orthonormal vectors $\mathbf{x}$ and $\mathbf{y}$ and a fixed point, usually the origin of the coordinate system provided by the unit vectors $\mathbf{x}$ and $\mathbf{y}$; thus, it may be thought of as rotation by $\omega \in (-\pi, \pi]$ about the axis $\mathbf{z}$ through the origin and orthogonal to vectors $\mathbf{x}$ and $\mathbf{y}$ spanning the invariant plane.

Rotations in $\mathbb{R}^n$, $n = 2, 3$, form the special orthogonal groups $SO(n)$, $n = 2, 3$, respectively, as they satisfy the following:

- Closure: For all $R_1, R_2 \in SO(n)$, the result of their concatenation, i.e., sequential composition R of first R_1 followed by R_2 , is an element of $SO(n)$, $R_2 R_1 = R \in SO(n)$.
- Associativity: For all $R_1, R_2, R_3 \in SO(n)$ it is $R_3(R_2 R_1) = (R_3 R_2)R_1$.
- Existence of an identity element: There exists an element $I \in SO(n)$ such that $IR = RI$ for every element $R \in SO(n)$.
- Existence of an inverse element: For each $R \in SO(n)$ exists an element $R^{-1} \in SO(n)$ such that $RR^{-1} = R^{-1}R = I$ where I is the identity element.

While the group $SO(2)$ is commutative, the group $SO(3)$ is not generally commutative as can be seen from the counter example

$$R_2(\pi/2, \mathbf{y})R_1(\pi/2, \mathbf{z})\mathbf{x} = R(\pi/2, \mathbf{y})\mathbf{y} = \mathbf{y}, \quad (1)$$

and

$$R_1(\pi/2, \mathbf{z})R_2(\pi/2, \mathbf{y})\mathbf{x} = R(\pi/2, \mathbf{z})(-\mathbf{z}) = -\mathbf{z}. \quad (2)$$

However, rotations in $\mathbb{R}^3$ about identical axes are commutative

$$R_2(\omega_2, \mathbf{n})R_1(\omega_1, \mathbf{n}) = R(\omega_1 + \omega_2, \mathbf{n}). \quad (3)$$

The Field $\mathbb{C}$ of Complex Numbers and Rotations in $\mathbb{R}^2$

A complex number $z \in \mathbb{C}$ is defined as

$$z = a + bi, i^2 = -1, \quad (4)$$

where a and $b \in \mathbb{R}$ are called a real and an imaginary part, respectively, and where i denotes the imaginary unit satisfying $i^2 = -1$. Complex numbers may be visualized as ordered pair $(\text{Re}(z), \text{Im}(z)) = (a, b)$ in the complex plane spanned by the basic units 1 and i , referred to as Argand diagram of complex numbers. The absolute value (modulus) and the argument (phase) of a complex number $z = a + bi$ are defined as

$$|z| = \sqrt{a^2 + b^2} = r, \quad \arg(z) = \theta. \quad (5)$$

The definition of the modulus implies the existence of an angle $\theta \in (-\pi, \pi]$ such that

$$a = r \cos \theta, \quad b = r \sin \theta, \quad (6)$$

and

$$z = r \cos \theta + ir \sin \theta = r(\cos \theta + i \sin \theta) = re^{i\theta}. \quad (7)$$

Especially for $r = 1$, it is

$$z = e^{i\theta}, \theta \in (-\pi, \pi], \quad (8)$$

carrying the visually inclined interpretation that a complex number z of modulus 1 varies along the unit circle in the complex plane centered at the origin as θ varies in $(-\pi, \pi]$, i.e., it is subject of a rotation by the angle θ about the axis $\mathbf{z} \in \mathbb{R}^3$ orthogonal to the complex plane through the origin of the coordinate system.

Multiplying $z_1 = \cos \theta_1 + i \sin \theta_1 = e^{i\theta_1}$, $z_2 = \cos \theta_2 + i \sin \theta_2 = e^{i\theta_2}$ results in

$$\begin{aligned} z_2 z_1 &= e^{i\theta_2} e^{i\theta_1} = e^{i(\theta_1 + \theta_2)} \\ &= \cos(\theta_1 + \theta_2) + i \sin(\theta_1 + \theta_2) \end{aligned} \quad (9)$$

reminiscent of the concatenation of two rotations $R_1(\theta_1, \mathbf{z})$ and $R_2(\theta_2, \mathbf{z})$ by angles θ_1 and θ_2 about the common axis of rotation $\mathbf{z} \in \mathbb{R}^3$, Eq. (3), leaving the complex plane invariant.

Real Quaternions Extending the Number System

A real quaternion $q \in \mathbb{H}$ is defined as

$$q = a + b_i + c_j + dk, \quad i^2 = j^2 = k^2 = ijk = -1, \quad (10)$$

where $a, b, c, d \in \mathbb{R}$ (Hamilton 1844). Then $a = \text{Sc}(q) = q_0$ is called scalar part, and $bi + cj + dk = \text{Vec}(q) = \mathbf{q}$ is called vector part of the real quaternion $q = q_0 + \mathbf{q}$. Hamilton's definition $i^2 = j^2 = k^2 = ijk = -1$ of the four basic units $1, i, j, k$, of real quaternions may be rewritten more instructively as a set of algebraic equations

$$(i) \quad i^2 = j^2 = k^2 = -1, \quad (11)$$

$$(ii) \quad ij = k, jk = i, ki = j, \quad (12)$$

$$(iii) \quad ij + ji = jk + kj = ik + ki = 0. \quad (13)$$

Real quaternions provide an extension of the hierarchical system of numbers beyond complex numbers.

If $\text{Sc}q = 0$, then q is called a pure quaternion; the subset of all pure quaternions is denoted $\text{Vec}\mathbb{H}$. For $q \in \text{Vec}\mathbb{H}$, q and $\mathbf{q}$ are identified, i.e., $q = \mathbf{q}$. The subset of all scalars may be denoted $\text{Sc}\mathbb{H}$. In this way, $\mathbb{R}$ and $\mathbb{R}^3$ are embedded in $\mathbb{H}$.

Given two quaternions, $p, q \in \mathbb{H}$, their quaternion product according to the algebraic rules, Eq. (11) above, is given by

$$pq = p_0q_0 - \mathbf{p} \cdot \mathbf{q} + p_0\mathbf{q} + q_0\mathbf{p} + \mathbf{p} \times \mathbf{q}, \quad (14)$$

where $\mathbf{p} \cdot \mathbf{q}$ and $\mathbf{p} \times \mathbf{q}$ represent the standard inner and cross product in $\mathbb{R}^3$; thus

$$\text{Sc}(pq) = p_0q_0 - \mathbf{p} \cdot \mathbf{q}, \quad (15)$$

$$\text{Vec}(pq) = p_0\mathbf{q} + q_0\mathbf{p} + \mathbf{p} \times \mathbf{q}. \quad (16)$$

Obviously, the quaternion product is not commutative.

The quaternion

$$q^* = \text{Sc}q - \text{Vec}q \quad (17)$$

is called the conjugate of q . With conjugated quaternions, it is possible to represent the scalar and vector part in an algebraic fashion as

$$q_0 = \text{Sc}q = \frac{1}{2}(q + q^*), \quad (18)$$

$$\mathbf{q} = \text{Vec}q = \frac{1}{2}(q - q^*). \quad (19)$$

It holds that

$$qq^* = q^*q = \|q\|^2 = q_0^2 + (q_1)^2 + (q_2)^2 + (q_3)^2, \quad (20)$$

and the number $\|q\|$ is called the norm of q . The norm of quaternions coincides with the Euclidean norm of q regarded as an element of the vector space $\mathbb{R}^4$. The scalar part of pq^* agrees with the usual Euclidean inner product in the space $\mathbb{R}^4$ and gives rise to the definition to the dot product of two real quaternions

$$p \cdot q = \text{Sc}(pq^*) = \sum_{i=0}^3 p_i q_i. \quad (21)$$

Since $(pq)^* = q^*p^*$, it is $\|pq\| = \|p\| \|q\|$. A quaternion q with $\|q\| = 1$ is called a unit quaternion. Furthermore, let $\mathbb{S}^2$ denote the unit sphere in $\text{Vec}\mathbb{H} \simeq \mathbb{R}^3$ of all unit vectors, and $\mathbb{S}^3$ the sphere in $\mathbb{H} \simeq \mathbb{R}^4$ of all unit quaternions. In complete analogy to $\mathbb{S}^2 \subset \mathbb{R}^3$, $\text{Sc}(pq^*) = p \cdot q$ provides a canonical measure for the spherical distance of unit quaternions $p, q \in \mathbb{S}^3$, i.e., the angle enclosed by p and q .

Each nonzero quaternion q has a unique inverse

$$q^{-1} = q^* / \|q\|^2 \quad (22)$$

with $\|q^{-1}\| = \|q\|^{-1}$. For unit quaternions, it is $q^{-1} = q^*$; for pure unit quaternions, it is $q^{-1} = -q$, implying $qq = -1$. The inverse of the unit quaternion $q = q_0 + \mathbf{q}$ is $q = q_0 - \mathbf{q}$.

Two quaternions p and $q \in \mathbb{H}$ are said to be orthogonal if pq^* is a pure quaternion, i.e., if $\text{Sc}(pq^*) = p \cdot q = 0$. According to Eq. (19), the condition of orthogonality may be rewritten as

$$2 \text{Sc}(pq^*) = pq^* + qp^* = 0. \quad (23)$$

If p and q are orthogonal unit quaternions, they are called orthonormal quaternions. It is emphasized that pure quaternions with orthogonal vector parts are orthogonal, but that the inverse is not generally true. Orthogonality of two quaternions does not imply orthogonality of their vector parts unless they are pure quaternions.

The Skew Field $\mathbb{H}$ of Real Quaternions and Rotations in $\mathbb{R}^3$

Any quaternion permits the representation

$$q = \|q\| \left(\frac{q_0}{\|q\|} + \frac{\mathbf{q}}{\|\mathbf{q}\|} \frac{\|q\|}{\|q\|} \right) \quad (24)$$

which implies the existence of an angle $\omega \in [0, \pi]$ such that

$$q = \|q\| \left(\cos \frac{\omega}{2} + \frac{\mathbf{q}}{\|\mathbf{q}\|} \sin \frac{\omega}{2} \right) \quad (25)$$

with $\omega = 2 \arccos(q_0/\|q\|)$, and $\|\mathbf{q}\|^2 = \mathbf{q}\mathbf{q}^*$ considering $\mathbf{q}$ as a pure quaternion. For an arbitrary unit quaternion, the representation Eq. (25) simplifies to

$$q = \cos \frac{\omega}{2} + \mathbf{n} \sin \frac{\omega}{2} \quad (26)$$

with the unit vector $\mathbf{n} = \mathbf{q}/\|\mathbf{q}\| \in \mathbb{S}^2$. The representation Eq. (26) suggests the association of a counterclockwise rotation in $\mathbb{R}^3$ by the angle ω about the axis $\mathbf{n} \in \mathbb{S}^2$. In fact, any rotation $R(\omega, \mathbf{n})$ by the angle $\omega \in [-\pi, \pi]$ about the axis $\mathbf{n} \in \mathbb{S}^2$ mapping the unit vector $\mathbf{u} \in \mathbb{S}^2$ onto the unit vector $\mathbf{v} \in \mathbb{S}^2$ according to

$$R(\omega, \mathbf{n})\mathbf{u} = \mathbf{v} \quad (27)$$

can be written in terms of the unit quaternion $q = \cos \frac{\omega}{2} + \mathbf{n} \sin \frac{\omega}{2}$ applying quaternion multiplication as

$$quq^* = \mathbf{v}, \quad (28)$$

where u and v are pure unit quaternions with vector parts $\mathbf{u}, \mathbf{v} \in \mathbb{S}^2$ and quaternion multiplication applies; for a proof, the reader is referred to (Downs 2003). Equation (28) reads explicitly

$$\mathbf{v} = \mathbf{u} \cos \omega + (\mathbf{n} \times \mathbf{u}) \sin \omega + (\mathbf{n} \cdot \mathbf{u})\mathbf{n}(1 - \cos \omega) \quad (29)$$

which is Rodrigues' formula for rotations in $\mathbb{R}^3$ (Rodrigues 1840). With $q^*vq = u$, the unit quaternion q^* provides the inverse rotation $R^{-1}(\omega, \mathbf{n})\mathbf{v} = R(\omega, -\mathbf{n})\mathbf{v} = \mathbf{u}$. Conjugation RR_0R^{-1} of a given rotation $R_0(\omega_0, \mathbf{n}_0)$ with an arbitrary rotation R is easily understood in terms of quaternions as the rotation

$$\begin{aligned} pq_0p^* &= p \left(\cos \frac{\omega_0}{2} + \mathbf{n}_0 \sin \frac{\omega_0}{2} \right) p^* \\ &= \cos \frac{\omega_0}{2} + p\mathbf{n}_0p^* \sin \frac{\omega_0}{2} \end{aligned} \quad (30)$$

about the rotated axes $p\mathbf{n}p^* \in \mathbb{S}^2$ by the same angle ω . Thus, conjugation may be applied to generate all rotations with a given angle of rotation.

Since the antipodal quaternions q and $-q$ yield the same rotation, the unit quaternion sphere $\mathbb{S}^4 \subset \mathbb{H}$ covers the special

orthogonal group $\text{SO}(3)$ twice. The association of rotations in $\mathbb{R}^3$ with unit quaternions in $\mathbb{S}^3$ implies the interpretation that the unit sphere of vectors $\mathbb{S}^2 \subset \mathbb{R}^3$ thought of as being embedded in $\text{Vec}\mathbb{H}$ is associated with all rotations by the angle π about arbitrary axes. Comprehensive expositions of real quaternions and their relationships to rotations in $\mathbb{R}^3$ are given in (Altmann 2013; Hanson 2006; Kuipers 1999).

Concatenation of Rotations

Let the rotations $R_1 = R_1(\omega_1, \mathbf{n}_1)$ and $R_2 = R_2(\omega_2, \mathbf{n}_2)$ be associated with unit quaternions

$$q = \cos \frac{\omega_1}{2} + \mathbf{n}_1 \sin \frac{\omega_1}{2} \quad (31)$$

$$p = \cos \frac{\omega_2}{2} + \mathbf{n}_2 \sin \frac{\omega_2}{2}. \quad (32)$$

Then the concatenation $R = R(\omega, \mathbf{n}) = R_2(\omega_2, \mathbf{n}_2)R_1(\omega_1, \mathbf{n}_1)$ is associated with the product pq as defined in Eq. (14)

$$pq = p_0q_0 - \mathbf{p} \cdot \mathbf{q} + p_0\mathbf{q} + q_0\mathbf{p} + \mathbf{p} \times \mathbf{q}, \quad (33)$$

and its scalar and vector parts, respectively,

$$\cos \frac{\omega}{2} = \text{Sc}(pq) = p_0q_0 - \mathbf{p} \cdot \mathbf{q}, \quad (34)$$

$$\mathbf{n} \sin \frac{\omega}{2} = \text{Vec}(pq) = p_0\mathbf{q} + q_0\mathbf{p} + \mathbf{p} \times \mathbf{q} \quad (35)$$

provide the angle ω and the axis $\mathbf{n}$ of the rotation R . It resolved the historical quest for angle and axis of sequential compositions of rotations in terms of quaternions in a straightforward way. An Earth science application is presented in detail in the entry ► “Euler Poles of Tectonic Plates” of this Encyclopedia where the path of a migrating relative Euler pole given two absolute Euler poles is determined.

Geodesics

Let q_1 and $q_2 \in \mathbb{S}^3$ be two orthonormal quaternions. The set of quaternions

$$q(t) = q_1 \cos t + q_2 \sin t, \quad t \in (-\pi, \pi], \quad (36)$$

parametrizes the great circle denoted $C(q_1, q_2) \subset \mathbb{S}^3$ spanned by q_1 and q_2 . Obviously, $-q_1$ and $-q_2 \in \mathbb{S}^3$ define the same great circle $C(q_1, q_2)$. Any $q(t) \in C(q_1, q_2)$ satisfies (Meister and Schaeben 2004)

$$q(t)(q_1^*q_2)q^*(t) = q_2q_1^*, t \in (-\pi, \pi]. \quad (37)$$

Since q_1 and q_2 are orthonormal, $q_1^*q_2$ and $q_2q_1^*$ define pure unit quaternions $\mathbf{u} = q_1^*q_2$, $\mathbf{v} = q_2q_1^*$. For example,

$$\begin{aligned} q_1 &= \cos\left(\frac{\arccos \mathbf{u}\mathbf{v}}{2}\right) + \frac{\mathbf{u} \times \mathbf{v}}{\|\mathbf{u} \times \mathbf{v}\|} \sin\left(\frac{\arccos \mathbf{u}\mathbf{v}}{2}\right) \\ &= \frac{1 - \mathbf{v}\mathbf{u}}{\|1 - \mathbf{v}\mathbf{u}\|} \end{aligned} \quad (38)$$

$$q_2 = \frac{\mathbf{u} + \mathbf{v}}{\|\mathbf{u} + \mathbf{v}\|} = \frac{u + v}{\|u + v\|} \quad (39)$$

for $\mathbf{u}, \mathbf{v} \in \mathbb{S}^2$ with $\mathbf{v} \neq -\mathbf{u}$.

The fiber $G(\mathbf{u}, \mathbf{v}) \subset \text{SO}(3)$ of all rotations R with $R\mathbf{u} = \mathbf{v}$ is associated with the great circle $C(q_1, q_2) \subset \mathbb{S}^3 \subset \mathbb{H}$ spanned by unit quaternions $q_1, q_2 \in \mathbb{S}^3$ given in terms of $\mathbf{u}, \mathbf{v} \in \mathbb{S}^2$ by Eqs. (38) and (39), for example. Since $q(t)$ and $-q(t)$ are associated with the same rotation, the circle $C(q_1, q_2)$ covers the fiber $G(\mathbf{u}, \mathbf{v})$ twice. The fibers $G(\mathbf{u}, \mathbf{v}), (\mathbf{u}, \mathbf{v}) \in \mathbb{S}^2 \times \mathbb{S}^2$ induce a double fibration or double covering of $\text{SO}(3)$ as

$$\bigcup_{\mathbf{u} \in \mathbb{S}^2} G(\mathbf{u}, \mathbf{v}) = \bigcup_{\mathbf{v} \in \mathbb{S}^2} G(\mathbf{u}, \mathbf{v}) = \text{SO}(3) \quad (40)$$

for any fixed $\mathbf{u}$ or fixed $\mathbf{v}$, respectively. In the same way, the great circles, $C_{\mathbf{u}, \mathbf{v}} \subset \mathbb{S}^3$, induce a double fibration of $\mathbb{S}^3$.

A major property of the great circle $C(q_1, q_2) \equiv C_{\mathbf{u}, \mathbf{v}}$ is that it can be parametrized by a pair $(\mathbf{u}, \mathbf{v}) \in \mathbb{S}^2 \times \mathbb{S}^2$ and its antipodally symmetric $(-\mathbf{u}, -\mathbf{v})$ in the way that it consists of all quaternions $q(t), t \in (-\pi, \pi]$, with $q(t)\mathbf{u}q^*(t) = \mathbf{v}$ for all $t \in (-\pi, \pi]$, and that it covers the fiber $G(\mathbf{u}, \mathbf{v})$ twice. While it is obvious that great circles $C(q_1, q_2)$ of $\mathbb{S}^3$ are its geodesics, their parametrization $C_{\mathbf{u}, \mathbf{v}}$ in terms of two unit vectors $\mathbf{u}$ and $\mathbf{v}$ such that the great circle corresponds to the fiber $G(\mathbf{u}, \mathbf{v}) \subset \text{SO}(3)$ of all rotations mapping $\mathbf{u}$ onto $\mathbf{v}$ reveals that the fibers are the geodesics of $\text{SO}(3)$. In fact, they are actually Hopf fibers.

Replacing $\mathbf{u}$ by $-\mathbf{u}$ in Eqs. (38) and (39) yields

$$\begin{aligned} q_3 &= \cos\left(\frac{\arccos(-\mathbf{u}\mathbf{v})}{2}\right) \\ &\quad + \frac{-\mathbf{u} \times \mathbf{v}}{\|-\mathbf{u} \times \mathbf{v}\|} \sin\left(\frac{\arccos(-\mathbf{u}\mathbf{v})}{2}\right) \\ &= \frac{1 + \mathbf{v}\mathbf{u}}{\|1 + \mathbf{v}\mathbf{u}\|} \end{aligned} \quad (41)$$

$$q_4 = \frac{-\mathbf{u} + \mathbf{v}}{\|-\mathbf{u} + \mathbf{v}\|} = \frac{-u + v}{\|-u + v\|} \quad (42)$$

and results in four mutually orthonormal quaternions $q_1, \dots, q_4$. Thus, the two circles $C(q_1, q_2)$ and $C(q_3, q_4)$ associated with the two fibers $G(\mathbf{u}, \mathbf{v})$ and $G(-\mathbf{u}, \mathbf{v})$, respectively, are

orthonormal to one another. Integration of suitable functions along geodesics gives rise to their Funk or Radon transforms, respectively. The Funk-Radon transform of an orientation probability density function defined on $\mathbb{S}^3$ provides the basic model of integral orientation measurements with X-ray or Neutron diffraction as applied in the analysis of “Crystallographic Preferred Orientation” (this Encyclopedia).

Conclusions

“Spoiler alert: unit quaternions provide ‘the’ way to represent rotations. Why? Unit quaternions allow a clear visualization (see Hanson 2006) of the space of rotations as the unit sphere $\mathbb{S}^3$ in four dimensions (with antipodal points identified). Unit quaternions make it possible to differentiate an expression ‘with respect to rotation’. Unit quaternions make it possible to find extrema of expressions by setting that derivative equal to zero! Unit quaternions make it easy to compose rotations (unlike, e.g., axis- and-angle notation). Unit quaternions do not suffer from singularities (as do, e.g., Euler angles when two axes line up see gimbal lock). Unit quaternions, while redundant (four parameters for three degrees of freedom), have only one constraint on their components (unlike orthonormal matrices, which have six non-linear constraints on the rows or columns plus one on the determinant). Unit quaternions make it possible to compute averages, to interpolate, to sample in the space of rotations and to represent densities over orientations. And, as demonstrated in Andrew Hansons (2020) article in the previous issue of Acta Crystallographica Section A, they also make possible insightful illustrations relating to orientations.” (Horn 2020)

Bibliography

- Altmann SL (1989) Math Mag 62:291
 Altmann SL (2013) Rotations, quaternions, and double groups. Dover Publications, Mineola/New York
 Downs L (2003). <http://www.cs.berkeley.edu/~laura/cs184/quat/quatproof.html>
 Gauss CF (1900) Mutationen des Raumes. In: Königlichen Gesellschaft der Wissenschaften (ed) Carl Friedrich Gauss Werke, vol 8. Springer, Berlin/Heidelberg, pp 357–362. https://doi.org/10.1007/978-3-642-92474-3_67
 Hamilton WR (1844) Phil Mag 25:489
 Hanson AJ (2006) Visualizing quaternions. Morgan–Kaufmann, San Francisco
 Hanson AJ (2020) Acta Cryst A76:432
 Horn BKP (2020) Acta Cryst A76:556
 Kuipers JB (1999) Quaternions and rotation sequences: a primer with applications to orbits, aerospace, and virtual reality. Princeton University Press
 Meister L, Schaeben H (2004) Math Methods Appl Sci 28:101
 Novelia A, O'Reilly OM (2015) Regul Chaotic Dyn 20:729
 Pujol J (2012) Commun Math Anal Commun Math Anal 13:1
 Rodrigues O (1840) J Math Pures Appl Liouville 5:380

Radial Basis Functions

Michael Hillier

Geological Survey of Canada, Ottawa, ON, Canada

Definition

Radial Basis Functions (RBFs) are used for scattered data approximation, a numerical method for approximating an unknown function $f(\mathbf{x})$ with an interpolant $s(\mathbf{x})$ from a distinct set of N sampled scattered data points $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\} \subseteq \mathbb{R}^d$ having data values $f(\mathbf{x}_j)$, $1 \leq j \leq N$. RBF interpolants provide optimal approximation of $f(\mathbf{x})$ and can be computed efficiently using RBF-based optimizations for domain decomposition, fast matrix solvers, and interpolant evaluators. In geoscience, RBF interpolation emerged as a useful method to rapidly compute geological features for various applications. RBF interpolants are expressed in terms of linear combinations of radially symmetric kernel functions of the form $\Phi(\mathbf{x}, \mathbf{x}_j) := \phi(r)$ centered on data point $\mathbf{x}_j$ where $r = \|\mathbf{x} - \mathbf{x}_j\|_2$ is the Euclidean distance between two points. These interpolants have the form

$$s(\mathbf{x}) = \sum_{j=1}^N w_j \Phi(\mathbf{x}, \mathbf{x}_j) + p(\mathbf{x}) \quad (1)$$

where $p(\mathbf{x})$ is a l -degree d -variate polynomial

$$p(\mathbf{x}) = \sum_k^Q c_k p_k(\mathbf{x}), Q = \frac{(l+d)!}{l!d!} \quad (2)$$

Coefficients w_j and c_k are determined from the linear system

$$\begin{bmatrix} A & P \\ P^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{c} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ 0 \end{bmatrix} \quad (3)$$

where $A = \Phi(\mathbf{x}_j, \mathbf{x}_k) \in \mathbb{R}^{N \times N}$ and $P = p_k(\mathbf{x}_j) \in \mathbb{R}^{N \times Q}$ is linearly independent and generated from the interpolation conditions $s(\mathbf{x}_j) = f(\mathbf{x}_j)$, $1 \leq j \leq N$. The block matrix on the left-hand side of Eq. 3 is referred to as the interpolation matrix. Optimal RBF interpolants providing the smoothest possible interpolation and minimal pointwise error are guaranteed from the solution of Eq. 3 when A is positive definite, defined by $\mathbf{w}^T A \mathbf{w} > 0$. RBFs can be split into two categories: strictly positive definite (SPD) and conditionally positive definite (CPD). Commonly used SPD and CPD RBFs are listed in Tables 1 and 2, respectively. Refer to Wendland (2005) for detailed analysis and properties of these functions. For SPD RBFs, A is always positive definite, whereas conditionally positive definite (CPD) RBFs of order m require the orthogonality constraints

$$\sum_j^N w_j p_k(\mathbf{x}_j) = 0, \quad 1 \leq k \leq Q \quad (4)$$

to ensure positive definiteness where $p(\mathbf{x})$ is at most a $(m-1)$ -degree polynomial. Furthermore, for SPD RBFs, low-degree polynomials are not required however can be useful for extrapolation. Notations used in this definition are given in Table 3.

Introduction

Scattered data approximation methods utilizing radial basis functions originally appeared in spatially dependent geoscience applications such as geodesy, hydrology, and geophysics in the early 1970s. A thorough review on these early works, using multiquadric and inverse multiquadric RBFs, is found in Hardy (1990). Beyond the 1980s, a myriad of uses for RBFs were found in a wide range of applications including artificial intelligence, solving PDEs, optimization, finance, signal processing, and 3D modeling. Their success can be

Radial Basis Functions, Table 1 Common SPD RBFs

Strictly positive definite (SPD) RBF	$\phi(r)$
Gaussian	$e^{-(\varepsilon r)^2}$
Inverse multiquadric	$(\varepsilon^2 + r^2)^{k/2} \quad k < 0$
Wendland (C^2)	$(1 - r_c)^4(4r_c + 1), r_c < 1$ $0, r_c \geq 1$
Wendland (C^4)	$(1 - r_c)^6(35r_c^2 + 18r_c + 3), r_c < 1$ $0, r_c \geq 1$
Matérn (C^0)	$e^{-\varepsilon r}$
Matérn (C^2)	$e^{-\varepsilon r}(1 + \varepsilon r)$
Matérn (C^4)	$e^{-\varepsilon r}(3 + 3\varepsilon r + (\varepsilon r)^2)$

Smaller shape parameters ε correspond to flatter functions as r increases. For Wendland's functions, the variable $r_c = r/R$ where R is a cutoff radius. For values $r_c > 1$, the function is zero

Radial Basis Functions, Table 2 Common CPD RBFs and their conditional order m

Conditionally positive definite (CPD) RBF	$\phi(r)$	Conditional order m
Multiquadric	$(\varepsilon^2 + r^2)^{k/2} \quad k > 0, k \notin 2\mathbb{N}$	$\lceil k/2 \rceil$
Radial powers	$r^k \quad k > 0, k \notin 2\mathbb{N}$	$\lceil k/2 \rceil$
Thin-plate splines	$r^{2k} \log r \quad k > 0, k \in \mathbb{N}$	$k + 1$

Radial powers and thin-plate splines are known as polyharmonic splines

Radial Basis Functions, Table 3 Notations and associated descriptions used in the RBF definition

Notations	Descriptions
d	Dimensionality
$\mathbf{x}$	d -dimensional point, e.g., 3D point $\mathbf{x} = (x, y, z)$
X	Set of scattered points $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$. Scattered points are irregularly distributed
N	Number of data points in point set X
$\mathbf{x}_j$	j -th data point from X
$f(\mathbf{x})$	Unknown scalar value at point $\mathbf{x}$
$s(\mathbf{x})$	Interpolated scalar value at point $\mathbf{x}$
$\Phi(\mathbf{x}, \mathbf{x}_j)$	Kernel function. Two points ($\mathbf{x}$ and $\mathbf{x}_j$) are input, and output is a scalar
r	Euclidean distance between two points
$\phi(r)$	Radial basis function
l	Degree of polynomial
$p(\mathbf{x})$	l -degree polynomial, e.g., 1-degree trivariate (3D): $p(\mathbf{x}) = c_1x + c_2y + c_3z + c_4$
$p_k(\mathbf{x})$	k -th monomial of polynomial, e.g., 1-degree trivariate polynomial has 4 monomials: $(x, y, z, 1)$, 2-degree trivariate polynomial has 10 monomials: $(x^2, y^2, z^2, xy, xz, yz, x, y, z, 1)$
Q	Number of monomials for polynomial
$\mathbf{w}$	RBF coefficients $\mathbf{w} = (w_1, \dots, w_N)$
$\mathbf{c}$	Polynomial coefficients $\mathbf{c} = (c_1, \dots, c_Q)$
A	Symmetric kernel matrix generated by radial kernel $\Phi(\mathbf{x}, \mathbf{x}_j)$ using all of point pairs of point set X
P	Polynomial matrix containing all monomials for point set X
$\mathbf{w}^T A \mathbf{w}$	Quadratic form associated with A . Interpolant smoothness is measured in terms of the quadratic form: $\ s\ ^2 = \mathbf{w}^T A \mathbf{w}$, where $\ s\ $ is the interpolant's norm. Positive definiteness of the quadratic form (e.g., $\mathbf{w}^T A \mathbf{w} > 0$) ensures the smoothest possible interpolation
m	Conditional order of CPD RBF
ε	SPD RBF shape parameter
R	Cutoff radius for Wendland's RBFs
r_c	Wendland's RBF variable

attributed to the robustness, versatility, and simplicity of the mathematical models that use them.

Important historical milestones beyond Hardy's pioneering work will be given. The variational approach

developed by Duchon (1976) demonstrated that polyharmonic splines, like thin plate splines, minimize the bending energy (e.g., aggregated squared curvature). These characteristics enabled modeling of complex surface

geometries with arbitrary genus topology in addition to desirable surface smoothness and hole-filling properties. In the mid-1980s, Micchelli (1986) proved the interpolation matrix is always invertible and laid a strong foundation for future work. By the late 1990s and early 2000s, these theoretical foundations coupled with dramatic increases in computing power, and RBF-based optimization techniques led to widespread use of implicit modeling of complex 3D objects from large datasets (Carr et al. 2001). The most common use of RBFs in geoscience is with implicit modeling.

Interestingly, the geostatistical method of interpolation known as kriging which takes a stochastic point of view of data observations as realizations of random variables belonging to a random field has been proven mathematically equivalent to RBF interpolation (Dubrule 1984; Scheuerer et al. 2013). Although these connections have been known by geostatisticians for some time, it has been only recently recognized by the numerical analysis community. RBF interpolation can be viewed as a subset of kriging with a chosen fixed covariance function and polynomial trend degree. In kriging, covariance functions are estimated from data.

RBF Implicit Modeling

Implicit modeling is a method by which iso-contours implicitly defined as some level set of a reconstructed scalar function are extracted from a discretized domain (e.g., triangulated meshes, voxel grids). Iso-contours, such as curves (2D) and surfaces (3D), require interpolants to be first evaluated throughout the discretized domain. Iso-contour extraction from evaluated domains is performed using computer graphic algorithms such as marching cubes (Lorensen and Cline 1987) (for 3D applications).

In geoscience, RBF-based implicit modeling methods are routinely applied to geologically derived structures and properties (Cowan et al. 2002) and provide a fast, automatic, and reproducible approach. Figure 1 illustrates RBF implicit modeling for 3D structural geology and ore grade estimation applications. Information regarding geological setting and data used for these examples can be found in de Kemp et al. 2016. Implicit modeling greatly improved upon shortcomings of previously widely used approaches for these applications called explicit modeling. Explicit-based modeling approaches required manual wireframe construction by Bezier and NURB objects by a modeler using CAD tools. Although explicit modeling offered great precision and flexibility, they suffered from lengthy nonreproducible user-dependent wireframe creation, as well as being difficult to update to newly acquired data – a common occurrence in geoscience.

Many geoscience applications suffer from limited data observations and give rise to nonunique models. There are an infinite set of solutions describing given observations. In

underdetermined settings, domain experts would like to test possible scenarios and hypotheses explaining physical phenomena apparent from data observations. Implicit modeling provides a framework in which many of these scenarios can be quickly tested and compared. Furthermore, recent developments in implicit model uncertainty characterization using Bayesian inference can quantify the degree to which scenarios are more likely (de la Varga and Wellmann 2016). These considerations make implicit modeling particularly advantageous in geoscience applications.

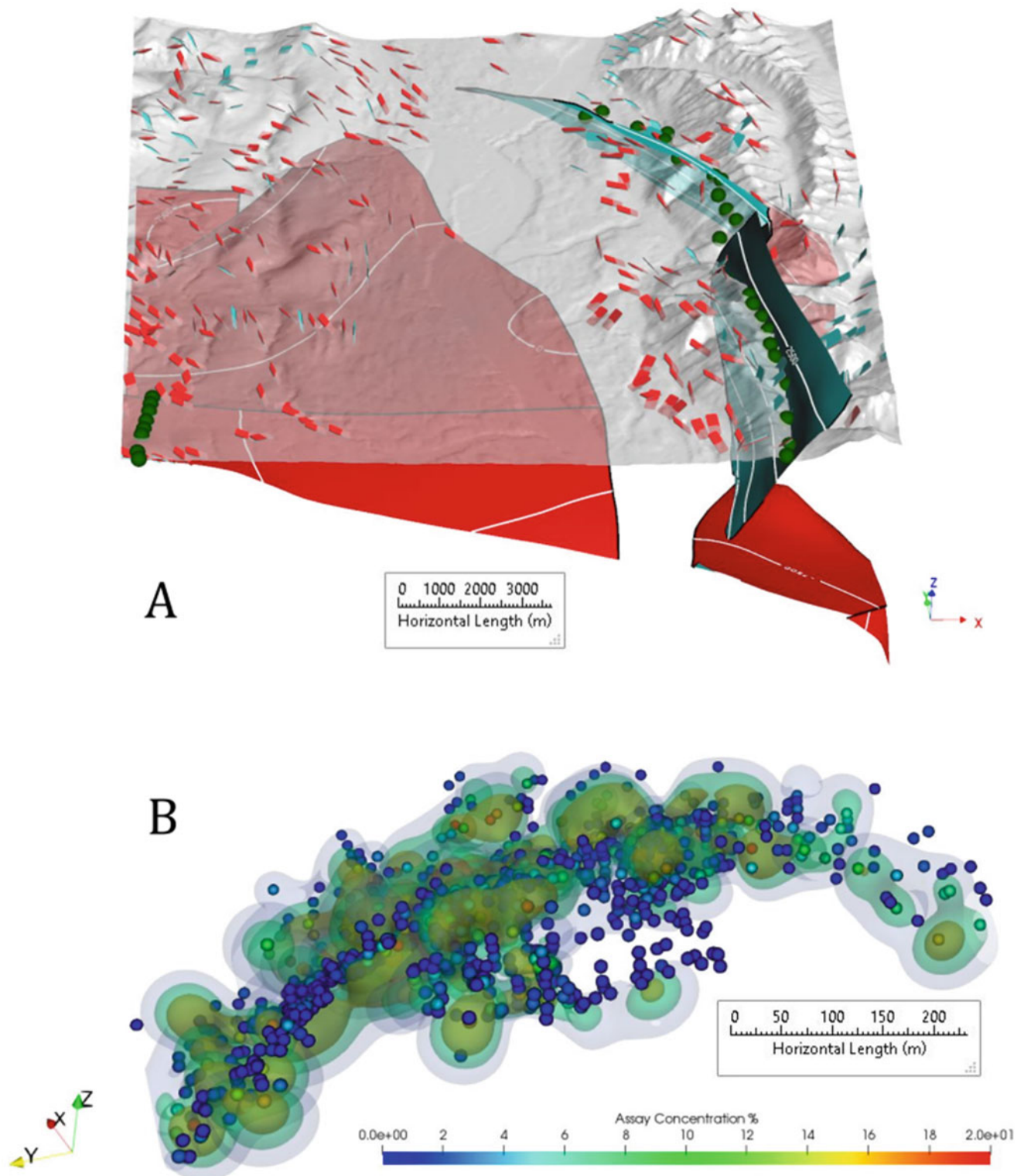
For structural RBF implicit modeling, geologically derived surfaces, such as interfaces between stratigraphic layers, are commonly implicitly defined as the zero-level set of the reconstructed scalar function (e.g., $f(\mathbf{x}) = 0$). In order to obtain nontrivial solutions to Eq. 3, additional points, called offset points, are generated by projecting known surface points attributed with planar orientation (e.g., normal), either measured or derived. Constraints for scalar values of offset points are set to its projection distance. Offset points above the surface distances are positive, whereas points below distances are negative. Geometrical artifacts in modeled surfaces using offset points can occur when modeling highly variant structures, since depending on the chosen projection distance points can be projected on the wrong side of the surface. Structural field observation datasets predominantly contain orientation measurements not directly sampling the surface of interest. Since off-surface orientation measurements influence the surface of interest's geometry, with closer measurements having greater impact than further measurements, the offset point approach to implicit modeling for such applications is far from optimal. Problems arising from the offset points were resolved by extensions to the basic RBF interpolation method explained in the next section.

RBF Methods and Enhancements for Geoscience

RBF interpolants of the form Eq. 1 can be expanded to include derivatives, inequality, surface points from multiple distinct surfaces, and fault constraints further enhancing modeling capacity for geoscience applications. Supplementary to the additional constraints, important RBF-based approximation methods will be described that may provide helpful for noisy datasets where exact interpolation should be avoided.

Derivative and Inequality Constraints

Incorporation of both derivatives and inequality constraints into RBF-interpolation can be achieved using generalized interpolation (Hillier et al. 2014) which employs a set of linearly independent functionals operating on RBFs. Linear functionals vary depending on constraint type. Derivative constraints such as gradients and directional derivatives



Radial Basis Functions, Fig. 1 Examples of RBF implicit modeling in geoscience: (a) overturned stratigraphic contact surface modeled using structural field observations: strike/dip orientations (red/blue colored tablets indicating stratigraphic top/bottom, respectively, of contact) and

green contact points from map traces and drill core; (b) ore grade modeling using metal assay data from drill core point samples (colored points). Iso-surfaces represent low, medium, and high ore grade

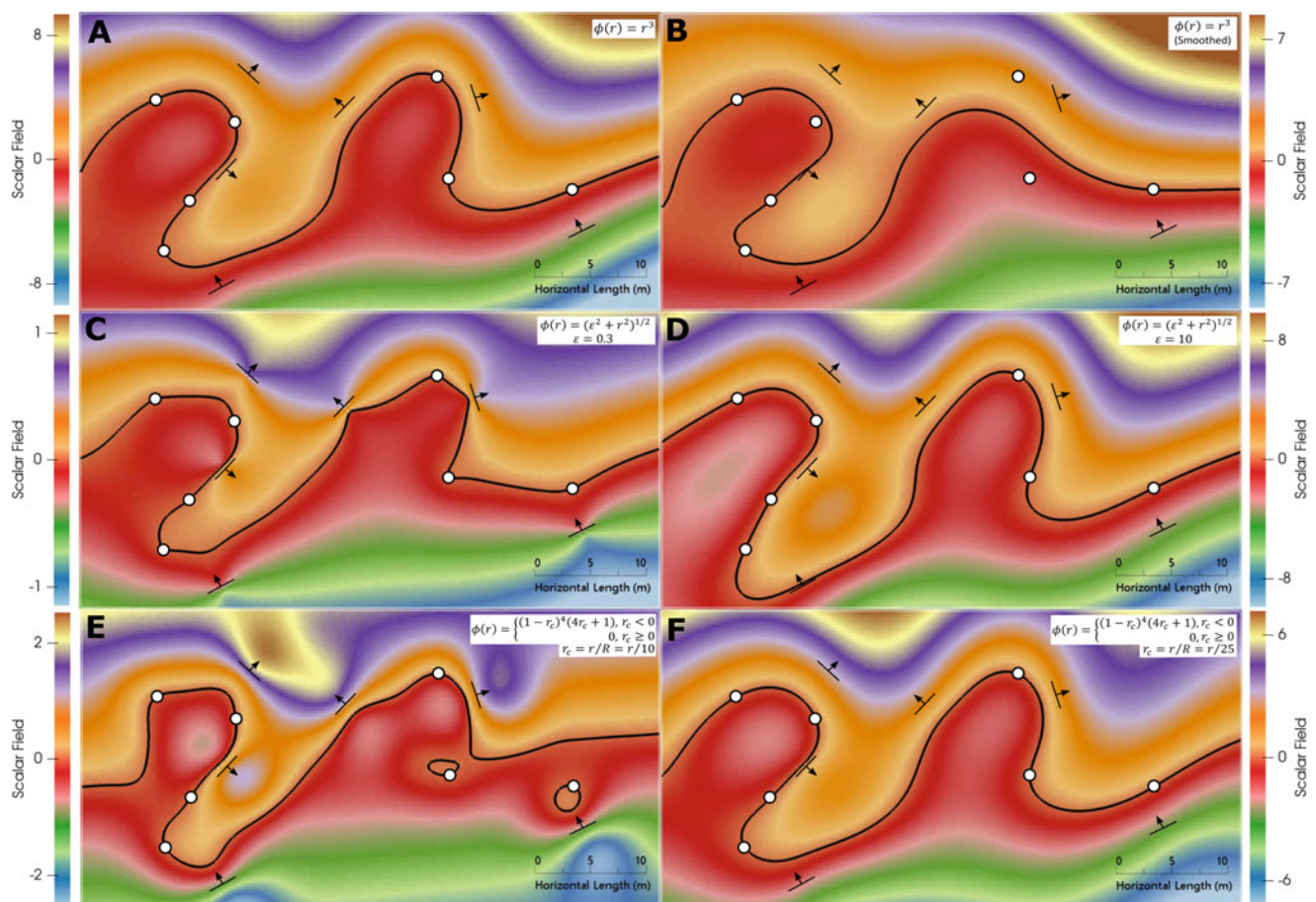
allow orientation data to be directly included into interpolation, overcoming challenges associated with the offset point approach for influencing modeled orientation. In addition, these constraints enhance continuity and extension of sampled structural features. Several scalar field interpolations using derivative constraints for orientation data (strike/dip with polarity) are shown in Fig. 2. Inequality constraints provide a means for imposing lower or upper bounds, as well as both, on data values. These constraints can also be leveraged for approximation purposes rather than exact interpolation (Fig. 2b). Furthermore, for structural modeling, inequality constraints can be used for rock unit observations where only their relative positions (e.g., above/below) to modeled surfaces are known. Determining optimal interpolants constrained by inequalities requires minimization of the interpolant's norm via quadratic optimization.

It is important to note that generalized RBF interpolants including geological interfaces, planar orientations, and tangent constraints (e.g., directional derivatives) are nearly identical to the interpolant from co-kriging in the dual form

(Lajaunie et al. 1997), differing only by the adoption of interface increments, indeed not surprising given the mathematical equivalence of RBF interpolation and kriging previously mentioned. Interface increments are constraints that permit the inclusion of multiple distinct surfaces simultaneously. To do so, a set of independent pairs of points for each distinct surface are constructed and added to the interpolant. The interpolation conditions for these constraints are the scalar field difference for each pair being equal to zero (e.g., points on the same iso-surface have equal scalar field values). Interface increments are particularly useful when modeling conformal stratigraphic geological volumes of rock and can also easily be implemented into the RBF interpolation using the same methodology.

Discontinuities

Discontinuous features can be incorporated into reconstructed scalar fields using RBF-based modeling. For 3D geological implicit modeling, discontinuous shifts in modeled rock volumes caused by geological faulting can be incorporated using



Radial Basis Functions, Fig. 2 Effect of different RBFs, methods, and parameters on scalar field interpolations for a synthetic dataset consisting of interface points (white circles) and strike/dip measurements with polarity (oriented line with arrow). Curve (black) represents the zero-valued iso-contour associated with the modeled interface. (a and b) cubic

radial power RBF that is fitted exactly and approximately (smoothed), respectively. (c and d) multiquadric RBF interpolations using different shape parameter ϵ values. (e and f) Wendland (C^2) RBF interpolations for two cutoff radii R

the three-step procedure developed by Calcagno et al. (2008). First, fault surface geometries represented by its own scalar field are modeled using corresponding fault constraints (e.g., fault locations, orientations). Second, descriptions for each fault's termination characteristics (e.g., infinite vs. finite) in addition to fault relationships (e.g., fault network), if applicable, using fault-fault relations are developed. The fault network relationships characterize the volumes of each fault's scalar field representation. Finally, to incorporate the sharp shifts in rock volumes across faults, discontinuous polynomial functions are added to the linear system Eq. 3 in a manner respecting the fault networks relationships.

Approximation Methods

Noisy and/or highly varying datasets can be problematic for exact RBF-interpolation. In these settings, geometrical artifacts and physically implausible solutions can be produced. Approximation methods are suggested for datasets possessing these attributes. There are four available approximation RBF-methods with varying degrees of implementation and computational complexity, including spline smoothing (Carr et al. 2001), use of inequality constraints, convolving an interpolant with a smoothing kernel (Carr et al. 2003), and greedy algorithms (Carr et al. 2001).

Spline smoothing simply requires user-specified values to be added to the diagonal elements of matrix A in Eq. 3. Larger values produce larger amounts of smoothing. However, choosing optimal values to add which are physically meaningful can be difficult. Obtaining the desired amount of smoothing requires a trial and error approach.

As previously mentioned, inequality constraints can be useful for approximation purposes (Fig. 2b). RBF-interpolants are optimal in the sense they have minimum norms where smaller interpolant norms correspond to smoother solutions. Since inequality constraints expand the range of possible solutions, interpolants that incorporate them will be smoother as compared to exact interpolation. The disadvantage of this approach is the additional computational complexity resulting from the quadratic optimization needed to obtain the minimum norm interpolants.

Convolution-based smoothing acts as a loss pass filter and is achieved by changing the interpolant's RBF with its associated smoothing kernel. A key advantage of this method is that the interpolant is only fitted once, and interpolant evaluations can use associated RBF-smoothing kernels with varying degrees of smoothing using the same fitted coefficients.

Greedy algorithms begin by fitting the interpolant using a small random subset of data. Next, residuals measuring the difference between data and interpolant values are computed on remaining points not included in the initial data subset by interpolant evaluation. Points corresponding to residuals beyond a user-defined threshold are added to the interpolant and refitted. The algorithm terminates when all residuals are

below the user-defined threshold. The disadvantage with this approach is the sensitivity to data outliers.

RBF Selection

RBF interpolation requires a specific RBF to be chosen for the interpolant (Eq. 1). Depending on data sampling, application, and model complexity (e.g., highly deformed geology), some RBFs may perform better than others. However, currently, there is a knowledge gap in determining the best RBF depending on these factors for geoscience applications and presents opportunities for future research. Nevertheless, awareness of the following two observations can be useful: (1) Interpolation differences between RBFs increase as data sparsity increases, (2) CPD RBFs listed in Table 2 generally perform well for surface modeling. The effect of using different RBFs (radial power, multiquadric, and Wendland) and associated parameters from Tables 1 and 2 on scalar field interpolations for a synthetic dataset is illustrated in Fig. 2. Interpolations presented in Fig. 2a, b using a cubic radial power RBF can be advantageous since they are parameter-free. For the multiquadric, smaller shape parameter values (Fig. 2c) ϵ correspond to flatter functions while larger values (Fig. 2d) correspond to smoother functions. For Wendland's compactly supported functions, the cutoff radii parameter R specifies the range of influence of sampled point data. Smaller R values are more locally based interpolations (Fig. 2e) whereas larger R values are more global (Fig. 2f).

The most common technique used for determining the optimal RBF and associated shape parameters (if applicable) given a dataset is leave one out cross-validation (LOOCV). This technique typically uses either error estimates or stability metrics of RBF interpolants. Although these metrics are useful to study mathematical properties of the interpolation, they are insufficient in determining which RBF and their parameters are best for producing physically plausible models for many geoscience applications.

Ideally, determining the best RBF for geoscience applications requires development of application-specific metrics computed on models produced by the evaluation of RBF interpolants over a discretized domain. The purpose of model metrics is to provide comparative measures for how much better one model compares to another. Model metrics can include various geometrical measures of smoothness and shape, topological measures computed from the Morse-Smale complex (Edelsbrunner et al. 2003), volume/area of regions bounded by function value ranges, and model uncertainty as measured by entropy from ensemble of models (de la Varga and Wellmann 2016). Combining model uncertainty with application-specific model metrics capable of measuring the performance and plausibility of relevant model features is recommended.

Software Implementations

There are numerous commercial and open source software-implementing RBF interpolation and associated methodologies. However, there are no implementations which offer comprehensive support for all possible optimizations and constraint types. Furthermore, some implementations obscure access and restrict fine tuning of various RBF parameters which can be beneficial for modeling. Access to specific RBF methods and control over their parameterization may require using open source code – albeit at the expense of ease of use. Additionally, there are potential synergies with combining RBF interpolation with other geomodeling tools which enable incorporation of additional geological information provided by interpretation, knowledge, and multi-disciplinary datasets.

Limitations

Although there has been great success with using RBF interpolation for spatially dependent geoscience applications, limitations exist, particularly with respect to geological modeling of structurally complex features. In these settings, physically implausible solutions can be produced. The predominant issue is limited control over global interpolation properties. A priori knowledge quite often available describing global properties such as topology, volume/area, and structural feature thickness cannot be incorporated as constraints, the reason being many global properties can only be computed after interpolants have been fitted. In addition, there are scalability problems. While RBF-based numerical optimizations enabled up to millions of data points using interpolants of the form Eq. 1, many of extensions previously described and their kriging counterparts have yet to be accommodated by these methods.

Recent Research

Potential for modeling capacity enhancement for geoscience applications using RBF interpolation may exist in leveraging recent research from the numerical analysis community. In particular, the variably scaled kernels introduced by Bozzini et al. (2015) transform the interpolation problem to a higher dimensional space permitting locally varying shape parameters for SPD RBF functions. The concept could provide a mechanism for incorporating locally varying anisotropy. For discontinuous applications, the work on variably scaled discontinuous kernels by De Marchi et al. (2020) would be of interest.

Conclusions

RBFs remain popular in geoscience for scattered data approximation for spatially dependent applications, particularly for implicit modeling. Although there has been widespread success in their use, there are limitations depending on the application. Notwithstanding the limitations, continued development of RBF-interpolation and their novel uses provides current and future geoscience applications with increasingly effective tools.

Cross-References

- [Interpolation](#)
- [Kriging](#)
- [Three-Dimensional Geologic Modeling](#)
- [Topology in Geosciences](#)

Acknowledgments Many thanks to Eric de Kemp for valuable feedback on this entry and many fruitful discussions involving the use of RBFs for constrained 3D geological modeling. SKUA-GOCAD® software was provided through the RING (Research for Integrative Numerical Geology) research consortia and Emerson. NRCan contribution number 20200068.

Bibliography

- Bozzini M, Lenarduzzi L, Rossini M, Schaback R (2015) Interpolation with variably scaled kernels. *IMA J Numer Anal* 35:199–219
- Calcagno P, Chiles JP, Courrioux G, Guillen A (2008) Geological modelling from field data and geological knowledge Part I. Modelling method coupling 3D potential-field interpolation and geological rules. *Phys Earth Planet Inter* 171(1-4):147–157
- Carr JC, Beatson RK, Cherrie JB, Mitchell TJ, Fright WR, McCallum BC, Evans TR (2001) Reconstruction and representation of 3D objects with radial basis functions. In: *ACM SIGGRAPH 2001, Computer graphics proceedings*. ACM Press, New York, pp 67–76
- Carr JC, Beatson RK, McCallum BC, Fright WR, McLennan TJ, Mitchell TJ (2003) Smooth surface reconstruction from noisy range data. In: *Proceedings of the 1st international conference on computer graphics and interactive techniques in Australasia and South, East Asia*, p 119–ff
- Cowan EJ, Beatson RK, Fright WR, McLennan TJ, Mitchell TJ (2002) Rapid geological modelling. In: *Applied structural geology for mineral exploration and mining, international symposium*, pp 23–25
- de Kemp EA, Schetselaar EM, Hillier MJ, Lydon JW, Ransom PW (2016) Assessing the workflow for regional-scale 3D geologic modeling: an example from the Sullivan time horizon, Purcell Anticlinorium East Kootenay Region, Southeastern British Columbia. *Interpretation* 4(3):SM33–SM50
- de la Varga M, Wellmann JF (2016) Structural geologic modeling as an inference problem: a Bayesian perspective. *Interpretation* 4(3):SM1–SM16
- De Marchi S, Marchetti F, Perracchione E (2020) Jumping with variably scaled discontinuous kernels (VSDKs). *BIT Numer Math* 60: 441–463. <https://doi.org/10.1007/s10543-019-00786-z>

- Dubrulle O (1984) Comparing splines and kriging. *Comput Geosci* 10(2-3):327–338
- Duchon J (1976) Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces. *RAIRO Analyse Numérique* 10(12):5–12
- Edelsbrunner H, Harer J, Natarajan V, Pascucci V (2003) Morse-Smale complexes for piecewise linear 3-manifolds. In: *Proceedings of the 19th annual symposium on Computational Geometry (SCG '03)*. ACM, New York, pp 361–370. <https://doi.org/10.1145/777792.777846>
- Hardy RL (1990) Theory and applications of the multiquadric-biharmonic method 20 year of discovery 1968–1988. *Comput Math Appl* 19(8-9):163–208
- Hillier MJ, Schetselaar EM, de Kemp EA, Perron G (2014) Three-dimensional modelling of geological surfaces using generalized interpolation with radial basis functions. *Math Geosci* 46(8):931–953
- Lajaunie C, Courrioux G, Manuel L (1997) Foliation fields and 3D cartography in geology: principles of a method based on potential interpolation. *Math Geol* 29(4):571–584
- Lorensen WE, Cline HE (1987) Marching cubes: a high resolution 3D surface construction algorithm. *Comput Graph* 21(4):163–169
- Micchelli CA (1986) Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constr Approx* 2:11–22
- Scheuerer M, Schaback R, Schlather M (2013) Interpolation of spatial data – a stochastic or a deterministic problem? *Eur J Appl Math* 24(4):601–629
- Wendland H (2005) *Scattered data approximation*. Cambridge University Press, New York

Random Forest

Emil D. Attanasi¹ and Timothy C. Coburn²

¹US Geological Survey, Reston, VA, USA

²Department of Systems Engineering, Colorado State University, Fort Collins, CO, USA

Definition

The random forest algorithm formalized by Breiman (2001) is a supervised learning method applied to predict the class for a classification problem or the mean value of a target variable of a regression problem. The fundamental elements of the algorithm consist of “classification and regression trees” (CART) that are applicable for modeling nonlinear relationships, particularly where the prerequisites of standard parametric regression, such as explicit functional expressions, independence of regressors, homoscedasticity, and normality of errors, cannot be met. Individual “trees,” sometimes referred to as decision trees, are built through the process of recursive partitioning of the predictor variable data space leading to a schematic of “nodes” and “branches” that depicts the relationships among the variables and their hierarchical importance. This iterative process splits the data into branches and then continues splitting each partition into smaller groups as the

partitions become more refined and the data more homogeneous within the partitions. The splitting rules may be set to maximize local-node homogeneity or optimize a criterion related to the predictive value of the target (dependent) variable.

The random forest algorithm constructs an ensemble of such decision trees using a training data set. This ensemble (“forest”) is instantiated with values of predictor variables to establish the class of the target variable (in the case of classification) or the value of the target variable (in the case of regression). The target variable represents the quantity to be determined and the predictors are variables whose known values are thought to be related to the target. The use of an ensemble approach, rather than a single tree, improves the predictive accuracy outside the training data set due to averaging the individual tree predictions and assures stability and robustness of the predictions because the ensemble encompasses more possibilities in the training data set (Hastie et al. 2009). In supervised learning, training data are the data used to train an algorithm to predict the target variable. If a separate validation data set is required, the training data may be a random subsample of the available data.

Description of Random Forest Algorithm

The random forest algorithm constructs an ensemble of trees in the following way. Each individual decision tree is developed from a bootstrap sample randomly selected with replacement from the training data set and subsequently built by recursively partitioning the predictor space. The recursive partitions are constructed by sampling a subset of the available predictors and then optimal splits in the predictor data space are identified that provide the greatest reductions in node impurity. The node impurity is a measure of the homogeneity of the target values at the node. For classification trees the standard algorithm minimizes Gini impurity, which is interpreted as the probability of misclassifying an observation. For regression trees, partitions are made to minimize the error variance of the target variables associated with the predictors. Bootstrap sampling of the training data to construct each tree, plus sampling the set of predictor variables prior to each partition, is intended to quantify uncertainty and reduce the prediction variance at the expense of introducing slight bias.

In the case of classification, the random forest ensemble prediction consists of the “majority vote” among all class assignments determined for the individual trees; and in the regression case, prediction consists of the mean of the predicted values of the target variable across all individual trees. For each tree, the part of the training data that is not included in the bootstrap sample used to construct the tree is known as the out-of-bag (OOB) sample. These data are used by the

algorithm to incorporate a validation step in the training procedure (Breiman 2001). The OOB error estimate is comparable to the error calculated with standard cross-validation procedures. The OOB error is an estimate of the random forest's "generalization error," which is defined as the prediction error expected for samples outside the training set.

Hyperparameters are used to control the learning process in the machine learning algorithms so that the predictive model minimizes the estimated generalization error. For random forests, they include the number of trees, the number of predictors sampled, the bootstrap sampling fraction of the training data, and various controls on tree complexity, such as the minimum number of observations to be used at a terminal node, the required observations for a partition (node size), and the maximum number of terminal nodes. A smaller subset of the predictors leads to lower correlation among trees. A smaller minimum node size leads to deeper trees with more splits. A smaller training sample fraction also leads to reduced correspondence among trees in the forest.

Model Interpretation

Model interpretation centers on the relative importance of the predictor variables and the responses of the target variable to incremental predictor variable changes. Two metrics for assessing the relative impact of predictors are permutation importance and Gini importance. Permutation importance ranks the predictor variables based on degradation of the model's predictive performance when their values in the OOB sample are randomly and successively permuted. Predictor variable importance rank is directly proportional to the degree of the loss in predictive accuracy for the OOB samples when predictor values are successively permuted. Alternatively, Gini importance counts the number of times a predictor is used in partitioning the predictor space weighted by the fraction of samples it splits (which approximates the probability of reaching that node) averaged over all trees of the ensemble. Gini importance rank corresponds to the relative magnitude of the calculated Gini importance. Because this traditional computational scheme for computing Gini importance produced biased results in favor of predictors with many possible split points, it was modified by Nembrini et al. (2018) and the revised version was shown to be unbiased. Other interpretive tools include permutation tests (Altmann et al. 2010) used to calculate p-values associated with the predictor variables. Innovative methods to compute prediction intervals (Zhang et al. 2019) and ways to generalize the algorithm to provide rigorous causal inference interpretations to the relationship between predictors and the target variable (Athey et al. 2019) are areas of recent research.

Because of their breadth, tree-based predictive models can capture the interaction effects of predictors. While identifying

and recognizing the nature of such interactions can be difficult, the use of various visual displays can be helpful. For example, the marginal effect that one predictor or two features have on the target variable can be shown by partial dependence plots (PDPs). By averaging predicted target values across trees, when marginal changes are made in one predictor, and other predictors are set at their average values, PDPs show whether the relationship between the target and a feature is linear, monotonic (i.e., entirely nonincreasing or non-decreasing), or more intricate. Another visual diagnostic is the individual conditional expectation plot (ICE). Whereas the PDPs show the predicted target values averaged across all trees, the ICE plot displays the individual tree values of the target relative to marginal changes in a predictor, thereby capturing the diversity of the predicted responses. Additional methods to identify, analyze, and interpret effects of predictor interactions are being pursued by researchers in a variety of other sciences, particularly medicine.

Analytical advantages include the nonparametric nature of the algorithm, its superior predictive performance, its capability to determine variable importance, built-in procedures to facilitate model validation using OOB observations, and the ability to compute variable proximities that can be used for data clustering to further improve prediction performance (Tyralis et al. 2019). Most implementations of the algorithm also provide imputation options to allow use of incomplete observations.

The analyst should be aware of the following limitations when applying the original random forest algorithm. First, the results are not as readily interpreted as those derived from a single regression or classification tree. Second, the reliability of the variable importance metrics is affected by correlation among predictors. Third, predictions cannot be reliably extrapolated beyond the range of the training data. Fourth, modifications are required to adequately model data sets in which the number of observations of the response variable belonging to one class is substantially different from the numbers of observations associated with the other classes (e.g., in the case of rare events such as some mineral occurrences). Finally, it is worth noting that the original algorithm is not suitable for causal inference. Further, variations of the random forest algorithm (e.g., extremely randomized forests or conditional inference random forests) do not consistently provide better performance (Tyralis et al. 2019).

Variations and Software

Tyralis et al. (2019) have published one of the most extensive reviews of the literature relating to random forests, all in the context of water and hydrogeological research. They list and reference 33 variations of the basic algorithm, with extremely randomized trees, quantile regression trees, and trees

associated with conditional inference being among those most frequently applied. Additional references for variations that lead to Bayesian, neural, dynamic, and causal random forests are also provided.

Computer software for modeling random forests is readily available in a number of open source domains and in many commercial packages. Tyrallis et al. (2019) list 55 packages related to random forests available in the R language, which is a free software environment for statistical computing and graphics. Python, also an open source language, has a random forest implementation in the Scikit-Learn package, and commercial software providers, such as Mathworks® and SAS®, commonly have packages for computing random forests (Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the US Government.).

Earth Science Applications

Numerous applications of the random forest algorithm and its variations have been published in the quantitative earth science literature. For example, applications of the algorithm to remote sensing are reviewed by Belgiu and Drăguț (2016). Another broad class of applications relate to predictive mapping at the regional level utilizing the classification function of the random forest algorithm. For example, mapping situations that use data generated from satellite images combined with various types of geophysical measurements (either from airborne sensors or field samples) can be used to classify rock lithology. In this case, the multiple classification scheme works on a pixel basis to generate the map image. Similar approaches have been applied to data on physiography, rainfall, and soil type to map areas having a propensity for landslides; and groundwater resources have been mapped using physiographic data, satellite imagery, and observed occurrences of springs (Tyrallis et al. 2019 and references therein). Other mapping applications include those involving land use and wetlands (Belgiu and Drăguț 2016).

During the last 50 years, geologists have developed an expert knowledge base identifying evidentiary data of individual classes of mineral occurrences. This information, along with spatial geologic data and occurrence/non-occurrence of minerals, can be addressed with the random forest algorithm to generate favorability maps for use in mineral prospecting (e.g., Carranza and Laborte 2016). Such applications, which include copper, gold, and polymetallic minerals, require adjustments to be made to the input data because the identified potentially marketable concentrations of mineral occurrences are quite rare.

Laboratory applications of the random forest algorithm include identification and location of organic matter and

pores in scanning electric microscope (SEM) images of organic rich shales. Similar image processing techniques can be used to analyze petrographic thin sections to predict (classify) rock type and mineral composition (Xie et al. 2020).

In the oil and gas industry, the random forest algorithm has been applied to well log images and digitally processed well logs to identify lithology and rock properties that become the basis for calculating oil and gas reservoir parameters. The algorithm is routinely used in field operations to classify rock lithology from well measurements taken while drilling in real time to identify the target formation as well as to predict rates of drilling penetration (Xie et al. 2020). Other uses include prediction of individual well flow rates and expected ultimate oil recovery by well. These latter applications of the regression tree function of random forests naturally lead to identifying the drivers of well productivity utilizing the variable importance function of random forests.

As well noted in Tyrallis et al. (2019), water science applications largely involve categorization of stream flow volumes and water quality features, as well as the analysis of predictor importance. Uses of the random forest algorithm to address environmental quality and ecological viability include mapping contaminants in estuarine systems, modeling nitrate concentrations in water wells, and prediction of algae blooms in tropical lakes. In these applications the algorithm is used to identify predictors that drive the models of such systems. In the atmospheric sciences, the use of random forests has been shown to improve severe weather forecasts when 2 and 3 days out from the event.

Summary

The random forest algorithm is broadly employed across multiple disciplines, with numerous variations having been developed for a wide variety of contexts and scenarios. Its use has become almost ubiquitous in the earth sciences, with applications ranging from research and planning to optimization of field and industrial operations, resource exploration and development, environmental monitoring, and beyond. Given the ready availability of software, execution of the basic algorithm is relatively straightforward, with computations and setup requiring minimal assumptions. Applications often encompass large data bases and thousands of predictors, and its superior predictive performance is well established in the earth sciences as well as in other fields such as the medical sciences. While not without limitations, the methodology underlying the algorithm is still an active field of research for generalization and statistical interpretation (Zhang et al. 2019; Athey et al. 2019).

Cross-References

- Artificial Intelligence in the Earth Sciences
- Decision Tree
- Machine Learning
- Predictive Geologic Mapping and Mineral Exploration

Bibliography

- Altmann A, Tolosi L, Sander O, Lengauer T (2010) Permutation importance: a corrected feature importance measure. *Bioinformatics* 26(10):1340–1347. <https://doi.org/10.1093/bioinformatics/btq134>
- Athey S, Tibshirani J, Wager S (2019) Generalized random forests. *Ann Stat* 47:1148–1178
- Belgiu M, Drăguț L (2016) Random forest in remote sensing: a review of applications and future directions. *ISPRS J Photogramm Remote Sens* 114, 24:–31. <https://doi.org/10.1016/j.isprsjprs.2016.01.011>
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32. <https://doi.org/10.1023/A:1010933404324>
- Carranza EJM, Laborte AG (2016) Data-driven predictive modeling of mineral prospectivity using random forests: a case study in Catanduanes Island (Philippines). *Nat Resour Res* 25(1):35–50
- Hastie T, Tibshirani R, Friedman J (2009) *Elements of statistical learning – data mining, inference, and prediction*, 2nd edn. Springer, Berlin
- Nembrini S, Keonig I, Wright M (2018) The revival of the Gini importance? *Bioinformatics* 34(21). <https://doi.org/10.1093/bioinformatics/bty373>
- Tyralis H, Papacharalampous G, Langousis A (2019) A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water* 11(5):910. <https://doi.org/10.3390/w11050910>
- Xie Y, Zhu C, Hu R, Zhu Z (2020) Lithology Identification with Extremely Randomized Trees. *Math Geosciences*, published online 12 August 2020. <https://doi.org/10.1007/s11004-020-09885-y>
- Zhang H, Zimmerman J, Nettleton D, Nordman DJ (2019) Random Forest prediction intervals. *Am Stat* 74(4):392–406. <https://doi.org/10.1080/00031305.2019.1585288>

Random Function

J. Jaime Gómez-Hernández
Research Institute of Water and Environmental Engineering,
Universitat Politècnica de València, Valencia, Spain

Synonyms

Stochastic process

Definition

A random function $Z(\mathbf{x})$, also known as a stochastic process, can be defined as a rule that assigns a realization to the outcome of an experiment

$$Z(\mathbf{x}) \sim \{z(\mathbf{x}, \theta)\}, \forall \theta \in \Omega, \mathbf{x} \in \mathcal{D}, \quad (1)$$

where $\mathcal{D}$ is the domain within which the realization is defined (it could be finite or infinite, discrete or continuous); in geosciences, $\mathcal{D}$ will be the one-, two-, or three-dimensional portion of the space where the variable of interest is being studied, or, in many occasions, the point set of one-, two-, or three-dimensional locations centered at the voxels discretizing the study area; Ω is the sample space containing all outcomes of the experiments; $\mathbf{x}$ is a point in the domain $\mathcal{D}$; and θ an outcome of the experiment. For the reminder and without loss of generality, $\mathbf{x}$ will be regarded as a spatial coordinate.

Realization Versus Random Variable

A random function as defined above can be analyzed also in terms of random variables. When the experiment outcome is fixed at θ_0 , the function $z(\mathbf{x}, \theta_0)$ is a realization of the random function within the study area. But when the location is fixed at $\mathbf{x}_0$, the function $z(\mathbf{x}_0, \theta)$ corresponds to a random variable. To alleviate the notation, θ is dropped hereafter, and then, a lower case z will represent realizations (and the values they can take) and an upper case Z a random variable or the random function itself.

The random function can then be characterized by all the n -variate distributions

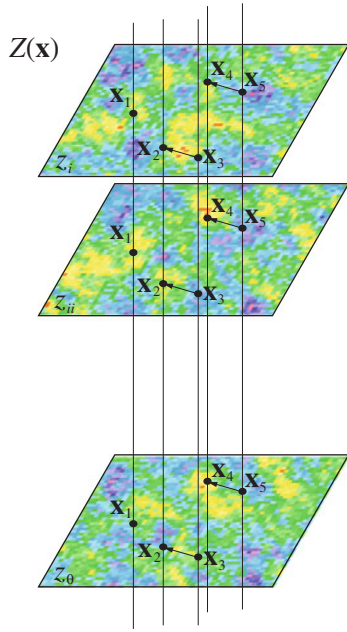
$$F(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n; z_1, z_2, \dots, z_n) = \text{Prob}(Z(\mathbf{x}_1) \leq z_1, Z(\mathbf{x}_2) \leq z_2, \dots, Z(\mathbf{x}_n) \leq z_n), \quad (2)$$

which are defined for any set of locations $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, and any set of values $\{z_1, z_2, \dots, z_n\}$.

To better understand what a random function is, it is convenient to consider this duality between realizations and random variables (Fig. 1). When considered as a collection of realizations, one needs to know which is the rule that associates a realization to the outcome of an experiment. When considered as a collection of random variables, one needs to know all its multivariate distributions.

The Random Function Model in Geosciences

Geoscience variables generally display patterns of spatial variability that cannot be modeled with deterministic models, and, at the same time, they display spatial patterns that denote that their distribution in space is not totally random. The random function model is a model suited for the modeling of these variables. A decision is taken that the specific, and unknown save for a few measurement locations, spatial distribution of the variable of interest corresponds with one of



Random Function, Fig. 1 Conceptualization of a random function as a collection of realizations, $Z(\mathbf{x}) \sim \{z_i, z_{ii}, \dots, z_0\}$. At any given location $\mathbf{x}$, the collection of z values through the realizations corresponds to a random variable

the realizations of a random function. Then, a random function is chosen and it is analyzed to build an understanding about the values of the variable at unsampled locations. The difficulty in adopting this model strives in how the random function is chosen, and then, how it is characterized, whether through the rule that associates realizations to experiment outcomes or through the multivariate distribution.

Some Random Function Summary Statistics

There are some summary statistics commonly used in the characterization of a random function. The first and second order ones are given next.

Expected Value

The expected value of a random function is a function of $\mathbf{x}$ with the expected values of each random variable

$$E\{Z(\mathbf{x})\} = \mu(\mathbf{x}). \quad (3)$$

Variance

Likewise, the variance of a random function is a function of $\mathbf{x}$ with the variance of each random variable

$$\text{Var}\{Z(\mathbf{x})\} = \sigma^2(\mathbf{x}) = E\{(Z(\mathbf{x}) - \mu(\mathbf{x}))^2\}. \quad (4)$$

Covariance

The covariance is computed between two locations and it corresponds to the covariance between the two random variables at those locations

$$C(\mathbf{x}_i, \mathbf{x}_j) = E\{(Z(\mathbf{x}_i) - \mu(\mathbf{x}_i))(Z(\mathbf{x}_j) - \mu(\mathbf{x}_j))\}, \forall \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}. \quad (5)$$

Variogram

Like the covariance, the variogram is also computed between two locations

$$\gamma(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{2} E\{(Z(\mathbf{x}_i) - Z(\mathbf{x}_j))^2\}, \forall \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}. \quad (6)$$

Stationary and Ergodicity

Refer back to Fig. 1, adopting a random function model implies the acceptance that one of the realizations $\{z_i, z_{ii}, \dots, z_0\}$ is the reality. Some data may have been sampled, but it is evident that, from these data, it is impossible to derive any of the common statistics described in the previous section, much less any of the n -variate distributions in Eq. (2). At $\mathbf{x}_1$, there is only one sample of the random variable $Z(\mathbf{x}_1)$, from which it is impossible to derive $\mu(\mathbf{x}_1)$, and the pair of values $z(\mathbf{x}_2)$ and $z(\mathbf{x}_3)$ are not sufficient to compute the covariance $C(\mathbf{x}_2, \mathbf{x}_3)$ or the variogram $\gamma(\mathbf{x}_2, \mathbf{x}_3)$. For these reasons, the random function models used in the geosciences are not as generic as their definition implies. It is typical to use random functions that are both stationary and ergodic.

Stationarity

A random function is said to be stationary of the first order if its expected value is constant

$$E\{Z(\mathbf{x})\} = \mu(\mathbf{x}) = \mu. \quad (7)$$

A random function is said to be stationary of the second order if its covariance depends only on the separation vector between the two locations and not in the actual locations. In Fig. 1, the pairs of points $z(\mathbf{x}_2)$ and $z(\mathbf{x}_3)$, and $z(\mathbf{x}_4)$ and $z(\mathbf{x}_5)$ are separated by the same vector; if the random function is stationary of the second order, the covariance would be the same for both pairs

$$C(\mathbf{x}_i, \mathbf{x}_j) = C(\mathbf{x}_i - \mathbf{x}_j). \quad (8)$$

When the separation vector is the null vector, the covariance becomes the variance of the random variable and, as a consequence of the previous equation, it will be constant

$$C(0) = \sigma^2. \quad (9)$$

Likewise, the variogram of a second-order stationary random function depends only on the separation vector

$$\gamma(\mathbf{x}_i, \mathbf{x}_j) = \gamma(\mathbf{x}_i - \mathbf{x}_j), \quad (10)$$

and the following two relationships hold

$$C(\mathbf{x}_i - \mathbf{x}_j) = \sigma^2 - \gamma(\mathbf{x}_i - \mathbf{x}_j) \quad (11)$$

$$\gamma(\mathbf{x}_i - \mathbf{x}_j) = \sigma^2 - C(\mathbf{x}_i - \mathbf{x}_j) \quad (12)$$

Ergodicity

A stationary random function defined on an infinite domain $\mathcal{D}$ is said to be ergodic on the mean when the constant expected value of all the random variables coincides with the mean computed on any realization

$$E\{Z(\mathbf{x}_j)\} = \int_{\mathcal{D}} z_{\theta}(\mathbf{x}) d\mathbf{x} = \mu, \quad \forall \mathbf{x}_j \in \mathcal{D}, \forall \theta \in \Omega. \quad (13)$$

Likewise, it is ergodic on the covariance when the expected value computed through the realizations can be replaced by the integral on any realization

$$\begin{aligned} C(\mathbf{x}_i - \mathbf{x}_j) &= C(\mathbf{h}) = E\{(Z(\mathbf{x}_i) - \mu)(Z(\mathbf{x}_i + \mathbf{h}) - \mu)\} \\ &= \int_{\mathcal{D}} (z_{\theta}(\mathbf{x}) - \mu)(z_{\theta}(\mathbf{x} + \mathbf{h}) - \mu) d\mathbf{x}, \\ &\quad \forall \mathbf{x}_i - \mathbf{x}_j \in \mathcal{D}, \forall \theta \in \Omega \end{aligned} \quad (14)$$

The Decision of Stationarity and Ergodicity

Given that, in practice, only limited information about one of the realizations of the random function is available, it seems convenient to use stationary and ergodic random function models since, in that case, the mean and covariance of the random function could be replaced by approximate estimates of the spatial integrals computed from the sample data. But the use of a stationary and ergodic covariance is only a matter of convenience, nothing that can be proved or disproved on the basis of sparse information from a given realization.

On certain occasions, the modeler may choose a non-stationary random function, for instance one with a space-varying expected value, but such a decision must be taken from ancillary data, like geological or process information but never on a statistical analysis of a single sample from multiple random variables.

Convenient Random Function Models

There are some random function models that are commonly used in mathematical geosciences for their convenience in specifying all n -variate distributions in Eq. (2).

The Multi-Gaussian Random Function Model

The multi-Gaussian random function model is probably the most commonly used because its multivariate distribution only depends on the mean and the covariance; all remaining higher-order moments are functions of these two (Anderson

1984). A stationary and ergodic multi-Gaussian random function has the advantage that mean and covariance can be estimated from the mean and covariance derived from the data using Eqs. (13) and (14).

The Multi- φ -Gaussian Random Function Model

The properties of the multi-Gaussian random function model extend to any monotonic transformation $Y(\mathbf{x}) = \varphi(Z(\mathbf{x}))$. The new random function $Y(\mathbf{x})$ is the collection of realizations $\{y_{\theta} = \varphi(z_{\theta})\}$ and has the same property of being fully characterized just by the mean and the covariance, as long as function φ is monotonic. A typical example is the multi-log-Gaussian distribution used frequently to model permeabilities.

The Indicator Random Function

One of the problems associated with the adoption of any Gaussian-related random function has to do with its convenience of being fully characterized by the mean and the covariance. The full dependency of all higher-order moments on the first two moments makes it impossible to control the higher-order moments, and, particularly, the multi-Gaussian models are characterized by little spatial connectivity for the extreme values (Gómez-Hernández and Wen 1998). Extreme-value continuity is fundamental in geological settings; in petroleum engineering, shale barriers or flow channels can depict much higher continuity in space than intermediate values, and such a feature cannot be reproduced by a Gaussian random function.

The continuity of extreme values is explicitly handled by the indicator formalism. Consider that the range of z values is binned into K classes and then K indicator variables are derived for any z value

$$i_k(\mathbf{x}; z) = \begin{cases} 1; & \text{if } z \in [z_{k1}, z_{k2}] \\ 0; & \text{if not} \end{cases}, k = 1, \dots, K \quad (15)$$

where $[z_{k1}, z_{k2}]$ are the limits of class k . As a result, now there are K random functions $I_k(\mathbf{x}; z)$ the continuity of which can be controlled independently of each other and can be used to build non-Gaussian realizations of $Z(\mathbf{x})$.

Algorithmically Defined Random Functions

According to Eq. (1), the set of realizations generated by current stochastic simulation computer codes define a random function, as long as an unequivocal rule identifying each realization is defined. Such a rule is the algorithm behind the computed code. The experiment consists of drawing random seed number to prime the algorithm. To each seed corresponds a unique realization.

In the beginning, computer simulation codes were built to generate realizations drawn from a specific random function model. This is the case for any sequential Gaussian simulation code (see, for instance, Deutsch and Journel 1992; Gómez-Hernández and Journel 1993; Gómez-Hernández and Srivastava 2021), which, by construction, will generate realizations drawn from a multi-Gaussian random function. Likewise, any sequential indicator simulation code (Gómez-Hernández and Srivastava 1990; Deutsch and Journel 1992) generates realizations from an indicator random function. But soon after these codes were of general use, practitioners demanded codes that generated realizations with other characteristics. For example, there was an interest in generating realizations from a stationary random function with univariate distributions far from a Gaussian bell and with a given covariance. This could not be achieved with any Gaussian simulator. To solve this problem, the direct sequential simulation was proposed (Soares 2001), this algorithm generates realizations with the desired univariate distribution and covariance; however, the other moments or n -variate distributions are unidentified, but they could be derived from the statistical analysis of a sufficiently large collection of realizations.

The concept of algorithmically defined random function was born and the need to have the analytical expression for the n -variate distributions of Eq. (2) disappears. The most recent multipoint geostatistics (Mariethoz and Caers 2014; Strebelle 2002) would be a paradigmatic example of these kinds of random functions defined by a collection of realizations, which are built on the basis of a training image and an algorithm to extract high-order statistics from the training image to infuse them onto the realizations.

Summary and Conclusions

The random function model is often used in the geosciences to model the spatial variability of properties that display a heterogeneity that cannot be described deterministically nor is completely random. From a practitioner point of view, the use of such a model is a decision that cannot be proved or disproved from the data. The random function model is the foundation of geostatistics and is behind standard algorithms such as kriging or sequential simulation.

Cross-References

- [Kriging](#)
- [Monte Carlo Method](#)
- [Multivariate Analysis](#)
- [Random Variable](#)
- [Realizations](#)
- [Simulation](#)
- [Spatial Statistics](#)

Bibliography

- Anderson TW (1984) Multivariate statistical analysis. Wiley, New York
- Deutsch CV, Journel AG (1992) GSLIB, geostatistical software library and user's guide. Oxford University Press, New York
- Gómez-Hernández JJ, Journel AG (1993) Joint sequential simulation of multi-Gaussian fields. In: Soares A (ed) Geostatistics Tróia '92, vol 1. Kluwer Academic, Dordrecht, pp 85–94
- Gómez-Hernández JJ, Srivastava RM (1990) ISIM3D: an ANSI-C three dimensional multiple indicator conditional simulation program. *Comput Geosci* 16(4):395–440
- Gómez-Hernández JJ, Srivastava RM (2021) One step at a time: the origins of sequential simulation and beyond. *Math Geosci* 53(2): 193–209
- Gómez-Hernández JJ, Wen XH (1998) To be or not to be multiGaussian: a reflection in stochastic hydrogeology. *Adv Water Resour* 21(1): 47–61
- Mariethoz G, Caers J (2014) Multiple-point geostatistics: stochastic modeling with training images. Wiley
- Soares A (2001) Direct sequential simulation and cosimulation. *Math Geol* 33(8):911–926
- Strebelle S (2002) Conditional simulation of complex geological structures using multiple-point statistics. *Math Geol* 34(1):1–21

Random Variable

Ricardo A. Olea
Geology, Energy and Minerals Science Center,
U.S. Geological Survey, Reston, VA, USA

Background

When Galileo dropped objects from the leaning tower of Pisa, he noted that, for a given height and meteorological condition, the falling time was the same. This is one example of a situation in which the result of the experiment is unique and predictable. These types of observations are called deterministic experiments.

Now flip a coin. The result is not unique anymore. More than one outcome is possible; the result is now uncertain. These experiments go by the name of random experiments.

Statistics uses the concept of probability to measure the chances of different outcomes in a random experiment. By convention, probabilities are always positive numbers including zero. There are three fundamental probability axioms (Hogg et al. 2018):

- The probability of an impossible outcome is 0.
- The maximum value of 1 denotes outcomes absolutely certain to occur.
- If two outcomes of the same experiment cannot occur simultaneously, their probabilities are additive.

Probabilities can also be multiplied by a factor of 100 and reported as percentages. A set containing all possible outcomes is called sample space.

Definition

A random variable is a function that assigns a value in a sample space to an element of an arbitrary set (James 1992; Pawlowsky-Glahn et al. 2015). It is a model for a random experiment: the arbitrary set is an abstraction of the experimental conditions, the values taken by the random variable are in the sample space, and the function itself models the assignment of outcomes, thus also describing its frequency of appearance. In simpler terms, for the purpose of this presentation, a random variable is a function that assigns to each of the outcomes of a random experiment a value with a certain probability. A random variable also goes by stochastic variable and aleatory variable. Random variables are usually annotated as Roman capital letters, such as X or Y .

The characteristics of the sample space may define a variety of random variables. When the sample space comprises the real numbers and the scale is assumed absolute, differences are computed by ordinary subtraction, and averages are computed using the ordinary sum. However, there is a need of alternative sample spaces. For instance, a set of functions of time; a set of bounded regions on a plane; the points on a simplex; points on a sphere; positive real numbers. They characterize the respective random variables as stochastic processes, random sets, random compositions, random

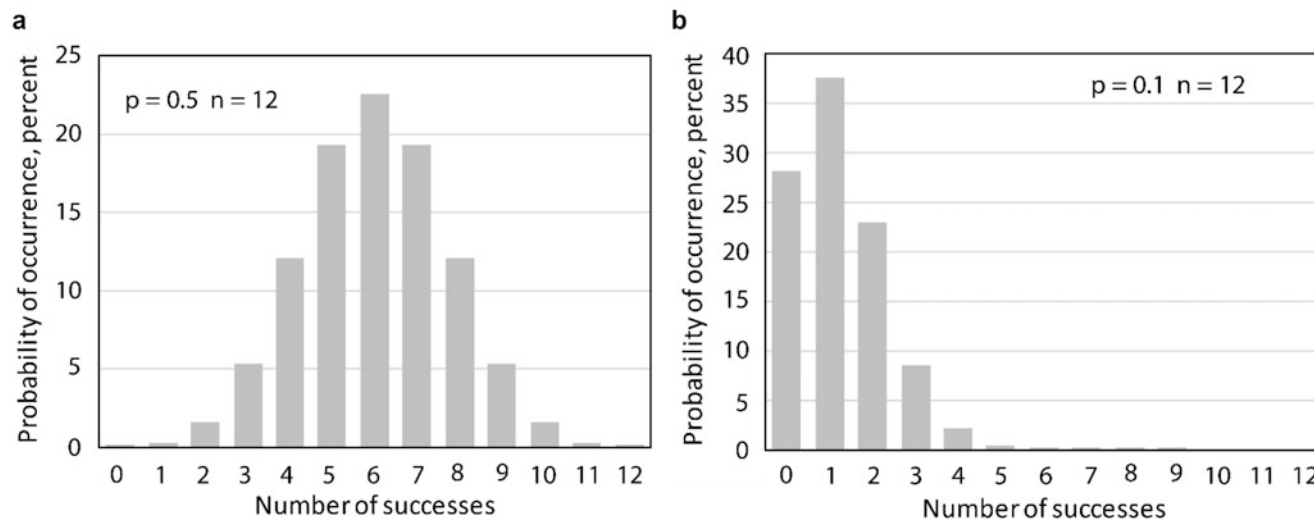
directions, and ratio scale variables. These types of random variables deserve separate development of their theories.

A random variate or realization is the result of randomly drawing one outcome from a random variable.

Discrete and Continuous Random Variables

In order to fully characterize a random variable, a description of the probability of the outcomes is necessary. The nature of the sample space determines the kind of description. In terms of the type of the sample space, the random variables can be discrete or continuous. If one can count the number of possible outcomes, then the random variable is discrete. Outcomes from game chances such as flipping coins, playing cards, or casting dice are examples of discrete random variable sample spaces. Fig. 1 shows two examples that provide chances in random experiments with only two possible outcomes per trial. The case when the chance of each outcome is the same can be associated with the cumulative number of heads expected after flipping a coin multiple times (Fig. 1a). When the value of the parameter is 0.1, it gives the expected number of producing wells when drilling wildcats in different plays (Fig. 1b). In the discrete case, a probability mass function assigns the probabilities to each outcome.

A continuous sample space has different mathematical characteristics. For this presentation, we center the attention in random variables whose sample space is identified with the real numbers. Commonly, the description of the probability in this continuous case is determined based on a probability



Random Variable, Fig. 1 Examples of discrete random variables when repeating a random experiment with only two possible outcomes 12 times: (a) number of heads or tails in flipping a fair coin; (b) number

of new reservoirs successfully discovered when drilling 12 wildcats at different plays when considering that the probability of success per well is 10%

density function, $f(x)$. If X is a univariate random variable, the probability of X to be in an interval (a, b) is given by

$$\Pr[a \leq X \leq b] = \int_a^b f(x) dx. \quad (1)$$

With this kind of definition, a single point is always assigned a null probability. It should be remarked that the probability density function does not always exist but when it does, it is always nonnegative. When the limits of integration in Eq. 1 are $(-\infty, \infty)$, the result is the area under the curve, which must be 1 to agree with the axioms in the Background section.

Analytical and Numerical Random Variables

The results in Fig. 1 can be obtained by actually running experiments or from mathematical analysis. In the latter case, the result is a probability mass function that in this case goes by the name of binomial distributions (Forbes et al. 2011). The binomial distribution is an example of a discrete analytical probabilistic model:

$$f_b(x) = \frac{n!}{x! \cdot (n-x)!} \cdot p^x \cdot (1-p)^{n-x}, \quad x = 0, 1, 2, \dots, n; \quad 0 < p < 1, \quad (2)$$

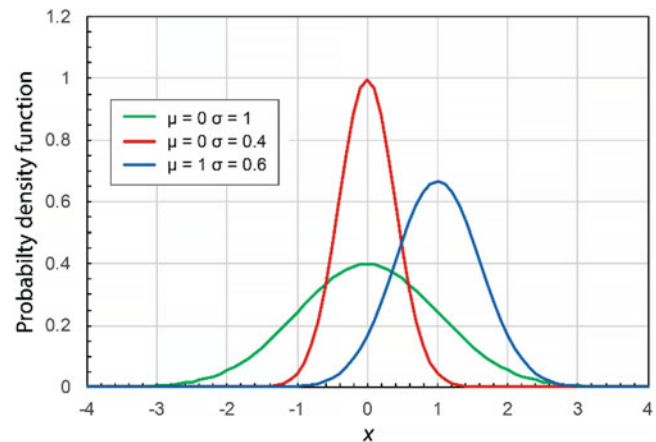
where n is a positive integer.

The normal or Gaussian distribution is one of the most widely used continuous analytical models because many attributes either follow or approximate a normal distribution (Papoulis 2002):

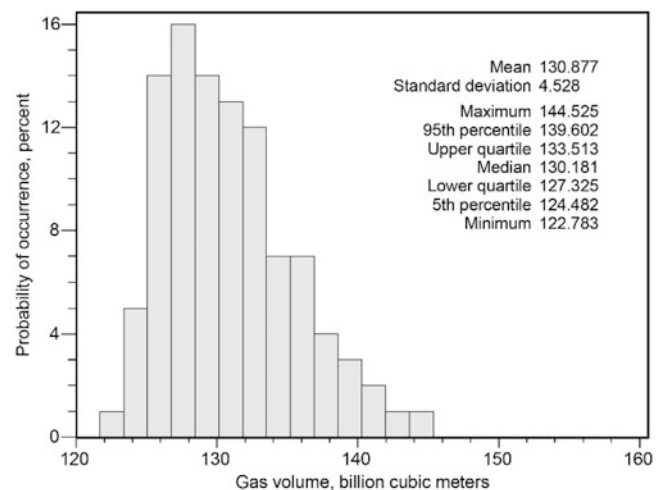
$$f_N(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \exp \left[-\frac{1}{2} \cdot \left(\frac{x - \mu}{\sigma} \right)^2 \right], \quad 0 < \sigma, \quad (3)$$

where μ and σ are two parameters allowing the same expression to take different shapes, such as those in Fig. 2. The distribution is always symmetrical with respect to parameter μ , which coincides with the mean of the distribution. Parameter σ controls the spread of the distribution and coincides with the value of its standard deviation. The domain of the normal distribution is $(-\infty, \infty)$, which sometimes is a problem in the geosciences because no attribute takes infinite values, and some do not take negative values.

The advent of computers is increasingly popularizing the generation of random variables by general purpose methods such as the Monte Carlo method or the bootstrap method. Figure 3 is an example of a repeated realization of a random variable – a random sample for short – describing the volume

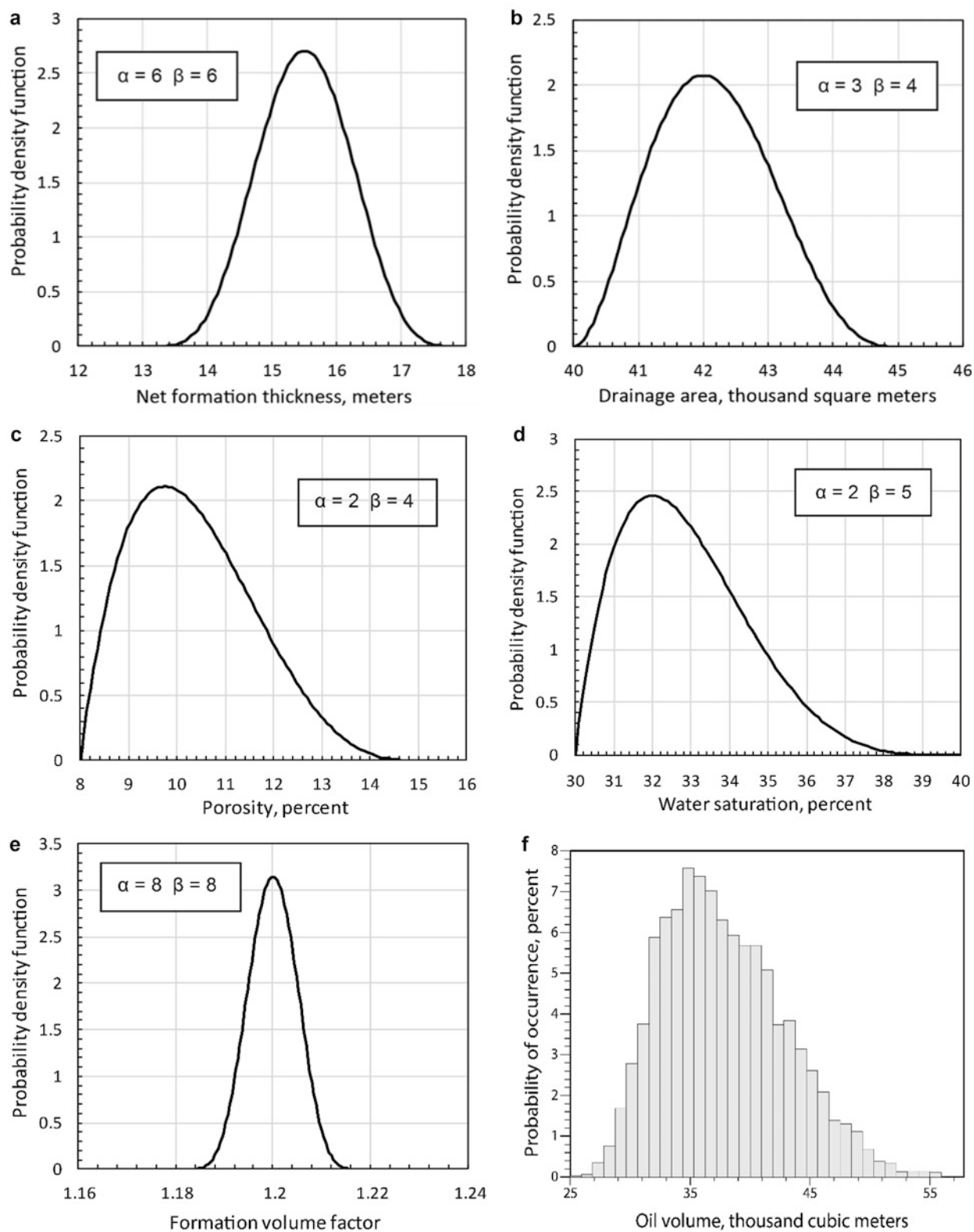


Random Variable, Fig. 2 Examples of normal distributions. Parameter μ determines the center of the distribution, and σ the spread and height of the peak



Random Variable, Fig. 3 Example of a numerical random variable showing all possible values of the true gas volume in a gas reservoir

of natural gas to be produced from a petroleum reservoir. In this case, the modeling was based on estimated ultimate recovery of natural gas from several wells (Olea et al. 2010). In this situation, the outcomes are all the possible values that the true volume could have. Note that the possible values are not equally likely. Relative to the coin example in Fig. 1a, the interpretation of the random variable is different now. By no means is the random variable implying that the gas volume in the one reservoir can take different values. In this case, there is just one unique true value, say, 125.2 billion cubic feet, but that value is unknown. The whole purpose of using a random variable is to learn what to expect about the magnitude of the true value.



Random Variable, Fig. 4 Example of numerical approximation of oil volume random variable: (a–e) input random variables, all following shifted and scaled beta distributions; (f) resulting oil volume using 5000 draws per attribute

Combination of Random Variables

Even in the simplest case of a real sample space, the arithmetic of random variables can be complex, so the analytical results are quite limited. The simplest cases for operation, for instance sum or product of real random variables, are those in which the random variables involved are independent. Two random variables are independent when their joint probability density function is the product of their individual probability density functions. This means that the result of one of the random variables does not influence the realizations of the other one. A well-known classical example is that of two independent normal random variables X and Y . In such a case, their sum follows another normal distribution with a mean equal to the sum of the means, and the variance is equal to the sum of the variance of X plus the variance of Y .

The advent of computers has radically increased the capabilities to work with combinations of random variables. Today, the Monte Carlo methods can be used to consider unlimited combinations of random variables. For example, suppose one were interested in a probabilistic assessment of the unknown volume of oil in place within the drainage area of an oil well. Let H be the random variable representing the net formation thickness, A the drainage area, P the porosity, W the water saturation, and B the conversion factor to standard conditions. The volume V of the oil in place is given by (Terry and Rogers 2014):

$$V = 0.0001 \cdot \frac{H \cdot A \cdot P \cdot (100 - W)}{B}. \quad (4)$$

This is an expression complex enough not able to be solved analytically regardless of the form of the five random variables. However, if the modeler can provide distributions for all five random variables, then it is possible to obtain a numerical distribution for the random variable V by repeatedly drawing random variates from those distributions by applying the Monte Carlo method. At least 1000 iterations are recommended for stable results. In the example, for the first set of draws the values were 16.353 m (H), 43,038 m² (A), 9.912% (P), 31.149% (W), and 1.2 (B). The oil volume for those values is 40.026 thousand m³ (V). The result of performing the draws and calculations for another 4999 times are presented in Fig. 4. Both α and β are shape parameters of the beta distribution whose domain can be shifted and linearly scaled to vary within any finite boundaries (a, b) (Olea 2011):

$$f_{\beta}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1} du}, \quad 0 < x < 1; \quad \alpha, \beta > 0. \quad (5)$$

Summary and Conclusions

A random variable is a convenient concept to link outcomes of random experiments to elements in another set. Random variables play a fundamental role in probabilistic modeling of uncertain attributes, from simple experiments like flipping coins, to more complex computer-aided calculations, such as determining the volume of crude oil resources in a reservoir.

Cross-References

- [Bootstrap](#)
- [Compositional Data](#)
- [Monte Carlo Method](#)
- [Probability Density Function](#)
- [Univariate](#)

Bibliography

- Forbes C, Evans M, Hastings N, Peacock B (2011) Statistical distributions, 4th edn. Wiley, 321 pp
- Hogg RV, McKean JW, Craig AT (2018) Introduction to mathematical statistics, 8th edn. Pearson, London, 768 pp
- James RC (1992) Mathematics dictionary, 5th edn. Springer, 560 pp
- Olea RA (2011) On the use of the beta distribution in probabilistic resource assessments. *Nat Resour Res* 20(4):377–388
- Olea RA, Cook TA, Coleman JL (2010) Methodology for the assessment of unconventional (continuous) resources with an application to the Greater Natural Buttes gas field, Utah. *Nat Resour Res* 19(4): 237–251
- Papoulis A (2002) Probability, random variables and stochastic processes, 4th edn. McGraw Hill Europe, 852 pp
- Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015) Modeling and analysis of compositional data. Wiley, 247 pp
- Terry RE, Rogers JB (2014) Applied petroleum reservoir engineering, 3rd edn. Pearson, 503 pp

Rank Score Test

Alejandro C. Frery
School of Mathematics and Statistics, Victoria University of Wellington, Wellington, New Zealand

Synonyms

Rank test

Definition

Rank score tests use the index of sorted observations to make inference, rather than the observations themselves. There are several rank tests in the literature. They are all based on the index which sorts the observations, rather than on the values. With this approach, rank tests gain robustness. We refer the reader to the book by Hollander et al. (2013) for a comprehensive account of nonparametric techniques. Girón et al. (2012) used such tests for edge detection in SAR imagery.

Consider the two-sample case in which we want to compare $\mathbf{x} = (x_1, x_2, \dots, x_m)$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)$. Denote $\mathbf{xy} = (x_1, x_2, \dots, x_m, y_1, y_2, \dots, y_n)$ the joint sample. Unless explicitly stated, we will assume that all observations are different. Denote by $\mathbf{R}_x$ and $\mathbf{R}_y$ the vectors of sizes m and n which correspond to the ranks of $\mathbf{x}$ and $\mathbf{y}$ in the joint sample, respectively.

The literature is not unanimous in the names for the following tests. We will follow the denomination presented by Hollander et al. (2013).

Wilcoxon Rank Sum Test

This is a generic test for the hypothesis that x and y are the result of drawing observations from the same distribution, i.e., $\mathcal{H}_0: F_X(t) = F_Y(t)$ for every $t \in \mathbb{R}$. Typical alternatives are of the form $\mathcal{H}_1: F_Y(t) = F_X(t - \Delta)$. These alternatives are called “location-shift model,” and describe a “treatment effect” measured by $\Delta = E(Y) - E(X)$. With this, the null hypothesis reduces to $\mathcal{H}_0: \Delta = 0$.

The Wilcoxon test uses the sum of the ranks of the second sample $\sum_{i=1}^n \mathbf{R}_y(i)$. It is possible to compute its expected value and variance, with which we form the test statistic:

$$W = \frac{\sum_{i=1}^n \mathbf{R}_y(i) - n(m+n+1)/2}{\sqrt{mn(m+n+1)/12}}. \quad (1)$$

Under the null hypothesis, and when both $m, n \rightarrow \infty$, then W is asymptotically distributed as a standard Normal random variable. This asymptotic result leads to the following decision rules:

- Reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1: \Delta > 0$ if $W \geq z_\alpha$;
- Reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1: \Delta < 0$ if $W \leq -z_\alpha$;
- Reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1: \Delta \neq 0$ if $|W| \geq z_{\alpha/2}$,

(in which z_α is the α quantile of the standard Normal distribution)

Mann-Whitney Test

A different approach for the same tests presented is based on the Mann-Whitney statistic:

$$U = \sum_{i=1}^m \sum_{j=1}^n \mathbb{1}_{x_i < y_j},$$

where $\mathbb{1}_C$ denotes the indicator function of condition C . The statistic U relates to Wilcoxon's W statistic by $W = U + n(n+1)/2$, and thus, they produce equivalent tests.

van der Waerden's Test

Consider the van der Waerden's test statistic given by

$$V = \sum_{j=1}^n \Phi^{-1} \left(\frac{R_y(j)}{m+n+1} \right),$$

where Φ^{-1} is the inverse of the cumulative distribution function of a standard Normal random variable. This test statistic has symmetric distribution under $\mathcal{H}_0$. Its critical values are available through, among others, the **waerden.test** function from **R**'s **agricolae** package. The van der Waerden's test is an attractive competitor of Wilcoxon's W test.

Summary

There are several tests that use ranks rather than observations for assessing the null hypothesis that two samples $\mathbf{x}$ and $\mathbf{y}$ were produced by the same distribution $\mathcal{H}_0: F_X(t) = F_Y(t)$. They have the advantage of less sensitivity to outliers than tests that use the values directly. The performance of nonparametric tests is comparable to their parametric versions when the samples are large.

Cross-References

- [Interquartile Range](#)
- [Univariate](#)

Bibliography

- Girón E, Frery AC, Cribari-Neto F (2012) Nonparametric edge detection in speckled imagery. *Math Comput Simul* 82:2182–2198. <https://doi.org/10.1016/j.matcom.2012.04.013>
- Hollander M, Wolfe DA, Chicken E (2013) *Nonparametric statistical methods*. Wiley. https://www.ebook.de/de/product/20514944/hollander_chicken_wolfe_nonparametric_statistical_meth.html

Rao, C. R.

B. L. S. Prakasa Rao

CR Rao Advanced Institute of Mathematics, Statistics and
Computer Science, Hyderabad, India



Fig. 1 C. R. Rao, courtesy of Dr. Tejaswini Rao, daughter of C. R. Rao

Biography

Statistical methods are widely employed in all areas of activities of society including industry and government. Professor Calyampudi Radhakrishna Rao has made distinct and extensive contributions to several branches of the subject of statistics and its applications leading to efficient methods of statistical analysis. These include the theory of estimation, multivariate analysis, characterization problems, combinatorics, and the matrix algebra. In the area of estimation, Rao established that, under specific conditions, the variance of an unbiased estimator of a parameter is not less than the reciprocal of the Fisher information contained in the sample. This result became known as the Cramer-Rao inequality. He has also shown that one can possibly improve an unbiased estimator by a process known as Rao-Blackwellization; Rao developed tests known as the Score and Lagrangian multiplier

tests based on large samples for testing simple and composite hypotheses under general conditions.

Based on the properties of suitable sample statistics, Rao obtained a number of characterizations of the normal, Poisson, Cauchy, stable, exponential, geometric, and other distributions. General classes of distributions known as weighted distributions were introduced in his work on discrete distributions arising out of the methods of ascertainment. The problem of defining a generalized inverse of a singular square or rectangular matrix was studied by Rao and he developed the calculus of generalized inverses of matrices with applications to unified theory of linear estimation of parameters in a general Gauss-Markov model.

In multivariate analysis, one has to deal with extraction of information from a large number of measurements made on each sample unit. Not all measurements carry independent information. It is possible that a subset of measurements may lead to procedures which are more efficient than using the whole set of measurements. Rao developed a test to ascertain whether or not the information contained in a subset is the same as that given in the complete set. He also developed a method for studying clustering and other interrelationships among individuals or populations. Using general diversity measures applicable to both qualitative and quantitative data, the method of analysis of diversity was developed by Rao for which he introduced the concept of quadratic entropy in the analysis of diversity.

Combinatorial arrangements known as orthogonal arrays were introduced by Rao for use in the design of experiments. These arrangements are widely used in multifactorial experiments to determine the optimum combinations of factors to solve industrial problems. These have also applications in coding theory. An important result of practical interest resulting from this novel approach is the Hamming-Rao bound associated with orthogonal arrays.

Calyampudi Radhakrishna Rao was born on September 10, 1920, in Huvvina Hadagall in the state of Karnataka in India. After finishing his high school and basic undergraduate education, he went to Andhra University in Visakhapatnam, a coastal city in the state Andhra Pradesh where he obtained his B.A.(Hons) at the age of 19. After deciding to pursue a research career in mathematics, Rao went to Calcutta where he joined the Indian Statistical Institute (ISI) for training in the subject of statistics. At the same time, he completed a master's degree program in statistics at Calcutta University. Later Rao obtained his Ph.D. degree at Cambridge University in the UK after which he became a full professor at ISI at the young age of 29. Upon retiring from the ISI in 1980, he moved to the USA where he worked for another 40 years and currently is a research professor at the University of Buffalo in New York State. He received several awards including the Bhatnagar award and the India Science award from the Government of India, the National Medal of Science from USA, and was

elected as fellow of several academies of Science in India and abroad. In total he received 39 honorary doctorates from universities in India and other countries. Several students received their Ph.D. degrees under his guidance. C. R. Rao is among the worldwide leaders in statistical science over the last several decades. He celebrated his 100th birthday on September 10, 2020.

Rao, S. V. L. N.

B. S. Daya Sagar

Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, Karnataka, India



Fig. 1 S.V. L. N Rao (1928–1999). (Courtesy of “Dr. Ramana Sonty, second son of SVLN Rao”)

Biography

Professor S. V. L. N. Rao was born on May 11, 1928, in India. He studied at Andhra University and obtained his B.Sc. (Hons) and M.Sc. degrees by 1950. He started his career in 1952, as an instructor in geology at the Indian Institute of Technology (IIT) Kharagpur. By 1975, he rose to the position of professor at the same institute. He obtained his Ph.D. degree in 1959. During 1985–1987, he served as Head of the Department of Geology and Geophysics. His role in the establishment of the Remote Sensing Service Centre at the IIT campus with the funding from the Department of Science and Technology (DST) was instrumental.

Rao, earlier worked on geochemistry problems, and as a postdoctoral fellow at Kingston, Canada, he learned the use of computers while he pursued crystallography studies.

From the late 1970s, his research interests have shifted to the applications of geostatistics and mathematical morphology in satellite and geological data analysis (Rao and Srivastava 1969; Rao and Prasad 1982). As a researcher and visiting professor, he visited several countries including Canada, Hungary, Germany, France, Norway, the UK, and the USA. He served as an adviser and consultant to provide solutions in mineral exploration and mine planning work for National Mineral Development Corporation and Hindustan Zinc Limited. On the invitation of the then Vice-Chancellor of Andhra University, he joined the Andhra University, where he headed the Center for Remote Sensing and Information Systems as a Director (1988–1992) and also served as Professor of Computer Science and Systems Engineering. He has guided about 40 Ph.D. students in geophysics, geology, remote sensing, mathematics, and computer science during his four decades of teaching and research. He played an instrumental role in starting an M.Tech course in remote sensing in 1988 as the Founding Director of the Center for Remote Sensing and Information Systems. From several conversations, it was obvious that he benefited from his collaboration with Late Professor Barth of Chicago University, Late Professor Berry, the then President of American Mineralogical Society, Prof. Jean Serra, Centre de Morphologie Mathematique, and Professor Jef Johnson of Open University. In 1984, he was invited to Paris School of Mines, where he worked with Jean Serra on optical microscopy, and he was involved in developing approaches for the quantitative treatment of polarized microscopy in geology. While he worked with Jef Johnson, Prof. Rao proposed the principle of dilational similarity.

His colleagues and students remember Professor Rao as a distinguished geochemist, geostatistician, and computer scientist, and as a selfless friendly thesis adviser. He was an associate editor for the *Mathematical Geology* that released a special issue in his memory (Sagar 2001).

Cross-References

► [Mathematical Morphology](#)

Bibliography

- Rao SVLN, Prasad J (1982) Definition of kriging in terms of fuzzy logic. *Math Geol* 14(1):37–42
- Rao SVLN, Srivastava GS (1969) Comparison of regression surfaces in geologic studies. *Trans Kansas Acad Sci* (1903–) 72(1):91–97
- Sagar BSD (ed) (2001) Special issue ‘In memory of the Late Professor SVLN Rao’. *J Math Geosci* 33(3):245–396

Realizations

C. Özgen Karacan

U.S. Geological Survey, Geology, Energy and Minerals
Science Center, Reston, VA, USA

Definition

In statistics, a realization is an observed value of a random variable (Gubner 2006). In mathematical geology, the most important realizations are those in the form of maps of spatially correlated regionalized variables.

Spatial description of random variables within complex domains and making certain decisions about those require complete knowledge of the attribute of interest at each point in space. However, it is virtually impossible to sample from every location within the domain to gain a complete spatial understanding of the random variables with certainty at different scales. Therefore, limited sampling leaves us with incomplete information, which is the source of uncertainty. Understanding the uncertainty and quantifying it are essential to minimize the risks of decision making. Geostatistical simulation techniques aim to quantify spatial uncertainty of random variables by numerically reproducing the reality, which we have limited knowledge of, in a discretized (node) domain. The process creates multiple alternative maps of the random variable, which are called realizations. These are numerical representations of reality constructed by obeying the constraints of spatial statistics of the sampled data. Realizations are discretized maps explaining the uncertainty of the property of interest.

Methods of Generating Realizations

Realizations can be generated through interpolation methods (e.g., kriging), or simulation methods that rely on two-point statistics (e.g., sequential Gaussian simulation, sequential indicator simulation), or multiple point statistics (e.g., FILTERSIM and SNESIM), which utilize a training image (TI) (Remy et al. 2009). The choice of the most appropriate method may depend on the objectives and the theoretical and practical aspects of the work. For instance, interpolation methods tend to be practical, but they generate only one realization where spatial variations are smoothed along with a kriging variance surface. The drawback is that quantifying joint uncertainty of the model at unsampled locations is not possible. Also, the local variations are suppressed through smoothing, which may not be a realistic representation of reality.

Simulation methods (either relying on a variogram or a TI) are more commonly used to generate realizations since these

methods can generate multiple realizations through a stochastic process. Depending on the needs, data limitations, spatial complexity, and the objectives of the work, either a variogram-based method or a multiple point simulation algorithm along with a TI as a conceptual model to define local patterns can be used to generate realizations. The realizations generated through multiple point simulations generally exhibit local patterns similar to those conceptualized in the TI (Pyrcz and Deutsch 2014).

Simulations may or may not be conditioned to the data in sampling locations, depending on the approach. Conditional simulations honor the data at sampling locations while estimated values vary outside the sampled locations constrained by spatial continuity, trend, and secondary information (Pyrcz and Deutsch 2014). On the other hand, in realizations generated using the unconditional approach, locations of high- and low-estimated values do not necessarily correspond to the actual locations of the high- and low- sampled data as they are not preserved at their locations. Therefore, conditional simulation is preferred more often to generate realizations.

The Concept of “Equiprobable”

Multiple realizations generated by repeating the entire simulation process are used to assess the joint uncertainty of the model at unsampled locations. Each realization obeys the input statistics and trends, as well as the data at sampling locations. Since each of these realizations are generated by repeating the same stochastic process during simulations and are constrained with the same data and spatial correlations, they are equally likely to occur and equally likely to be the representation of reality. Therefore, they are often called “equiprobable.” However, equal probability refers and applies to generation and representativeness of realizations when they are considered individually. In reality, when all realizations that represent the same attribute are evaluated with a criterion that is important for the goals of the project, some realizations are more similar to others and that class has a higher probability within the entire population of realizations. Therefore, the term equiprobable is true for individual realizations in the sense that each of them is one of the outcomes of the simulation process with equal probability. However, they may not have the same probability when they are ranked.

Number of Required Realizations, Screening, and Ranking

With the vast number of realizations that can be generated through stochastic simulations, it is important to determine the optimum number of realizations needed as well as the

minimum screening and ranking criteria. Simulation algorithms can generate tens or hundreds of realizations depending on the need. However, as the number of nodes and the number of realizations increase, computational time and data storage requirements increase and may become physical limitations. For the purposes of generating a realization population for ranking and for practical purposes of spatial uncertainty assessments, usually 100 realizations can be considered optimal. This number of realizations forms a robust population for visualization, uncertainty assessment, and ranking for use in other models.

There are a set of criteria discussed by Leuangthong et al. (2004) for screening and acceptance of realizations. If these criteria are not met within acceptable range, realizations cannot be considered as representations of reality as they are intended to. According to these authors, realizations should reproduce (1) *data values at their location* – a crossplot of the data and the simulated values at the data locations should be generated to see if the data follows the 1:1 line; (2) *the histogram of the attribute of interest and basic statistics* – A “Q-Q” plot of realization data versus the data at sampling locations can be utilized to compare the two distributions, which should follow the 1:1 line. Alternatively, global reproduction of the histogram can be examined on a few randomly selected realizations. The key features to note are the histogram shape, the range of the simulated values, and the summary statistics, such as the mean, median, and the variance; (3) *the spatial continuity of the data* – the variogram should be calculated for multiple realizations and compared to the input variogram model in the same directions. The model variogram should be reproduced within acceptable ergodic fluctuations.

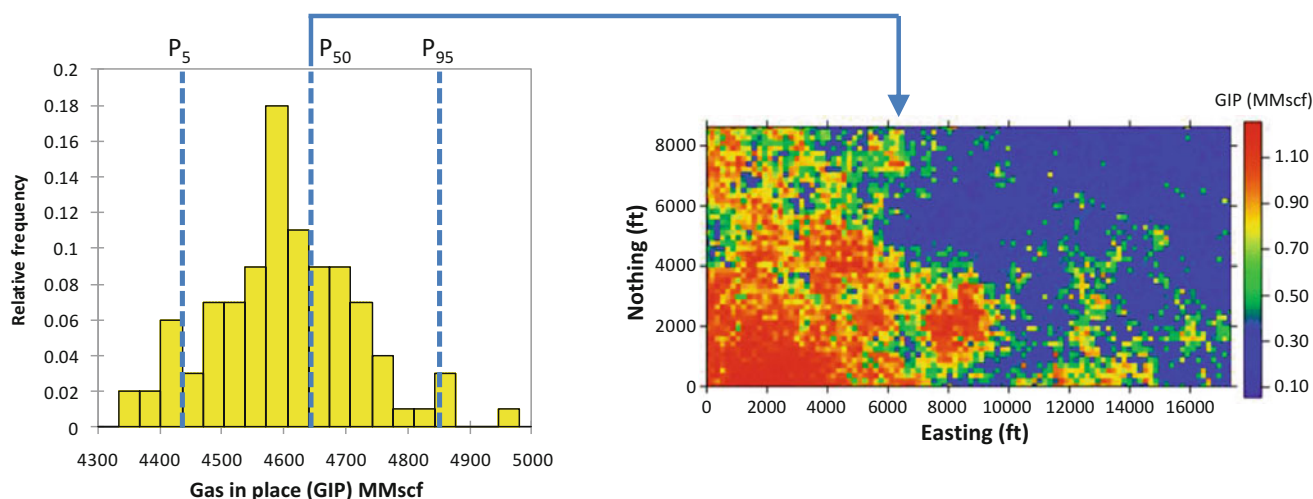
Ordering and ranking of the realizations, for both visualization and use in other models, can be achieved by capturing

the similarities and differences between them. De Barros and Deutsch (2017) argue that calculating the Euclidean distance between the realizations can be an effective approach, as well as other distance measures including the use of connectivity and other flow-related properties, or weighting the realizations according to calculated resources, such as gas quantity. For example, Fig. 1 shows a ranking of 100 total gas in place (GIP) realizations of a coal-bearing strata and the realization corresponding to P_{50} of the probability distribution.

Further Evaluation of Realizations for Mechanistic Model Uncertainty

Generation of realizations is rarely the only and the final objective of geostatistical simulations, unless realizations are for resource assessments (e.g., gold quantity or coal tonnage). For example, in petroleum engineering studies, the final product can be an uncertainty model defining reservoir architecture to aid in making optimum decisions for given injection/production scenarios. Therefore, once realizations are compiled and pass the minimum screening and acceptance criteria, they may be ranked or subjected to post-processing for use in mechanistic models, such as reservoir simulations (e.g., realizations of porosity, permeability, or gas in place) for conducting sensitivity analysis of reservoir response.

When utilizing realizations in flow models, for example, either all realizations, or ranked realizations with defined marginal probabilities (P_5 , P_{50} , and P_{95}), or the E-type models or local percentile models can be used (Karacan and Olea 2015; Thanh et al. 2020). An E-type model, which is the local average of all realizations at each node, provides an indication of the most likely value of each reservoir property at a specific node. Similarly, local percentile models can be calculated



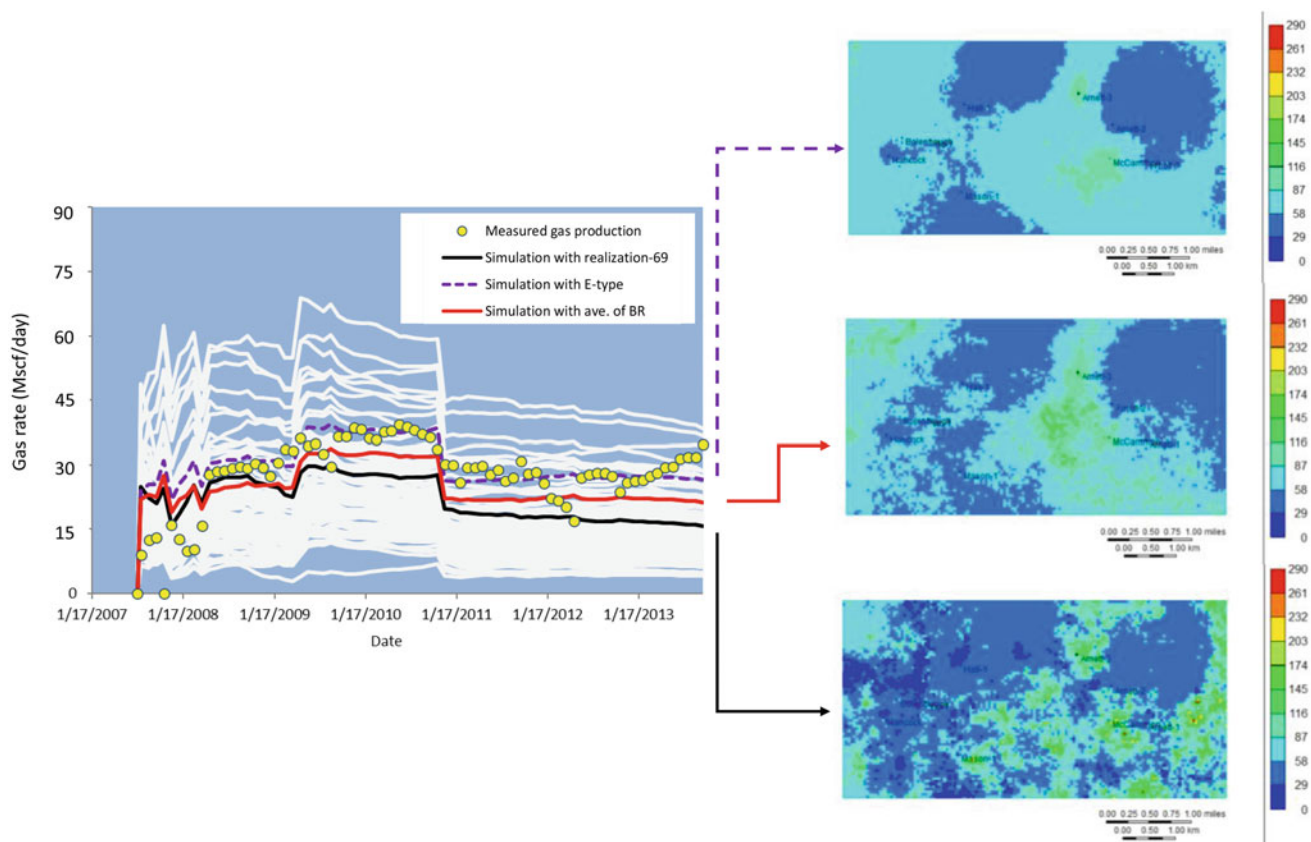
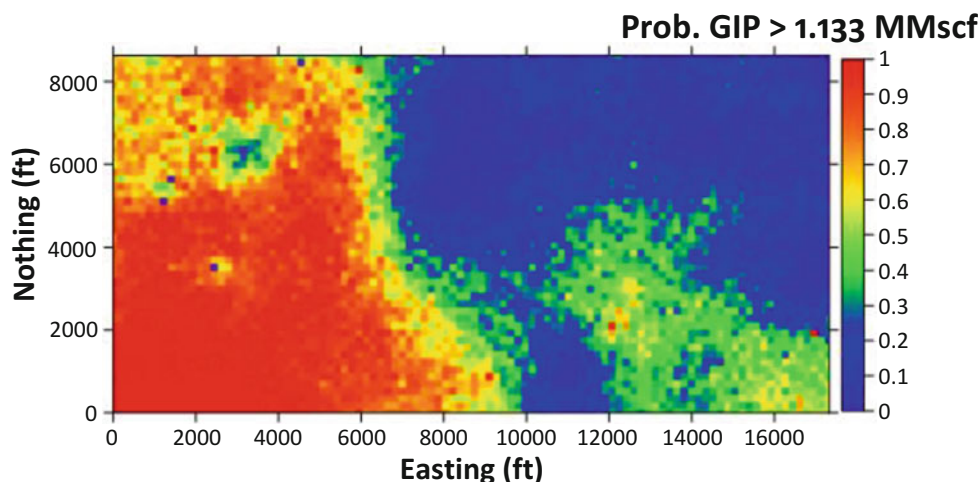
Realizations, Fig. 1 Histogram of total gas in place (GIP) realizations generated using sequential Gaussian simulation for ranking, and the P_{50} realization

from the realizations and are good indicators of areas with definitive high or low values (Fig. 2). The advantage of sequentially using all realizations of properties, on the other hand, is the ability to fully assess uncertainty and to investigate the sensitivity of reservoir response to all equiprobable representations of reality. If this is not possible, or practical,

E-type models or ranked realizations can also be used to define model architecture and assess uncertainty of reservoir behavior.

Historical oil or gas production data can be extremely useful for further assessing realizations to refine the architecture of the reservoir and to assess its uncertainty through

Realizations, Fig. 2 Local probabilities calculated for GIP value > 1.133 (mean) using 100 realizations for the same coal-bearing strata from Fig. 1



Realizations, Fig. 3 Comparison of predicted gas production rates of a coalbed methane well with the measured rates using 100 realizations generated for different properties (white lines). The results with realization number 69, the E-type model, and the average of the best

realizations (BR) are marked. Permeability realizations (in md) corresponding to those production profiles are shown on the right-hand side. (Modified from Karacan and Olea 2015)

history matching. If production data are available, the simulation results generated either using all or ranked realizations, such as E-type or different percentiles, can be compared to well-production data. For instance, if a certain combination of realizations does not produce similar production values, then these realizations can be removed from the set as possible options, thereby reducing model uncertainty (Pyrz and Deutsch 2014). Figure 3 shows a comparison of predicted coalbed methane gas production with the measured production when all 100 realizations of different properties are used in the model. Three predictions are shown on Fig. 3 (1) Predictions provided by the E-type model; (2) the realization that gives the minimum error for all wells (realization number 69); (3) the average of the realizations that minimize the error for each of the wells (best realizations). The permeability realizations corresponding to those three gas production profiles are also shown. E-type models generate reasonable production results. However, using all realizations allows for fully assessing uncertainty and for selecting individual models that best describe the reservoir architecture and contained resources (Karacan and Olea 2015). Therefore, although each realization is an equiprobable representation of reality at the geostatistical simulation level, the realizations are not equiprobable when it comes to defining the architecture of mechanistic models, as shown by the reservoir response (Fig. 3).

Summary and Conclusions

Geostatistical simulation methods generate a model of uncertainty with multiple sets of spatially distributed possible values. One set of possible outcomes is referred to as a realization. Realizations can be generated using variogram-based or TI-based methods using either conditional or unconditional approaches. Since simulations use a stochastic approach, they can generate a range of tens to hundreds of realizations, depending on the objective and physical limitations of the modeled system. However, regardless of the number of realizations, they must pass certain screening and acceptance criteria constrained by the data at sampled locations to be considered as a representation of reality and deemed useful for further post-processing or analysis. Realizations can be ranked using different distance criteria. Ranked or post-processed realizations can then be used for many different purposes spanning from resource assessments to building mechanistic models for assessing uncertainties.

Cross-References

- [Random Variable](#)
- [Simulation](#)
- [Variogram](#)

Bibliography

- De Barros G, Deutsch CV (2017) Optimal ordering of realizations for visualization and presentation. *Comput Geosci* 103:51–58
- Gubner JA (2006) Probability and random processes for electrical and computer engineers. Cambridge University Press, 363 pp
- Karacan CÖ, Olea RA (2015) Stochastic reservoir simulation for the modeling of uncertainty in coal seam degasification. *Fuel* 148:87–97
- Leuangthong O, McLennan JA, Deutsch CV (2004) Minimum acceptance criteria for geostatistical realizations. *Nat Resour Res* 13: 131–141
- Pyrz MJ, Deutsch CV (2014) Geostatistical reservoir modeling. Oxford University Press, New York. 433 pp
- Remy N, Boucher A, Wu J (2009) Applied geostatistics with SGeMS. Cambridge University Press, Cambridge, UK. 264 pp
- Thanh HV, Sugai Y, Nguete R, Sasaki K (2020) Robust optimization of CO₂ sequestration through a water alternating gas process under geological uncertainties in Cuu Long Basin, Vietnam. *J Nat Gas Sci Eng* 76:103208

Reduced Major Axis Regression

Muhammad Bilal¹, Md. Arfan Ali¹, Janet E. Nichol², Zhongfeng Qiu¹, Alaa Mhawish¹ and Khaled Mohamed Khedher^{3,4}

¹School of Marine Sciences, Nanjing University of Information Science and Technology, Nanjing, China

²Department of Geography, School of Global Studies, University of Sussex, Brighton, UK

³Department of Civil Engineering, College of Engineering, King Khalid University, Abha, Saudi Arabia

⁴Department of Civil Engineering, High Institute of Technological Studies, Mrezgua University Campus, Nabeul, Tunisia

Definition

Ordinary least squares (OLS) regression assumes that the independent variable on the X-axis is measured without uncertainty or error, and the dependent variable on the Y-axis is measured with error. However, reduced major axis (RMA) regression assumes errors in both the dependent as well as in the independent variables and minimizes the sum of the diagonal (vertical and horizontal) distances of the data points from the fitted line (Harper 2016). In geosciences, the measurements are usually exposed to several types of error that vary in magnitude and sources. For example, remotely sensed aerosol optical depth (AOD) measured with either ground instruments (such as Sunphotometers) or satellite-based sensors have errors with different magnitudes, although errors are higher in satellite-based than ground-based sensors. However, irrespective of the magnitude of the error in the independent variable X-axis (ground-based measurements) and dependent variable Y-axis (satellite-based measurements), the errors in both variables

should be considered in regression analysis. For modeling the relationship between dependent and independent variables both having an error, RMA regression is preferable to OLS regression, which assumes that the X-axis variable is measured without error (Sokal and Rohlf 1981). Moreover, if the error rate in the independent variable surpasses one-third of the error rate in the dependent variable (which is common in geosciences), then RMA regression should be considered (McArdle 1988). Firstly, RMA normalizes the variables and then fits a line of slope (β), which is the geometric mean of lines degenerating X on Y and Y on X (Sokal and Rohlf 1981). The RMA slope can be positive and negative, depending on the correlation (r) sign. However, the RMA slope is unaffected by the decreasing value of r .

Introduction

Data representation, whether geophysical observations, numerical models, or laboratory results, using the best fit line is a conventional method in the scientific and other fields. In this regard, the most common statistical methods such as ordinary least squares (OLS) regression and reduced major axis (RMA) regression are used to demonstrate a relationship between two variables (dependent and independent). However, in geosciences, both the dependent and independent variables are assumed to have errors. Specifically, remote sensing data provided by both satellite-based and ground-based instruments, have errors due to instrument calibration, algorithm uncertainties, and changing environmental conditions, etc. Therefore, to define the best fit line between remote sensing data used in geosciences applications, RMA is recommended rather than OLS (Bilal et al. 2019, 2021). The utility of RMA is demonstrated in a case study, where both RMA and OLS regression were applied to aerosol remote sensing data obtained from the MODIS (Moderate Resolution Imaging Spectroradiometer) operational Level 2 aerosol products and ground-based AERONET (Aerosol Robotic Network) Sunphotometer Version 3 Level 2.0 measurements. The remotely sensed MODIS aerosol data have pre-calculated expected errors depending on the inversion method, and similarly, the AERONET measurements have some uncertainties. The AERONET measurements are considered as standard (independent variables) for the validation of satellite-based aerosol remote sensing data (dependent variable). Therefore, the use of RMA is more suitable and recommended, as both the dependent variable (satellite remote sensing data) and independent variable (AERONET Sunphotometer measurements) have errors.

Methods

Ordinary Least Squares (OLS) Regression

Ordinary Least Squares (OLS) regression analysis is based on the assumption that the X (independent) variable is measured without an error or uncertainty and the Y (independent) variable is having an error or uncertainty. However, both X and Y variables (data) may have errors, especially, remote sensing data in the geosciences. Generally, in OLS regression, the errors in the dependent variable are ignored, but these errors should be considered for remote sensing data in the geosciences (Curran and Hay 1986). The OLS regression equation can be expressed as Eq. (1):

$$Y = (\alpha_{(OLS)} + \beta_{(OLS)}X + \varepsilon) \quad (1)$$

where

$\beta_{(OLS)}$ is the slope of OLS, $\alpha_{(OLS)}$ is the intercept of OLS, and ε is the error. The $\beta_{(OLS)}$ and $\alpha_{(OLS)}$ are calculated using Eqs. (2) and (3), respectively (Harper 2016):

$$\beta_{(OLS)} = \frac{S_{XY}}{S_{XX}} \quad (2)$$

$$\alpha_{(OLS)} = \bar{Y} - (\beta_{(OLS)})\bar{X} \quad (3)$$

where

$\bar{X}$ is the mean of the X variable, $\bar{Y}$ is the mean of the Y variable. The S_{XY} and S_{XX} can be calculated using Eqs. (4) and (5), respectively:

$$S_{XY} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \quad (4)$$

$$S_{XX} = \sum_{i=1}^n (X_i - \bar{X})^2 \quad (5)$$

Reduced Major Axis (RMA) Regression

Reduced major axis (RMA) regression is based on the assumption that both the X (independent) and Y (independent) variables are measured with error or uncertainty. RMA has many potential applications in numerous disciplines including remote sensing applications in geosciences. For validation of remote sensing observations against reference (in situ) data, where both observations and reference data have errors, RMA is a better choice compared to OLS (Curran and Hay 1986). Several studies in the fields of remote sensing applications in geosciences (Drury et al. 2008;

Li et al. 2019; Lin et al. 2014) have used RMA regression to define the best fit line between dependent and independent variables. The slope ($\beta_{(RMA)}$) and intercept ($\alpha_{(RMA)}$) of the best fit line of the RMA regression method are calculated using Eqs. (6) and (7), respectively (Harper 2016):

$$\beta_{(RMA)} = \left(\frac{S_{XY}}{S_{XX}} \right) / |r| \quad (6)$$

$$\alpha_{(RMA)} = \bar{Y} - \left(\beta_{(RMA)} \right) \bar{X} \quad (7)$$

where

$|r|$ is the absolute value of the Pearson correlation coefficient (r), and the slope ($\beta_{(RMA)}$) calculated using Eq. (6) can be positive or negative. The slope ($\beta_{(RMA)}$) can also be calculated using Eq. (8) and the positive or negative sign of the slope depends on the sign of the Pearson correlation coefficient (r).

$$\beta_{(RMA)} = \pm \frac{\sigma_Y}{\sigma_X} \quad (8)$$

where

σ_Y is the standard deviation of the Y (dependent) variable and σ_X is the standard deviation of the X (independent) variable.

Evaluation Method

To report the performance and errors in the RMA and OLS regression models, the mean bias error (MBE; Eq. 9), the root-mean-squared difference (RMSD; Eq. 10), and the mean systematic error (MSE; Eq. 11) are used. The MSE is considered as a significant method to show the difference between the trend-line of the independent (X) and dependent (Y) data, where a small value of MSE indicates a good trend and a large value of MSE denotes a trend has an error.

$$MBE = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - X_i) \quad (9)$$

$$RMSD = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - X_i)^2} \quad (10)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - X_i)^2 \quad (11)$$

Where $\hat{Y}$ is the predicted value obtained from the RMA and OLS regression models.

Results and Discussion

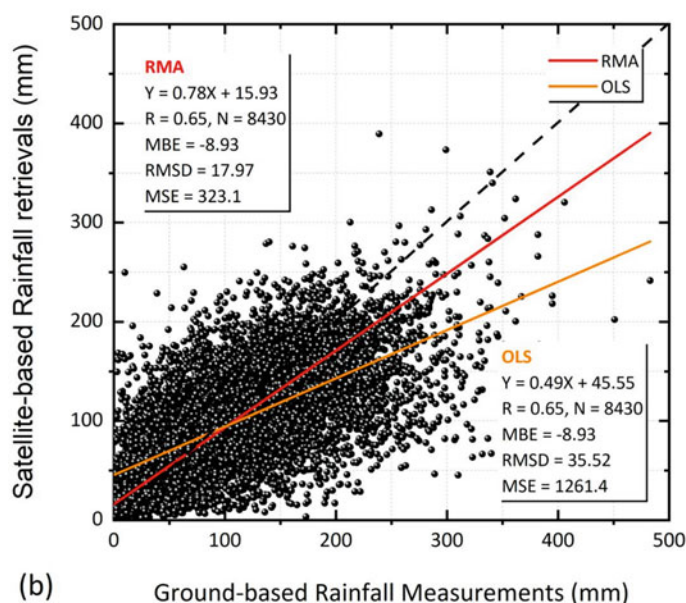
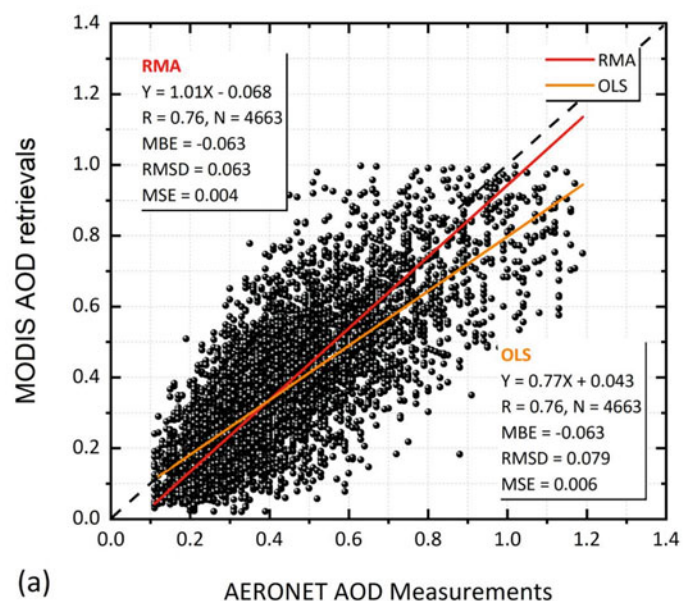
This study used remotely sensed aerosol optical depth (AOD) data from both satellite and ground-based Sunphotometer to demonstrate the performance of both the RMA and OLS models. Figure 1a shows the scatterplot and linear fit between MODIS AOD retrievals against AERONET Sunphotometer AOD measurements (reference data) calculated with both RMA (red line) and OLS (orange line) methods. The results show a better slope ~ 1.01 for RMA compared to the ~ 0.77 slope for OLS, which indicates a significant underestimation by 23%. However, no significant over- and underestimations were observed in the slope calculated by RMA. The results also showed a smaller MSE of 0.004 and RMSD of 0.063 for the RMA model compared to the OLS model (MSE = 0.006 and RMSD = 0.079). However, both the regression models showed the same MBE of -0.063 . To further demonstrate the performance of RMA vs. OLS, satellite-based rainfall retrievals were validated against ground-based rainfall measurements (Fig. 1b). A significant difference was also observed in slope and intercept calculated by RMA compared to those calculated by OLS with small errors (RMSD, and MSE). These results suggest the better performance of RMA regression, especially in terms of slope calculation, compared to OLS, and recommend the use of the RMA regression model for geoscience research applications.

Summary and Conclusions

In geosciences, satellite remote sensing data are usually required to validate against ground-based instrumental measurements. However, both datasets depend on instrumental calibration, which may introduce errors in the datasets. Therefore, the use of appropriate regression analysis is required to model the relationship between dependent and independent variables, which can consider errors in both variables. However, in general, the OLS regression method is most commonly used in geosciences to model the relationship between two variables, and it is assumed that the independent variable is measured without error, although this rarely occurs. On the other hand, the RAM method models the relationship between two variables considering errors in both the dependent as well as independent variables. Therefore, RMA regression should be considered more suitable for satellite remote sensing applications. To demonstrate the performance of RMA against OLS, remotely sensed satellite-based aerosol and rainfall data were validated against ground-based measurements. The results showed that the RMA regression is better able to model the relationship between the variables

Reduced Major Axis

Regression, Fig. 1 Reduced major axis (RMA) regression vs. ordinary least squares (OLS) regression. (a) Validation of MODIS AOD retrievals against AERONET AOD measurements and (b) Validation of satellite-based rainfall retrievals against ground-based rainfall measurements. Where the dashed black line represents the 1:1 line, the red line represents the line of the reduced major axis (RMA) regression, and the orange line represents the line of the ordinary least squares (OLS) regression



compared to OLS regression in terms of slope, small MBE, RMSD, and MSE. This study recommends that selecting a proper regression method is essential for modeling relationships in geoscience data and that RMA regression is more suitable than OLS.

Acknowledgments The authors would like to acknowledge NASA's Level-1 and Atmosphere Archive and Distribution System (LAADS) Distributed Active Archive Center (DAAC) (<https://ladsweb.modaps.eosdis.nasa.gov/>) for MODIS data and Principal Investigators of AERONET sites for AOD measurements. This research was funded by the National Key Research and Development Program of China (2016YFC1400901), the Jiangsu Provincial Department of Education for the Special Project of Jiangsu Distinguished Professor (R2018T22), the Deanship of Scientific Research at King Khalid University (RGP.1/372/42), and the Startup Foundation for Introduction Talent of NUIST (2017r107).

Bibliography

- Bilal M, Nazeer M, Nichol JE, Bleiweiss MP, Qiu Z, Jäkel E, Campbell JR, Atique L, Huang X, Lolli S (2019) A simplified and robust Surface Reflectance Estimation Method (SREM) for use over diverse land surfaces using multi-sensor data. *Remote Sens* 11:55
- Bilal M, Mhawish A, Nichol JE, Qiu Z, Nazeer M, Ali MA, de Leeuw G, Levy RC, Wang Y, Chen Y, Wang L, Shi Y, Bleiweiss MP, Mazhar U, Atique L, Ke S (2021) Air pollution scenario over Pakistan: characterization and ranking of extremely polluted cities using long-term concentrations of aerosols and trace gases. *Remote Sens Environ* 264:112617
- Curran P, Hay A (1986) The importance of measurement error for certain procedures in remote sensing at optical wavelengths. *Photogramm Eng Remote Sens* 52:229–241
- Drury E, Jacob DJ, Wang J, Spurr RJD, Chance K (2008) Improved algorithm for MODIS satellite retrievals of aerosol optical depths over western North America. *J Geophys Res* 113:D16204

- Harper WV (2016) Reduced major axis regression. In: Wiley StatsRef: statistics reference online. Wiley Online Library, Hoboken
- Li Z, Roy DP, Zhang HK, Vermote EF, Huang H (2019) Evaluation of Landsat-8 and Sentinel-2A aerosol optical depth retrievals across Chinese cities and implications for medium spatial resolution urban aerosol monitoring. *Remote Sens* 11:122
- Lin J, van Donkelaar A, Xin J, Che H, Wang Y (2014) Clear-sky aerosol optical depth over East China estimated from visibility measurements and chemical transport modeling. *Atmos Environ* 95:258–267
- McArdle BH (1988) The structural relationship: regression in biology. *Can J Zool* 66:2329–2339
- Sokal RR, Rohlf FJ (1981) *Biometry*, 2nd edn. Freeman, New York

Regression

D. Arun Kumar¹, G. Hemalatha² and M. Venkatanarayana¹

¹Department of Electronics and Communication Engineering and Center for Research and Innovation, KSRRM College of Engineering, Kadapa, Andhra Pradesh, India

²Department of Electronics and Communication Engineering, KSRRM College of Engineering, Kadapa, Andhra Pradesh, India

Definition

Regression analysis is a statistical approach to find the mathematical relationship between input independent variables and output dependent variables using linear/nonlinear equations. The regression model (also called as linear or nonlinear equation) is used to predict the values of output variable for the given values of input variables.

Introduction

The Earth's resources like vegetation, minerals, water, etc. are monitored using remote sensing. Remote sensing data analysis is based on statistical methods like regression, classification, and clustering (Richards and Jia 2006). Regression (also identified as prediction) involves obtaining output values of themes such as temperature, slope, pressure, elevation, etc., for the respective input values (Jensen 2004). One of the major applications of regression analysis is the prediction of weather parameters like pressure, temperature, and humidity using the data obtained from the sensors. Data classification is labeling the pixels in images to the classes like land use/land cover, weather parameters, etc. Clustering is an unsupervised data analysis approach in which the data points (also called as pixels in the image) are grouped into clusters without class labels (Theodoridis and Koutroumbas 1999). The clustered data points are assigned with corresponding class labels using ancillary data and ground truth.

The information related to Earth resources is represented in the form of digital images (Matus-Hernández et al. 2018). A digital image is mathematically defined as 2D function. The digital image consists of finite elements called pixels (Lillesand and Kiefer 1994). The digital images in remote sensing vary from monochromatic to multispectral and hyperspectral images (Campbell 1987). In multispectral and hyperspectral images, the pixels are also defined as pixel vectors or patterns (Bishop et al. 1995). A pattern is an entity associated and characterized by features (Duda et al. 2000). The pixel vector is a point in N-dimensional feature space. Regression analysis is predicting the output values for the unknown patterns based on the existing output values of patterns in the dataset (Doan and Kalita 2015). Regression analysis is implemented using a mathematical model by fitting a linear/nonlinear equation from the training data (Matus-Hernández et al. 2018).

There exist various methods of regression analysis for remote sensing application. These methods are broadly classified as linear and nonlinear approaches. The detailed description of these methods is given in following sections. Different methods of regression analysis are discussed in section “[Type of Regression Analysis in Remote Sensing](#).” The evaluation methods of regression analysis are given in section “[Evaluation of Regression model](#).” The results of linear regression model for an example dataset is provided in section “[Example: Linear Regression](#).” The relevant conclusion of regression analysis is given in section “[Conclusions](#).”

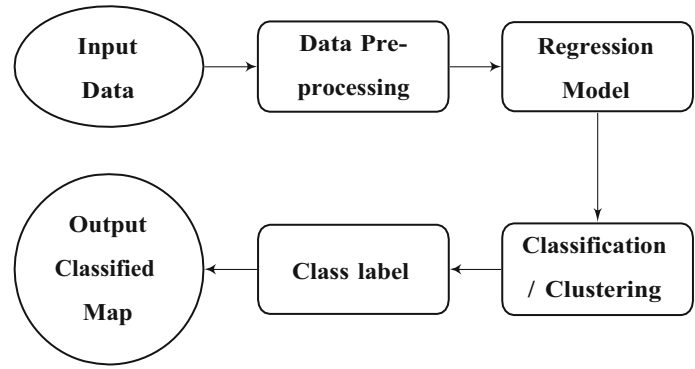
Type of Regression Analysis in Remote Sensing

The functional block diagram of the regression model is provided in Fig. 1. The input pixels/patterns are used to prepare the required dataset to fit the regression model. The data is preprocessed, and later, the dataset is divided into training set, testing set, and validation set. The training set is used to fit the equation for the regression model. The regression models are classified as linear or nonlinear models based on the relationship between input variables and output variable. A linear equation representing linear regression model is used to predict the values of the output variable if there exists a linear relationship between input variables and output variable. The equation for a single input variable linear regression model is given as:

$$y = a_0 + a_1x \quad (1)$$

where y is the output variable, x is a single input variable, and a_0 and a_1 are coefficients. Similarly, a nonlinear equation is used to fit a regression model if the relationship between input feature and output feature is nonlinear. In non-linear

Regression, Fig. 1 Regression model for remote sensing data analysis. The model considers input data, preprocessing of input data, fitting the linear or nonlinear equation, classifying/clustering the data, assigning the class label, and obtaining the classified map



regression analysis, a polynomial equation is used to fit the relationship between the dependent variable (output variable) and independent variable (input variable). The n th order polynomial equation for the nonlinear single input regression model is given as:

$$y = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n, \quad (2)$$

where a_0 , a_1 , a_2 , and a_n are coefficients.

The regression models are also classified based on the dependence of the output variable on the number of input variables (features). The models are classified as single output-single input regression model, single output-multi-input regression model, multi input-multi-output regression model. The single output single input regression model predicts the values of the output variable which is dependent on the single input variable. The mathematical equation for single output-single input regression model is given in Eq. 1. The single output-multi-input regression model is used to predict the values of output variable which depends on multiple inputs. The equation for single output multi-input regression model is given as:

$$y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n, \quad (3)$$

where a_0 , a_1 , a_2 , $\dots$, a_n are coefficients and x_1 , x_2 , and x_n are input variables or features. The models such as neural networks are used for multi-input-multi-output regression analysis.

The regression model is tested with the test dataset by passing each pattern/pixel vector and predicting the values of the output variable. The regression model is validated using the validation dataset. The predicted values of the patterns/pixels are then classified by using various supervised/unsupervised classification techniques. The major supervised classification techniques include K-NN, minimum distance to mean, neural networks, decision trees, and random forest. The unsupervised classification techniques include clustering techniques like K-means clustering, hierarchical clustering,

density-based SCAN, etc. The patterns are labeled with the corresponding class labels to obtain the output classified map.

Evaluation of Regression Model

The performance of the regression model is evaluated using three important performance metrics like R square, root mean square error (RMS), and mean absolute error (MAE). R square is the square of the correlation coefficient (R), and it is the measure of variability in dependent/output variable. Root mean square error is the square root of mean square error. Mean square error is the sum of the square of predicted output value and the actual value divided by the total number of samples. RMS error is calculated as:

$$RMS = \frac{\sum_{i=1}^n (y_i - y_i')^2}{N} \quad (4)$$

where N is the total number of patterns/pixels and n is the n th pixel in N number of pixels. y_i is the actual value of output variable, and y_i' is the predicted value of output variable.

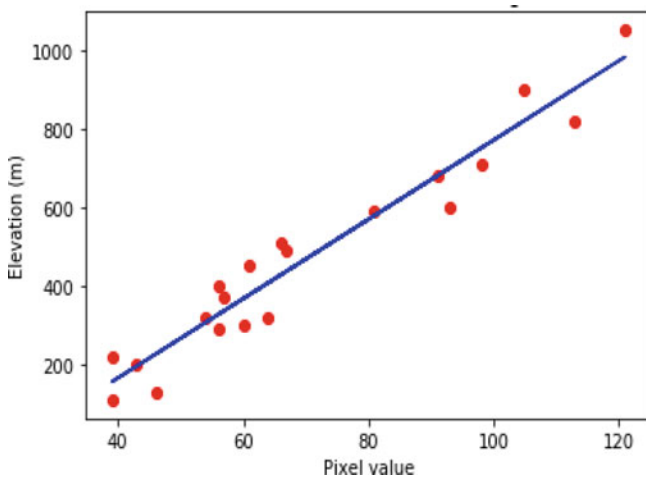
The mean absolute error (MAE) is the sum of absolute value of error divided by total number of pixels.

$$MAE = \frac{\sum_{i=1}^n |y_i - y_i'|}{N} \quad (5)$$

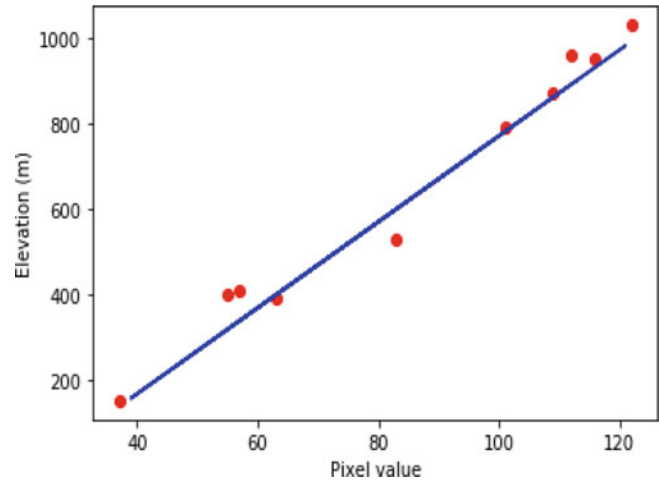
The values of evaluation metrics for the regression model lies between 0 and 1 while 0 indicates less error and 1 indicates the maximum error.

Example: Linear Regression

In the present study, two datasets were considered to demonstrate model fitting for linear regression and polynomial regression. The first dataset consists of pixel values and



Regression, Fig. 2 Linear regression model to predict elevation values for the given pixel values (training set). The pixel values and the corresponding elevation values are collected from the SRTM digital elevation model (DEM) image



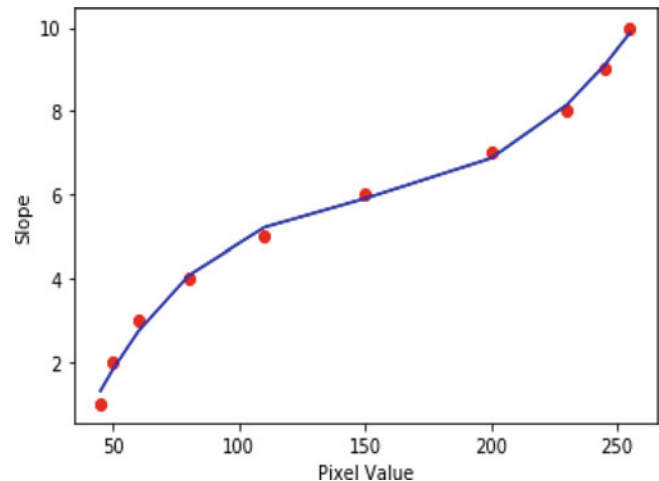
Regression, Fig. 3 Linear regression model to predict elevation values for the given pixel values (test set). The pixel values and corresponding elevation values are collected from the SRTM digital elevation model (DEM) image

corresponding ground elevation (*m*). The dataset is prepared by considering the SRTM DEM image with a spatial resolution of 90 m. The dataset is divided into the training set and test set. The training set is used to fit the linear regression model, and the test set is used to test the performance of the model. The linear regression model for the training dataset is given in Fig. 2. The linear regression model is tested on the test dataset, and the relationship between the predicted values and the input pixel values is given in Fig. 3.

The second dataset consists of pixel values and corresponding slope values. The dataset is generated from the raster slope map using QGIS. The polynomial regression model for the second dataset is given in Fig. 4. In Fig. 4, the *x*-axis represents the digital number (DN) or pixel value, and the *y*-axis represents the slope value. The fourth-order polynomial is used to fit the polynomial regression model for the second dataset. The performance of regression models was evaluated using three metrics such as RMSE, MAE, and *R* square. The values of three metrics were obtained between 0 and 1 for two datasets.

Conclusions

In the present study, the basic regression models for data prediction were presented with example datasets. Emphasis is given on regression models such as linear, polynomial, single input-single output, and single output-multi-input. In the present study, two datasets were considered as an example to demonstrate the functioning of linear and polynomial regression analysis for both training and test cases. The model was evaluated using the metrics like *R* square, MSE, and MAE, and the values were obtained between 0 and



Regression, Fig. 4 Polynomial regression model to predict slope values for the given pixel values. The pixel values and corresponding pixel values are collected from a raster slope map generated using QGIS

1. Furthermore, there exist advanced regression models like logistic regression, robust regression, probit regression, ridge regression, etc., to predict the values of the output variable.

Acknowledgments The author is thankful to K. Madan Mohan Reddy, vice chairman, Kandula Srinivasa Reddy College of Engineering, and A. Mohan, director, Kandula Group of Institutions for establishing the Machine Learning group at Kandula Srinivasa Reddy College of Engineering, Kadapa, Andhra Pradesh, India – 516003.

Bibliography

Bishop CM et al (1995) Neural networks for pattern recognition. Oxford University Press, Oxford

- Campbell JB (1987) Introduction to remote sensing. The Guilford Press, New York
- Doan T, Kalita J (2015) Selecting machine learning algorithms using regression models. In: 2015 IEEE international conference on data mining workshop (ICDMW), IEEE, pp 1498–1505
- Duda RO, Hart PE, Stork DG (2000) Pattern classification, 2nd edn. Wiley Interscience Publications, New York
- Jensen JR (2004) Introductory digital image processing: a remote sensing perspective, 3rd edn. Prentice Hall, Upper Saddle River
- Lillesand TM, Kiefer RW (1994) Remote sensing and image interpretation. John Wiley and Sons, Inc., Chichester
- Matus-Hernández MÁ, Hernández-Saavedra NY, Martínez-Rincón RO (2018) Predictive performance of regression models to estimate chlorophyll-a concentration based on landsat imagery. PLoS One 13(10):e0205682. <https://doi.org/10.1371/journal.pone.0205682>
- Richards JA, Jia X (2006) Remote sensing digital image analysis: an introduction, 4th edn. Springer Verlag, Berlin, Heidelberg
- Theodoridis S, Koutroumbas K (1999) Pattern recognition and neural networks. In: Advanced course on artificial intelligence. Springer, pp 169–195. Pattern Recognition, 4th edn. Academic Press, Inc., USA

Remote Sensing

Ranganath Navalgund¹ and Raghavendra P. Singh²

¹Indian Space Research Organisation (ISRO), Bengaluru, India

²Space Applications Centre (ISRO), Ahmedabad, India

Definition

Remote sensing (RS) is the science of making inference from measurements made at distance without any physical contact with the objects under study. It refers to identification of earth features by detecting electromagnetic (EM) radiation that is reflected and/or emitted by the surface at various wavelengths. It is based on the scientific rationale that every object reflects/scatters a portion of electromagnetic energy incident on it. The reflection depends on physical and chemical properties of the material. The object also emits radiation depending upon its temperature and emissivity.

Introduction

Visual perception of objects by the human eye is an example of remote sensing. However, it is confined to visible portion of EM spectrum. Modern remote sensors are capable of detection of electromagnetic radiation extending from gamma rays, x rays, and ultraviolet to infrared and microwave regions. While, in general, the medium of interaction for RS is through the electromagnetic radiation, gravity and acoustics are also employed in some specific circumstances. However, acoustic RS is very limited because it requires a physical medium for transmission. Aerial remote sensing started way

back in 1858 from balloon-based photography of Paris and later through rocket-propelled camera systems by the Germans in 1891. The first image of the Earth from space was taken by the Explorer 6 satellite of the United States in August 1959. Systematic Earth observation started with the launch of Television Infrared Observation Satellite (TIROS-1) on April 1, 1960, designed primarily for meteorological observations. Remote Sensing as we understand today gained momentum after the launch of Earth Resources Technology Satellite-1 (ERTS, later renamed as LANDSAT 1) in 1972 by the National Aeronautics and Space Administration, USA. The basic objective of earth remote sensing is to obtain inventory of natural resources in a spatial format at different time intervals, to ascertain their condition, to observe dynamics of the atmosphere and oceans, and to help understand various physical processes in an integrated manner. Subsequent to the launch of LANDSAT-1, many space agencies of different countries and some private players have successfully launched a number of RS satellites.

Physical Basis and Signatures

Study of the interaction of EM radiation with matter at different wavelength regions is paramount to remote sensing. Electromagnetic radiation covers a large spectrum from the very short wavelength gamma rays ($<10^{-10}$ m) to long radio waves (10^6 m). Mostly visible (0.4–0.7 μm), the reflected infrared (IR) (0.7–3 μm), the thermal IR (3–5 μm and 8–14 μm), and the microwave region (0.3–300 cm) are used in RS. The sun acts as a source of EM radiation used in passive optical remote sensing. As per Planck's law, all bodies at absolute temperatures above zero degrees Kelvin emit EM radiation which is a function of wavelengths (λ) and temperature.

$$E(\lambda, T) = \frac{2hc^2}{\lambda^5} \frac{1}{e^{hc/\lambda kT} - 1} \quad (1)$$

where $E(\lambda, T)$ is spectral radiant emittance in ergs/s, $h = 6.625 \times 10^{-27}$ erg sec (Planck's constant), $k = 1.38 \times 10^{-16}$ erg/K (Boltzmann constant), $c = 3 \times 10^{10}$ cm/s (Speed of light), and T is absolute temperature in °K.

The earth may be considered as a blackbody at ~ 300 °K emitting EM radiation with peak emission at around 9.7 μm . While EM radiation at shorter wavelengths (visible and near IR) interacts with a few micrometers of the earth surface, EM radiation at microwave wavelengths has greater subsurface penetration capability. Different land cover features, like water, soil, vegetation, rocks, cloud, and snow, reflect visible and infrared light in different ways. Vegetation reflectance is mostly governed by photosynthetic pigments in visible region

(400–700 nm), in near-infrared region (0.75–1.35 μm) by the leaf structure, and in the shortwave infrared by the presence of water. As vegetation becomes stressed or senescent, chlorophyll absorption decreases (blue and red regions). The reflectance ratio of the near infrared to red is known as a sensitive indicator of vegetation growth/vigor. Typical soil spectral signature shows lower reflectance in lower wavelength and high reflectance at larger wavelength. Water is relatively an absorber in the near-infrared and middle-infrared wavelength. This property helps in delineation and inventory of even small wetlands. Phyto-planktons present in inland wetlands, the coastal water, and ocean are associated with absorption due to chlorophyll pigment in narrow spectral regions facilitating their estimation. Snow exhibits high reflectance up to 0.8 μm and relatively low reflectance in middle infrared. Clouds look uniformly bright throughout the range 0.3 to 3 μm due to, nonselective scattering. Emissivity signature in thermal infrared region provides information on mineral composition. Most of the rock-forming silicate minerals have their diagnostic features in 8–14 μm . Remote sensing in the microwave region (frequencies between 10^9 and 10^{12} Hz) is highly useful as it provides observation of the earth's surface regardless of day/night and the atmospheric conditions. It also provides certain unique properties of targets in view of its ability to penetrate and its interaction with electrical properties of the targets. Microwave response is dependent upon its frequency, incidence angle, and polarization and also on emissivity characteristics of targets. The radar equation relates received signal with radar parameters and the target characteristics. For monostatic radars, it is given as

$$P_r = \frac{\lambda^2}{(4\pi)^3} \int \frac{P_t G^2 \sigma^0}{R^4} dA \quad (2)$$

where P_r = average power returned to the radar antenna from an extended target, P_t = power transmitted by the radar, G = gain of the antenna, R = distance from the antenna to the target, λ = wavelength of the radar, and σ^0 = the radar-scattering coefficient of the target.

Spectral response measured at the ground level is modified by the intervening atmosphere when measured by spaceborne sensor. Therefore, it is essential to understand and quantify the atmospheric effects on the radiation transfer between Sun, Earth, and sensor. Spectral regions where the EM radiation is transmitted through the atmosphere with little attenuation are called atmospheric windows. Manifestation of interaction of EM radiation with nonpoint targets is characterized by signatures comprising wavelength-wise spectral variations; changes in the reflectance or emittance, spatial variations; changes in the reflectance/emittance determined by the target characteristics' (shape, size, and texture) temporal variations; diurnal and/or seasonal changes in reflectance or emittance

and polarization variations; and variations in the polarization of EM wave reflected or emitted. Signatures are statistical in nature and not deterministic.

Sensors, Platforms, and Orbits

One needs a variety of instruments/sensors with a great degree of sensitivity and viewing geometries to study the interaction of EM radiation with the earth surface, atmosphere, and oceans. Sensors are broadly categorized in two categories, i.e., passive or active sensors. Passive sensors detect natural radiations either emitted or reflected from the earth. Active sensors have their own source of radiation to illuminate the target materials and detect the scattered EM wave. Initially the photographic cameras were widely used as imaging system. Currently, they are used in unmanned aerial vehicles (UAV) for various field surveys. Later, opto-mechanical scanners became a popular method of RS after photographic and TV cameras. Linear imaging self-scanning systems using charge-coupled devices are currently used to provide very high-resolution imagery and to avoid possible failures of the mechanical scan mechanisms. Sensor is characterized by spatial, spectral, and radiometric resolution. Spatial resolution refers to the ability of the sensor to distinguish and resolve the smallest object on the ground. Spectral resolution represents the spectral bandwidth (Full Width at Half Maxima) of imaging system. Radiometric resolution of sensor is quantified in terms of its ability to distinguish between smallest changes in the reflectance/emittance between various targets, generally quantified as noise-equivalent change in reflectance ($\text{NE}\Delta\rho$). Dynamic range is another important sensor characteristic, which refers to the measurements associated with minimum to maximum reflectance.

Hyperspectral instruments combine the concept of optical imaging (camera-/multi-spectral sensor) with well-established tools such as diffraction grating/Michelson/Sagnac interferometer, tunable filters for splitting light into narrow spectral components in airborne/spaceborne missions. Hyperspectral signature can detect the individual absorption features, since all materials are bound by chemical bonds; thus, they can be identified by their spectral characteristics more accurately as compared to broadband multispectral imagers.

Passive Microwave radiometers are used to measure the energy emitted by the objects in the microwave region. Antenna receives the emitted energy, and the signal is processed to derive brightness temperature of the surface. The brightness temperature is a product of the absolute physical temperature and emissivity. Generally, microwave radiometers have coarse spatial resolution. Side-Looking Airborne Radar (SLAR) is a type of active imaging sensor, which measures backscattered microwave radiation. Small real

aperture antenna limits the spatial resolution of SLAR. Synthetic Aperture Radar (SAR) is a coherent side-looking radar system, which utilizes the flight path of the platform to simulate an extreme large antenna or aperture electronically and generates high-resolution remote-sensing image. SAR interferometry is a technique based on combining two SAR images of similar region imaged from different positions and/or times. It is used for the generation of digital elevation model/topographic mapping and detection of small coherent movements (differential interferometry). In case of optical sensors, heights are determined using along track or across track stereo viewing. Lidar is also used to determine surface heights. While a microwave altimeter uses a microwave pulse in nadir direction to estimate surface heights of sea, scatterometer detects the backscattered signal from the sea surface to quantify surface roughness of sea and in turn surface wind velocity. Microwave sounders are radiometers operating at suitably placed wavelengths of an absorption band of oxygen, water vapor, etc. to estimate vertical profiles of some physical parameters.

Platforms and Orbits: The sensors are mounted on suitable platforms. Platforms can be a tripod, balloon, aircraft, drone, and a spacecraft. The height of the platform dictates the spatial resolution and swath of image. In general, higher platform allows wide area coverage with lower spatial resolution. Spacecrafts are mostly three-axes stabilized to facilitate nadir viewing, though there were some spin-stabilized systems in the initial years. A spacecraft comprises various subsystems such as mainframe, payloads, attitude and orbit control, power, thermal, telemetry, tracking and control, propulsion, etc. In order to provide greater revisit/repeat observations, many identical satellites are flown in tandem. Geostationary and Near-Earth orbit are two main types of orbits used in remote sensing. In a geostationary orbit, the satellites are located at 36,000 km in the equatorial plane of the earth. Satellite appears stationary to the earth as a period of revolution of satellite and rotation of the earth matches in this orbit. Geostationary orbit has the advantage of high revisit period (15–20 min) but is limited by spatial resolution (few km.). Near-polar, sun-synchronous orbit located at few hundreds of km height is preferred for high-resolution imaging with constant solar illumination condition. Such orbits are used to take images of few meter resolution and ensure repetitive observations (few days) of a scene at same local time for various terrestrial applications. In addition, inclined orbits to the equator that are not sun synchronous are also employed in certain specific cases to obtain greater visits per day.

Data Preprocessing, Products, and Analysis

RS data acquired from the sensors onboard the satellites by the earth station is in digital form. The acquired data has to be

preprocessed and corrected for various factors such as (i) image features of the sensor; (ii) stability and orbit parameters of the platform; (iii) scene surface characteristics; (iv) motion of the earth including its curvature; and (v) atmospheric effects. Normally standard data products are generated applying geometric and radiometric corrections. The procedures employed for geometric correction generally treat distortions in two groups, i.e., those which are systematic (or predictable), and those which are essentially random. The former includes the effect due to the earth rotation, the earth curvature, deviation from nominal altitude and attitude, variations of the above deviations during the imaging of a scene, etc. Nonlinear behavior of detector, responsivity variation between the detectors, pixel dropouts etc. result in radiometric distortions. Data thus obtained is subjected to image processing/classification techniques to extract information on earth surface features. Inversion of spectral radiance measured by the sensor at the top of the atmosphere is used to derive various geophysical parameters using radiative transfer modeling, which in turn are used to understand physical processes related to biosphere, atmosphere, hydrosphere, cryosphere, etc. RS data are growing in scope, volume, and variety. Processing of such data involves development in field of data science and associated techniques, i.e., neural network, genetic algorithm, machine learning, artificial intelligence, and cloud computing. Calibration is important for generating precise RS data product and associated geophysical parameters. Correction factors for sensor-related radiometric errors are normally generated by extensive calibration measurements during laboratory tests as well as postlaunch campaigns over natural or artificial targets. Standard lamps, solar diffuser reflectance panel, and blackbody are used for onboard calibration. Similarly, ground-based corner reflectors and spectrometric measurements are used for postlaunch calibration. Large patches of uniform and pseudoinvariant features like forest, desert, are also used for vicarious calibration. Data products/geophysical/biophysical parameters derived from system outputs are validated through field campaigns.

Remote-Sensing Applications

RS data are widely used in (i) inventory, monitoring, and condition assessment of natural resources viz. agriculture, forestry, minerals, water, and marine ecosystem, snow and glaciers, desertification, etc.; (ii) generation of thematic maps and updating of topographic maps; (iii) studying the atmosphere, cloud movements, and meteorology; (iv) locating living and nonliving ocean resources and in ocean state forecasting; and (v) disaster monitoring, mitigation, and early warning and in providing inputs to climate change studies. A variety of application programs, viz. agricultural crop production forecasts at national/regional level, forest

extents, identifying areas of deforestation, inventory of surface water bodies, snow cover, glaciers and their retreat, coastal zones, urban areas and their sprawl, exploration of geological resources, desertification, updating of topographic and thematic maps etc., help in planning resource management in many countries. Satellites-based observation of ocean and atmospheric is used to improve ocean-state forecasting and weather forecasting, and reduce exploitation of some of the ocean resources, in particular marine fisheries. Satellite observations have found an important and crucial role in facilitating early warning of some of the disasters, in monitoring and mitigation exercises, and in damage assessment. RS data have also been used in studying earth system science and in understanding various earth processes.

Summary and Future Direction

While earth observation from space has made significant contributions in studying different facets of the planet earth and its environment, there is a need to study the earth as a unified system of land, oceans, and atmosphere with its interlinkages and biogeochemical processes. Development of smaller and agile satellites in form of constellation and flying in formation are needed to achieve multidisciplinary/multitemporal observation capability. Technology has also helped in the collection of large volume of data, transmission, processing, and dissemination/distribution efficiently across the globe. With technological advances in Earth observation (resolutions, bands, coverage, sophistication of measurements, etc.), the concept of user segment has also seen a dramatic change. Although space technology has tremendously advanced the observation domain for many applications, there are still some challenges. Regional scale soil moisture, ocean salinity, surface pressure, vertical profile of winds, and vegetation florescence are some of the challenges, which need attention.

Bibliography

- Colwell RN, Estes DJE, Thorley GA (eds) (1983a) Manual of remote sensing. vol II: Interpretation and applications. American Society of Photogrammetry, Falls Church
- Colwell RN, Simonnet DS, Ulaby FW (eds) (1983b) Manual of remote sensing. vol I: Theory, instruments and techniques. American Society of Photogrammetry, Falls Church
- Joseph G (ed) (2003) Fundamentals of remote sensing. Universities Press (India) Private Limited, Hyderabad
- Joseph G, Navalgund RR (1991) Remote sensing – physical basis and its evolution. In: Srivastava US (ed) Glimpses of science in India. Malhotra Publishing House, New Delhi, pp 357–383
- Liang S (ed) (2018) Comprehensive remote sensing, vol 1–9. Elsevier, Amsterdam
- Navalgund RR (2002) Remote sensing. Resonance 7(1):37–46

- Navalgund RR, Jayaraman V, Roy PS (2007) Remote sensing applications: an overview. Curr Sci 93:1747–1766
- Slater PN (1980) Remote sensing, optics and optical system. Addison-Wesley Publishing Company, Reading
- Ulaby FT, Moore RK, Fung AK (1981) Microwave remote sensing: fundamentals and radiometry, vol 1. Addison Wesley, Reading

Reproducible Workflow

Anirudh Prabhu and Peter Fox

Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY, USA

Synonyms

Computational workflow; Reproducibility; Research workflows

Definition

Science advances on a foundation of trusted discoveries, reproducing an experiment is one important approach that scientists use to gain confidence in their conclusions. (McNutt 2014)

Reproducibility has been consistently identified as an important component of scientific research. Although there is a consensus on the importance of reproducibility along with the other commonly used “R” terminology (i.e., replicability, repeatability, etc.), there is some disagreement on the usage of these terms, including conflicting definitions used by different parts of the research community. In this article, we explore the different definitions used in scientific literature (specifically pertaining to computational research), whether there is a need for a single standardized definition and provide an alternative based on nonfunctional requirements. We also describe the role of reproducibility (and other R’s) in scientific workflows.

Reproducibility Versus Replicability: Various and Potentially Conflicting Definitions

Reproducibility has been extensively discussed in recent literature and been consistently mentioned as an important component of scientific research (Bechhofer et al. 2013; Stodden 2015; Plesser 2018; Goble 2016; Jasny et al. 2011; Miceli 2019; Peng 2009; Peng 2011). The first appearance of the phrase “reproducible research” in a scholarly publication appears in an invited paper presented at the 1992 meeting of the Society of Exploration Geophysics (Claerbout and Karrenbach 1992; Barba 2018). While there is widespread

agreement on the importance of reproducibility along with the commonly used “R” terminology, e.g., replicability and repeatability. There is some disagreement in the usage of these terms and their semantics: often their definitions are contradictory.

Broadly reproducibility and replicability have been defined along four different perspectives (Barba 2018; National Academies of Sciences, Engineering, and Medicine 2019):

1. There is no distinction between the terms and hence they can be used interchangeably (Bechhofer et al. 2013; Open Science Collaboration 2012; Stodden 2015). A few papers that do not distinguish between reproducibility and replication define various levels of reproducibility. These levels are defined based on either the specific domain of research or the scientific approach (Stodden 2011, 2013; Goodman et al. 2016).
2. Reproducibility refers to instances in which the similar results are obtained by a different team using a different experimental setup. For computational research, this focuses on an independent group obtaining similar outcomes using artifacts they develop independently (Plesser 2018; Goble 2016; Jasny et al. 2011; Association for Computing Machinery Version 1.0 2016; Drummond 2009).
3. Reproducibility refers to the ability of an independent team of researchers to obtain consistent results (the same results) with the same experiment as the original study/experiment. For computational research, this implies using the same data resources, algorithms, source code, and programming environment as the original study (Claerbout and Karrenbach 1992; Miceli 2019; Peng 2009; Association for Computing Machinery Version 1.1 2020; Walters 2013).
4. A reproducibility spectrum starts with a minimum standard for judging scientific claims and ends with full replication of the same results using the same data resources, algorithms, source code, and programming environment (Peng 2011).

Papers using approaches 2 and 3 have contradictory interpretations of the terms reproducibility and replicability. What approach 2 refers to as reproducibility is referred to as replicability in approach 3, and vice versa.

Is One Definition Needed?

With at least the four conflicting definitions, those cited in the previous section and many uncited in other review papers, the commonly asked question is: “it is possible to define reproducibility across domains and applications?”. Varying

answers are provided to this question, ranging from either picking a side in the conflicting definitions to just presenting an overview of the situation regarding terminology and its history (Barba 2018; Baker 2016; Plesser 2018). There does however exist some commonality between all of the definitions and interpretations of reproducibility, replicability, repeatability, and the other “Rs.” The commonality observed in all the interpretations are **how** reproducibility and replicability (or however it is referred to across scientific literature) are achieved. Thus, instead of choosing one of the existing definitions, or creating a new one, focus would be on **how** to design and execute “a more complete” scientific experiment/study. While the Association for Computing Machinery’s definitions (Association for Computing Machinery Version 1.0 2016; Version 1.1 2020) have taken one side and then the other in terms of defining reproducibility over the years, they also provide one aspect of introspection into the final task at hand. ACM appropriately divides the tasks for reproducibility or replicability (hereafter simply referred to as the “Rs”) as follows:

1. Same Team, Same Experimental Setup (STSE)
2. Different Team, Same Experimental Setup (DTSE)
3. Different Team, Different Experimental Setup (DTDE)

If we focus on the nonfunctional requirements, the **how** of these “Rs” rather the **what** they are called, then a more general consensus in science at times even across domains is possible. That is because at the highest level, the “Rs” will always fall into one of the three categories mentioned above. How researchers obtain consistent results for each of these three categorical conditions then becomes a way to distinctly define the “Rs” for each approach, be it statistical, computational, empirical, or experimental. For example, in STSE data-driven experiments (statistical and computational), one step includes setting a seed to reproduce results in methods that include randomization, especially statistical sampling or generating layouts of visualizations, etc. And in DTSE setups, a common step is to store all intermediate results along with detailed documentation of any and all data processing that takes place in the experiments.

Workflows

Dramatic increases in the volume of data collected in the geosciences in the last few decades have meant that there has been a trend involving the exploration of new data-driven research methods. In the field of informatics, scientific explorations usually follow a set of steps to get from accessing the raw data, processing it, visualizing, analyzing, and deriving insights to answer scientific questions. Such a sequence of steps is called a workflow.

Workflow Systems and the “Rs”

Workflow systems (or workflow management systems) have become increasingly popular as a way of specifying and executing data-intensive analysis (Davidson et al. 2008). A workflow system is a problem-solving environment, tuned to a distributed and service-oriented computational infrastructure (Ludäscher et al. 2006). Unlike general-purpose scripting languages, workflow systems automatically record the means by which results are produced (Davidson et al. 2008).

During the initial development of the scientific workflow systems, the “Rs” were not explicitly included or highlighted as part of the desiderata for workflow systems (Ludäscher et al. 2006; McPhillips et al. 2009; Oinn et al. 2004; Hull et al. 2006). With the avalanche of data caused by the development of new experimental technologies, scientists increased their focus on data-driven and computational research methods (Cohen-Boulakia et al. 2017). This change in focus resulted in a renewed interest in the reproducibility (or replicability) of these experiments. Cohen-Boulakia et al. (2017) examined popular scientific workflow systems in the context of the “Rs.” The authors do this by defining the various “Rs” (what they call levels of reproducibility) and introducing a set of criteria that play a major role in the ability of a workflow systems to be “reproducibility-friendly” for a given computational or data-driven experiment. Popular workflow systems such as Taverna, Galaxy, OpenAlea, VisTrails, and Nextflow were then examined in the context of reproducibility and their respective limitations were documented.

More recent platforms that help capture scientific workflows including the creation, execution, and publication of computational and data-driven experiments, such as Whole Tale (Chard et al. 2020), consider and promote the “Rs” as a central component of workflow design and execution.

“Rs” in Scientific Workflows

In reviewing extant literature and practice but without explicit statements confirming, we assert “While workflow systems have become popular recently, there is still significant portion of the scientific community that do not use workflow systems.” In order to make sure data-driven experiments achieve any one of more of the “Rs,” decisions need to be made during the workflow design process. This is especially important in case the workflows need to be reproduced (or replicated, etc.) by researchers in different research domains (which is fairly common in interdisciplinary scientific explorations). Design decisions must be made based on the nonfunctional requirements of the scientific experiments, i.e., based on **how** to design a reproducible (or replicable) workflow.

The categories mentioned in Peng’s reproducibility spectrum (Peng 2011) are as follows (starting with* the least reproducible to the gold standard):

1. Publication + Code
2. Publication + Code + Data

3. Publication + Linked and executable code + Data
4. Full replication (i.e., gold standard)

Since, in informatics, achieving the “Rs” involve releasing data and code, we focus the next section on how to design workflows and share the required data and code to get closer to the gold standard.

As put forward by Sandve et al. (2013), “rules” for considerations in workflow design for ensuring the “Rs” are fulfilled in the experiments are:

1. Keep track of how every result was produced
2. Avoid manual data manipulation
3. Archive the exact versions of external programs used
4. Version control all custom scripts
5. Intermediate results need to be recorded, preferably in formats standardized throughout the experiment
6. Note seeds set for analyses that include randomness
7. Store raw data behind plots
8. Generate detailed analysis outputs, allowing for inspections for summarization plots
9. Connect text statements to underlying results
10. Provide public access to scripts, runs, and results

Each of the above ten rules capture a specific aspect of the “Rs” and present what is needed in terms of handling the information and tracking procedures in order to release data and code for a scientific experiment (Sandve et al. 2013).

Releasing Data

Data used in an experiment/study can be obtained in a number of ways. It can range anywhere from being generated by the research team conducting the experiment/study, to being directly retrieved from an existing repository without making any changes to the data parameters. Usually, the process lies somewhere in the middle, where people need to retrieve the specific data required for their experiments from a few data repositories and make small additions or modifications to the data. In all of these cases though, there are some good practices to be followed so that the experiments achieve the criteria for the “Rs.” These steps are as follows:

1. Credit the creators of the data resource(s) being used. This may include citing a data paper, a dataset PID, or any citation/accreditation method stated by the dataset creators.
2. As the data is being processed, save versions of the dataset that are created during the execution of the workflow.
3. Document the code used for processing the data.

The datasets used in the experiment/study should be included along with the publication in many ways, such as:

- Releasing to data repositories like Zenodo, Dryad, Dataverse, Figshare, etc. (Assante et al. 2016).
- Adding data files as part of the supplementary materials for the main publication. Some journals host data associated with their publications.
- Publishing data papers in scientific journals or the more nascent, but specific, data journals (Callaghan et al. 2012).

Releasing Code

Releasing code is not only a central aspect of achieving the “Rs” for an experiment/study, but also helps highlight the algorithms used, and decisions made throughout the research workflow. Similar to releasing data, code has been released in many ways:

- Adding the code/script files used in experiment along with the papers in the supplementary materials.
- Releasing code files and scripts as a single research product to repositories like Zenodo, Figshare, etc. (Note: Unlike Dryad and Dataverse, Zenodo and Figshare are repositories that accept any research product, including code, presentations, videos, images, and softwares).
- Saving executable code for an experiment in an interactive environment like Jupyter or R notebooks, which are designed to support the workflow of scientific computing (Kluyver et al. 2016). Such notebooks can then be published to a repository for download and use or be linked to a working JupyterHub server. Services like mybinder.org enable sharing of projects containing live notebooks including a computational environment where user can execute code (Jupyter et al. 2018; Kluyver et al. 2016).

Writing code or scripts for an experiment/study are one of the steps where the research workflow is clearly visible. Hence, it is important to provide documentation with detailed instructions on how the individual code files tie into the larger research workflow. And also, correctly archive all the outputs (according to the rules put forward by Sandve et al. (2013)) required to achieve the “Rs.”

Summary

In conclusion, even though there are some disagreements on defining reproducibility or replicability, we observe some consensus on *how* to achieve reproducibility or replicability (the “Rs”). Also, there has been progress in the area of *data* and *code* releasing in the recent years which not only contributes to the “Rs” of scientific research, but also provides researchers who work in these areas much needed credit and an incentive to share their work outside of a scientific journal paper.

Bibliography

- Assante M, Candela L, Castelli D, Tani A (2016) Are scientific data repositories coping with research data publishing? *Data Sci J* 15:6. <https://doi.org/10.5334/dsj-2016-006>
- Association for Computing Machinery (2016) Artifact review and badging. Available online at: <https://www.acm.org/publications/policies/artifact-review-badging>
- Association for Computing Machinery (2020) Artifact review and badging. Available online at: <https://www.acm.org/publications/policies/artifact-review-and-badging-current>
- Baker M (2016) Reproducibility crisis. *Nature* 533(26):353–366
- Barba LA (2018) Terminologies for reproducible research. CoRR, abs/1802.03311. Retrieved from <http://arxiv.org/abs/1802.03311>
- Bechhofer S, Buchan I, De Roure D, Missier P, Ainsworth J, Bhagat J, Couch P, Cruickshank D, Delderfield M, Dunlop I, Gamble M, Michaelides D, Owen S, Newman D, Sufi S, Goble C (2013) Why linked data is not enough for scientists. *Futur Gener Comput Syst* 29(2):599–611. <https://doi.org/10.1016/j.future.2011.08.004>
- Callaghan S, Donegan S, Pepler S, Thorley M, Cunningham N, Kirsch P, Ault L, Bell P, Bowie R, Leadbetter A, Lowry R, Moncoiffé G, Harrison K, Smith-Haddon B, Weatherby A, Wright D (2012) Making data a first class scientific output: data citation and publication by NERC’s Environmental Data Centres. *Int J Digit Curation* 7(1): 107–113. <https://doi.org/10.2218/ijdc.v7i1.218>
- Chard K, Gaffney N, Hategan M, Kowalik K, Ludäscher B, McPhillips T, Nabrzyski J, Stodden V, Taylor I, Thelen T, Turk MJ, & Willis C. (2020). Toward Enabling Reproducibility for Data-Intensive Research Using the Whole Tale Platform [JB]. *Advances in Parallel Computing*, 36(Parallel Computing: Technology Trends), 766–778. <https://doi.org/10.3233/APC200107>
- Claerbout JF, Karrenbach M (1992) Electronic documents give reproducible research a new meaning. In: SEG technical program expanded abstracts 1992. Society of Exploration Geophysicists, pp 601–604. <https://doi.org/10.1190/1.1822162>
- Cohen-Boulakia S, Belhajjame K, Collin O, Chopard J, Froidevaux C, Gaignard A, Hinsin K, Larmande P, Bras YL, Lemoine F, Mareuil F, Ménager H, Pradal C, Blanchet C (2017) Scientific workflows for computational reproducibility in the life sciences: status, challenges and opportunities. *Futur Gener Comput Syst* 75:284–298. <https://doi.org/10.1016/j.future.2017.01.012>
- Davidson S, Cohen-Boulakia S, Eyal A, Ludäscher B, McPhillips T, Bowers S, Anand MK, Freire J (2008) Provenance and scientific workflows: challenges and opportunities. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pp 1345–1350
- Drummond C (2009) Replicability is not reproducibility: nor is it good science. [Conference Paper] *Cogprints ID* 7691. <http://cogprints.org/7691/>
- Goble C (2016) What is reproducibility? The R* Brouhaha. In: Reproducible open science workshop, 20th international conference on theory and practice of digital libraries. Hannover, Germany. Retrieved from <http://repscience2016.research-infrastructures.eu/img/CaroleGoble-ReproScience2016v2.pdf>
- Goodman SN, Fanelli D, Ioannidis JPA (2016) What does research reproducibility mean? *Sci Transl Med* 8(341):341ps12 LP–341ps12. <https://doi.org/10.1126/scitranslmed.aaf5027>
- Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, Oinn T (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res* 34:W729–W732. <https://doi.org/10.1093/nar/gkl320>. (Web Server)
- Jasny BR, Chin G, Chong L, Vignieri S (2011) Data replication & reproducibility. Again, and again, and again....Introduction. *Science* 334(6060):1225. <https://doi.org/10.1126/science.334.6060.1225>. PMID: 22144612

- Jupyter P, Bussonnier M, Forde J, Freeman J, Granger B, Head T, Holdgraf C, Kelley K, Nalvarte G, Osheroff A, Pacer M, Panda Y, Perez F, Ragan-Kelley B, Willing C (2018) Binder 2.0 – reproducible, interactive, sharable environments for science at scale. In: Proceedings of the 17th Python in science conference. <https://doi.org/10.2580/majora-4af1f417-011>
- Kluyver T, Ragan-Kelley B, Perez F, Granger B, Bussonnier M, Frederic J, Kelley K, Hamrick J, Grout J, Corlay S, Ivanov P, Avila D, Abdalla S, Willing C, & Jupyter Development Team. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. Stand Alone, 0(Positioning and Power in Academic Publishing: Players, Agents and Agendas), 87–90. <https://doi.org/10.3233/978-1-61499-649-1-87>
- Ludäscher B, Altintas I, Berkley C, Higgins D, Jaeger E, Jones M, Lee EA, Tao J, Zhao Y (2006) Scientific workflow management and the Kepler system. *Concurrency Comput: Pract Exp* 18(10):1039–1065. <https://doi.org/10.1002/cpe.994>
- McNutt M (2014) Reproducibility. *Science* 343(6168):229LP–229. <https://doi.org/10.1126/science.1250475>
- McPhillips T, Bowers S, Zinn D, Ludäscher B (2009) Scientific workflow design for mere mortals. *Futur Gener Comput Syst* 25(5): 541–551. <https://doi.org/10.1016/j.future.2008.06.013>
- Miceli S (2019) Reproducibility and replicability in research, September 3. <https://www.nationalacademies.org/news/2019/09/reproducibility-and-replicability-in-research>. Retrieved 12 Oct 2020
- National Academies of Sciences, Engineering, and Medicine (2019) Reproducibility and replicability in science. The National Academies Press, Washington, DC. <https://doi.org/10.17226/25303>
- Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 20(17):3045–3054. <https://doi.org/10.1093/bioinformatics/bth361>
- Open Science Collaboration (2012) An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspect Psychol Sci* 7(6):657–660. <https://doi.org/10.1177/1745691612462588>
- Peng RD (2009) Reproducible research and Biostatistics. *Biostatistics* 10(3):405–408. <https://doi.org/10.1093/biostatistics/kxp014>
- Peng RD (2011) Reproducible research in computational science. *Science* 334(6060):1226–1227. <https://doi.org/10.1126/science.1213847>
- Plesser HE (2018) Reproducibility vs. replicability: a brief history of a confused terminology. *Front Neuroinform* 11:76. <https://doi.org/10.3389/fninf.2017.00076>
- Sandve GK, Nekrutenko A, Taylor J, Hovig E (2013) Ten simple rules for reproducible computational research. *PLoS Comput Biol* 9(10):e1003285. <https://doi.org/10.1371/journal.pcbi.1003285>
- Stodden V (2011) Trust your science? Open your data and code. *Amstat News*, July 1. <http://magazine.amstat.org/blog/2011/07/01/trust-your-science/>
- Stodden V (2013) Resolving irreproducibility in empirical and computational research *IMS Bull.* November 17. <http://bulletin.imstat.org/2013/11/resolving-irreproducibility-in-empirical-and-computationalresearch/>
- Stodden V (2015) Reproducing statistical results. *Annu Rev Stat Appl* 2(1):1–19. <https://doi.org/10.1146/annurev-statistics-010814-020127>
- Walters WP (2013) Modeling, informatics, and the quest for reproducibility. *J Chem Inf Model* 53(7):1529–1530. <https://doi.org/10.1021/ci400197w>

Rescaled Range Analysis

Gabor Korvin

Earth Sciences Department, King Fahd University of Petroleum and Minerals, Dhahran, Kingdom of Saudi Arabia

Definition

Rescaled-range analysis (R/S analysis) is a statistical method to detect and quantify long-term correlations in records of natural processes. It is a method to estimate the *Hurst exponent* H , which describes how the *rescaled range* of a *random process* X grows with time. If $X = \{\xi_t\}$ is a random process with mean value $\langle \xi \rangle$, its *range* R is the difference between the maximum and minimum of the cumulative X values:

$$R(l) = \max_{1 \leq t \leq l} X(t, l) - \min_{1 \leq t \leq l} X(t, l)$$

where l is length of the period considered and t is discretized time. Let $S(l) = \left\{ \frac{1}{l} \sum_{t=1}^l [\xi_t - \langle \xi \rangle]^2 \right\}^{1/2}$ be the standard deviation, then for many real-world processes the ratio R/S, called *rescaled range*, grows with time l as $R(l)/S(l) \propto l^H$. The exponent H is between 0 and 1, it is called “*Hurst exponent*” (to honor the British hydrologist H. E. Hurst, 1880–1978). When $0.5 < H < 1$, the behavior is *persistent* (the observed trend continues in the future), $0 < H < 0.5$ implies an *antipersistent* behavior, for $H = 0.5$ there is no long-range dependence between past and the future.

Rescaled Range and Its Scaling

Rescaled Range Analysis, the phenomenon later called *Hurst effect*, and the celebrated *Hurst exponent* H were discovered by the British hydrologist H. E. Hurst who worked in Egypt on designing an ideal reservoir. He used the following model (Hurst 1951; Korvin 1992: 328–329): Let X_t ; $t = 1, 2, \dots, N$ denote the net inflow of water into the reservoir in $N \gg 1$ consecutive years, $\bar{X}_N = \frac{1}{N} \sum_{t=1}^N X_t$ the average influx over the period, and suppose that every year we release this average period $\bar{X}_N$. In a given year t , $1 \leq t \leq N$, the reservoir will contain $X(t, N) = \sum_{i=1}^t (X_i - \bar{X}_N) = \sum_{i=1}^t X_i - t\bar{X}_N$ volume of water. If the reservoir is designed for N years, it must have a sufficiently large storage capacity $C(N)$ to assure that it never

overflows or gets dry, that is $C(N)$ must be larger than the difference $R(N)$ between the maximum and minimum amounts of water ever contained in the reservoir: $C(N) > R(N)$

$$R(N) = \max_{1 \leq t \leq N} X(t, N) - \min_{1 \leq t \leq N} X(t, N) \quad (\text{Eq. 1}).$$

$R(N)$ is called the *range* of the random variables X_i ; $t = 1, 2, \dots, N$, it depends on N and on the standard deviation $S(N) =$

$$\left\{ \frac{1}{N} \sum_{t=1}^N [X_t - \bar{X}_N]^2 \right\}^{1/2}. \text{ To eliminate this dependence, Hurst}$$

(1951) introduced the *rescaled range* $R^*(N) = \frac{R(N)}{S(N)}$ (Eq. 2). To understand how $R^*(N)$ changes with N , Hurst considered tossing k coins N times, and defined the random variable X_t as the difference between the number of heads and tails in the t -th experiment. For this process $R(N) = (\frac{\pi}{2}kN)^{1/2} - 1$; $S = k^{1/2}$, that is $R^*(N) = \frac{R(N)}{S(N)} \propto (\frac{\pi}{2}N)^{1/2}$ (Eqs. 3a, b, c; Korvin 1992: 329). Hurst analyzed some 120 geophysical and economic time series (river discharges, rainfall, water level, temperature, wheat prices, etc.; see Korvin 1992, for modern examples Hamed 2007) and found that R/S approximately scales as $R^*(N) = \frac{R(N)}{S(N)} \propto N^H$ (Eq. 4), where the exponent H always lied in the range $H = 0.73 \pm 0.09$, that is, significantly differed from the theoretically expected $H = 1/2$. Later, the exponent H was called *Hurst exponent*, the non-trivial scaling with exponents $H \neq 1/2$ is the *Hurst phenomenon* or *Hurst effect*, and the determination of H is the task of the *Rescaled Range* (R/S) analysis.

Random Processes with Hurst Exponent $H \neq 1/2$

As seen (Eq. 3c), the expected *Rescaled Range* of a sequence $\{X_t\}$, $t = 1, \dots, N$, of $N \gg 1$ random coin-tossing grows as $E[R^*(N)] = E\left[\frac{R(N)}{S(N)}\right] \propto (\frac{\pi}{2}N)^{1/2}$ (Eq. 4), this also holds asymptotically for any sequence of *identically and independently distributed (iid) normal variables* $E\left(\frac{R}{S}\right) = \sqrt{\frac{\pi}{2}}N^{0.5}$ (Korvin 1992: 329). Anis and Lloyd (1976) expressed the *exact* expected value of the rescaled range for *iid* normal vari-

$$\text{ables as } E\left[\frac{R(n)}{S(n)}\right] = \begin{cases} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n}{2}\right)} \sum_{i=1}^{n-1} \sqrt{\frac{n-i}{i}} & \text{for } n \leq 340 \\ \frac{1}{\sqrt{n\pi}} \sum_{i=1}^{n-1} \sqrt{\frac{n-i}{i}} & \text{for } n > 340 \end{cases}$$

(Eq. 5a, b), where $\Gamma(x)$ is the gamma function. Hamed (2007) noted that substituting $x = N/2$, $a = 1/2$, $b = 0$ into the known

$$\text{relation } \lim_{x \rightarrow \infty} x^{b-a} \frac{\Gamma(x+a)}{\Gamma(x+b)} = 1 \quad \text{yields } \frac{\Gamma(n/2 - 1/2)}{\Gamma(n/2)} \approx \left(\frac{2}{n}\right)^{1/2}$$

(Eq. 6a) in the first factor in the right-hand side of

Eq. (5a) and, while $\sum_{r=1}^{N-1} \sqrt{\frac{N-r}{r}} \sim \frac{\pi N}{2}$ for $N \gg 1$, for small values

of N the expression $\sum_{r=1}^{N-1} \sqrt{\frac{N-r}{r}} \approx \frac{\pi N}{2} - 1.46N^{0.5}$ (Eq. 6b) gives

better approximation. Substituting Eqs. (6 a, b) to Eq. (5), Hamed got the improved estimate $E\left(\frac{R}{S}\right) \approx \sqrt{\frac{\pi}{2}}n^{0.5} - 1.164$ (Eq. 7).

There have been many speculations about what kind of processes exhibit the anomalous scaling $R/S \propto N^H$ with $H \neq 1/2$ (Korvin 1992). A feasible model is the *fractional Brownian noise (fBn)* model of Mandelbrot and Van Ness (1968). They defined first *fractional Brownian motion (fBm)* of Hurst exponent H which is a *non-stationary self-affine random process* $B_H(t) = \frac{1}{\Gamma(H+0.5)} \left\{ \int_{-\infty}^0 [t-l]^{H-0.5} - [l]^{H-0.5} \right\} dB(l) + \int_0^t [t-l]^{H-0.5} dB(l)$ (Eq. 8a)

where $B(t)$ is a Gaussian process with zero mean and σ^2 variance, and proved that the continuous-time process $B_H(t)$ satisfies: (i) its increments are stationary; (ii) $B_H(t)$ is Gaussian, $\langle B_H(t) \rangle = 0$; $\langle [B_H(1)]^2 \rangle = \sigma^2$ (Eq. 8b); (iii) $B_H(t)$ is statistically invariant (*self-affine*) under the transformation $\{\tau \rightarrow \lambda\tau; B_H \rightarrow \lambda^H B_H\}$ (Eq. 8c). Self-affinity implies that $B_H(\lambda t) - B_H(0) = \lambda^H \{B_H(t) - B_H(0)\}$ (Eq. 8d). (The process $B_{0.5}(t)$ is the ordinary *Brownian motion*). It is easy to prove that $\langle [B_H(\tau + T) - B_H(\tau)]^2 \rangle \propto \sigma^2 T^{2H}$ (Eq. 9). Next, they constructed the process $X_t^{(H)}$ from the increments of $B_H(t)$, as $X_t^{(H)} = B_H(t) - B_H(t-1)$ (Eq. 10) and called it *fractional Brownian noise (fBn)* with exponent H (Mandelbrot and Van Ness 1968), it is stationary, of zero mean and variance σ^2 . By simple algebra (using Eqs. 8 and 9) the correlation between $X_1^{(H)}$ and $X_K^{(H)}$ is: $\langle X_1^{(H)} \cdot X_K^{(H)} \rangle \propto \sigma^2 H(2H-1)K^{2H-2}$ if $K \gg 1$ and $H \neq 1/2$ (Eq. 11a); and $\langle X_1^{(H)} \cdot X_K^{(H)} \rangle = 0$ for $\forall K \neq 1$ if $H = 1/2$ (Eq. 11b). In Eq. (11a) the correlations are always positive, and fall off as a power of K . Equation (11b) (for $H = 1/2$) describes the *uncorrelated white noise*. Let us prove that $\{X_t\} = \{X_t^{(H)}\}$ satisfies Hurst's Law, $\frac{R(N)}{S(N)} \propto N^H$. If for a process $\{X_i\}$ we denote by $X^*(t)$ the partial

sum $\sum_{i=1}^t X_i$, then the range of $\{X_i\}$ can be re-written in terms of $X^*(t)$ as $R(N) = \left\{ \max_{0 \leq t \leq N} - \min_{0 \leq t \leq N} \right\} [X^*(t) - \frac{t}{N} X^*(N)]$ (Eq. 12).

In particular, for the process $\{X_t^{(H)}\}$, Eq. (10) implies that $X^{*(H)}(t) = B_H(t)$ (Eq. 13). By the affine property, Eq. (8), changing the variable t to $\tau = t/N$ in Eq. (12), we obtain $R(N) =$

$$\left\{ \max_{0 \leq \tau \leq N} - \min_{0 \leq \tau \leq N} \right\} [N^H B_H(\tau) - \tau N^H B_H(1)] = \text{const} \cdot N^H,$$

and because $S(N) \propto \sigma$ for the *fBn* $\{X_i\}$, Hurst's rule $\frac{R(N)}{S(N)} \propto N^H$ (Eq. 14) holds indeed.

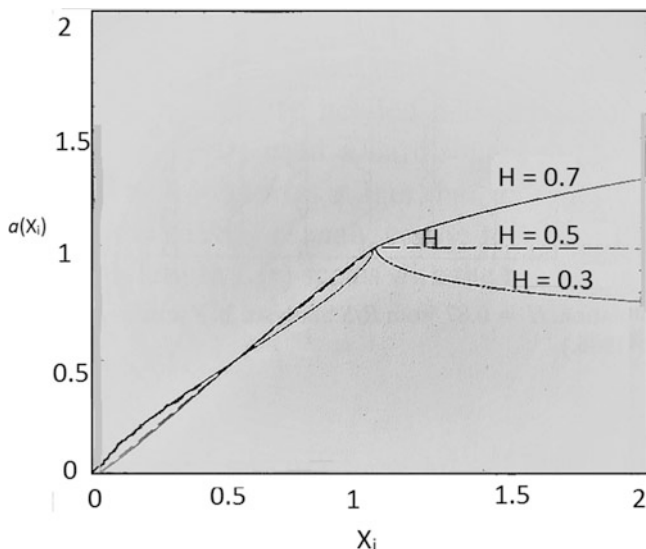
The Hurst Exponent H as an Indicator of Long-Term Behavior

Let $Z(x)$ be an fBm function which is only known at the two endpoints $x = 0$ and $x = L$ of an interval $[0, L]$, and suppose we need to estimate $Z(x)$ inside the interval $0 \leq x \leq L$ (“interpolation”) and for values $L \leq x$ (“extrapolation,” “prediction”) using the *optimal unbiased linear combination* $Z(x) = \lambda_1 Z(0) + \lambda_2 Z(L)$ (Mandelbrot and Van Ness 1968 Hewett 1986 Korvin 1992: 336–337). The coefficients λ_1 & λ_2 are found from the system

$$\left. \begin{aligned} \lambda_2 \gamma(0, L) + \mu &= \gamma(0, x) \\ \lambda_1 \gamma(0, L) + \mu &= \gamma(L, x) \\ \lambda_1 + \lambda_2 &= 1 \end{aligned} \right\} \text{ (Eq. 15), where } \mu$$

is a *Lagrange parameter*, and $\gamma(x, y) = \frac{1}{2} \langle |Z(x) - Z(y)|^2 \rangle$ the *semivariogram*. By Eq. (9) for an fBm with Hurst exponent H and variance σ^2 we have $\gamma(x, y) = \frac{1}{2} \sigma^2 |x - y|^{2H}$ (Eq. 16), the system of Eq. (15) is easily solved and yields the interpolation-extrapolation formula $Z^*(x) = Z(L) \times \left\{ \frac{1}{2} [1 - |1 - \xi|^{2H} + \xi^{2H}] \right\} = Z(L) \cdot Q(\xi)$, where $\xi = \frac{x}{L}$ (Eq. 17).

The extrapolation-interpolation function $Q(\xi)$ is shown in Fig. 1 for different values of H . We have $Q(0) = 0$, $Q(1) = 1$, for $0 \leq \xi \leq 1$ the function is almost linear. For $\xi > 1$, note that $|1 - \xi| = (\xi - 1)$ and $\xi^{2H} - (\xi - 1)^{2H} \approx \frac{d}{d\xi} \xi^{2H} \sim \xi^{2H-1}$ that is for $x > L$ the function $Z(x)$ is expected to continue as $Z(x) \propto Z(L) \cdot \left(\frac{x}{L}\right)^{2H-1}$ (Eq. 18). For $H = 1/2$ the best prediction is the last functional value; for $H > 1/2$ the trend over the interval will be continued (“persistence”), for $H < 1/2$ the trend will be reversed and the predicted value tends to the mean over the interval $[0, L]$ (“antipersistence”).



Rescaled Range Analysis, Fig. 1 The interpolation-extrapolation function $Q(\xi) = Q(x/L)$ for different values of H . (From Korvin (1992: 337, Fig. 5.3, after Hewett 1986))

Traditional R/S Analysis

Rescaled Range (R/S) analysis is used to find H for a time-series. It was invented and widely applied by Hurst (1951) and, in its classic form, consists of the following steps (Korvin 1992). (i) A time series of length $N \gg 1$ is divided into a great number of shorter sub-intervals of respective lengths $n = n_1, n_2, n_3, \dots$ where $n_1 = N > n_2 \geq n_3 \geq \dots$ (Eq. 19a). One way to do this is to select, in turn, $n = N, N/2, N/4, \dots$ or, more generally, $n = N, N/v, N/v^2, \dots$ (Eq. 19b); (ii) for a partial time series X of length n ; $\{x_1, x_2, \dots, x_n\}$ we calculate its mean: $\mu = \frac{1}{n} \sum_{i=1}^n x_i$ and (iii) subtract it from all data to get *mean-adjusted values* $y_i = x_i - \mu$ ($i = 1, 2, \dots, n$); (iv) we compute the cumulative sums $Z_t = \sum_{i=0}^t y_i$ ($t = 1, 2, \dots, n$); (v) calculate the range $R(n) = \max(Z_1, Z_2, \dots, Z_n) - \min(Z_1, Z_2, \dots, Z_n)$; (vi) determine the *Standard Deviation* $S(n) = \sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2}$; (vi) The process is repeated for each sub-period having the same length n , yielding the average rescaled ranges $\left\langle \frac{R(n)}{S(n)} \right\rangle$ over all sub-intervals of length n . (vii) The *Hurst exponent* is estimated by fitting the power law $\left\langle \frac{R(n)}{S(n)} \right\rangle = C \cdot n^H$ (Eq. 20) to the data. Least squares regression on the logarithms of each side of Eq. (20) yields $\log \frac{R(n)}{S(n)} \approx \log C + H \log n$ (Eq. 21a), the sign “ $\approx$ ” of approximate equality refers to the problematic issues and inevitable errors of logarithmic fitting. (viii) If the lengths of subintervals are powers of some integer v , as in Eq. (19b), it is recommended (Kriřtoufek 2010) to use base- v logarithm, that is $\log_v \frac{R(n)}{S(n)} \approx \log_v C + H \log_v n$ (Eq. 21b).

Recent Improvements and Variants

Detrended Hurst Analysis

Detrended Hurst analysis, also called *Detrended Fluctuation Analysis* DFA (Bassingthwaighte and Raymond 1994; Kriřtoufek 2010) uses a different measure of dispersion – *squared fluctuations* around the *trend of the signal*, rather than around the local mean, as in ordinary R/S analysis. Because we de-trend the sub-periods, the method works for non-stationary time-series, contrary to R/S. The procedure is labeled DFA-0, DFA-1, and DFA-2 according to the order of polynomials fitting the trend. The polynomials are seldom higher than second order (Kriřtoufek 2010).

The procedure starts as in R/S: the whole series is divided into sub-periods of length v (with increasing lengths v), then for each subperiod $\Pi_i = \{\xi_1, \dots, \xi_v\}$ a polynomial fit $P_{v,l}^{(i)}(t)$ is established (where l is the order of the polynomial, v the

length of the sub-period). The *de-trended signal* $Y_{v,l}^{(i)}(t) = \xi_t - P_{v,l}^{(i)}(t); t = 1, 2, \dots, v$ (Eq. 22) has a *fluctuation* $F_{(i)}^2(v, l) = \frac{1}{v} \sum_{t=1}^v \left(Y_{v,l}^{(i)}(t)\right)^2$ (Eq. 23). We do not divide the fluctuations by the variance but take their average over all subperiods Π_i of length v , in order to get $F_{DFA}^2(v, l)$, which scales as $F_{DFA}^2(v, l) \sim v^{2H(l)}$ (Eq. 24). The argument “ l ” in the estimated H indicates that it might depend on the degree of the fitting polynomial. Finally, an ordinary least squares regression on the logarithms of both sides of Eq. (24) yields $H(l)$ as slope of the straight line $\log F_{DFA}(v, l) \approx \log c + H(l) \log v$ (Eq. 25).

The Structure Function Method

The *Structure Function Method* utilizes that the q th-order structure functions $S_q(\tau) = \langle |W(t_i + \tau) - W(t_i)|^q \rangle$ of processes $W(t)$ showing the Hurst effect scale as $S_q(\tau) \sim C_q \cdot \tau^{qH(q)}$ (Eq. 26). The exponent $H(q)$ is determined by plotting $S_q(\tau)$ vs. τ on a log-log plot, and finding the slope $qH(q)$. Gilmore et al. (2002), in their analysis of *plasma turbulence*, eliminated the effect of q on $H(q)$ by using $q = 0.5, 1.0, 1.5, \dots, 5.0$ and averaging the $H(q)$ exponents.

A Modified Scaling Law

Hamed (2007) noted that traditional R/S analysis gives biased estimates of the Hurst exponent for finite samples, and more precisely computed the exact expression of Anis and Lloyd (1976) for the expected rescaled range of independent normal summands,

$$E\left[\frac{R(n)}{S(n)}\right] = \begin{cases} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n}{2}\right)} \sum_{i=1}^{n-1} \sqrt{\frac{n-i}{i}} & \text{for } n \leq 340 \\ \frac{1}{\sqrt{n\pi}} \sum_{i=1}^{n-1} \sqrt{\frac{n-i}{i}} & \text{for } n > 340 \end{cases}.$$

He derived, instead of the customary asymptotic formula $E\left(\frac{R}{S}\right) = \sqrt{\frac{\pi}{2}} N^{0.5}$, an improved scaling law $E\left(\frac{R}{S}\right) \approx \sqrt{\frac{\pi}{2}} n N^{0.5} - 1.164$. Inspired by this equation, he proposed the $\frac{R(n)}{S(n)}$ data should be non-linearly fitted as $E\left(\frac{R(n)}{S(n)}\right) = an^H + b$ (Eq. 27). He applied this to a large set of data and found: (a) For the annual mean temperature data from 56 Historical Climatological Network (HCN) stations in Midwest USA the proposed estimator gave an average estimated Hurst coefficient 0.645 compared to 0.725 with the old method (−11% change); (b) for annual rainfall data from 60 HCN stations in Midwest USA it gave $H = 0.568$ instead 0.650 (−13% change); (c) for annual river flow data from 49 U.S. Geological Survey (USGS) river stations in Midwest USA $H = 0.631$ as compared to 0.621 (+2% change); for tree-ring data from

88 locations over the world $H = 0.596$ increased to 0.639 (+7% change).

Anis-Lloyd Corrected R/S Hurst Exponent

Anis and Lloyd (1976) derived different theoretical expressions for the expected rescaled range of independent normal summands for small and large values of n :

$$E\left[\frac{R(n)}{S(n)}\right] = \begin{cases} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n}{2}\right)} \sum_{i=1}^{n-1} \sqrt{\frac{n-i}{i}} & \text{for } n \leq 340 \\ \frac{1}{\sqrt{n\pi}} \sum_{i=1}^{n-1} \sqrt{\frac{n-i}{i}} & \text{for } n > 340 \end{cases}.$$

The *Anis-Lloyd corrected R/S Hurst exponent* is calculated as 0.5 plus the slope of $\frac{R(n)}{S(n)} - E\left[\frac{R(n)}{S(n)}\right]$. A slight modification of the original Anis-Lloyd formula by a factor $(n - 1/2)/n$,

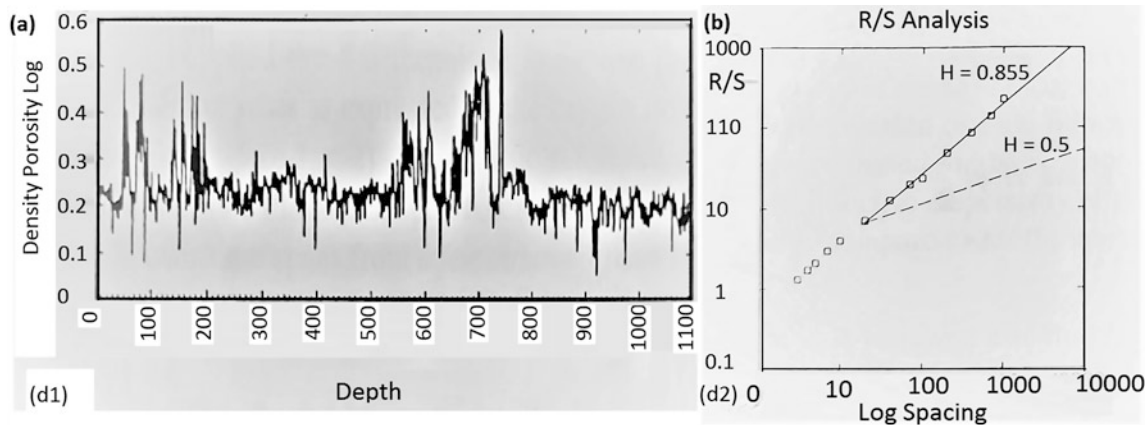
$$E\left[\frac{R(n)}{S(n)}\right] = \begin{cases} \frac{n-1/2}{n} \cdot \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n}{2}\right)} \sum_{i=1}^{n-1} \sqrt{\frac{n-i}{i}} & \text{for } n \leq 340 \\ \frac{n-1/2}{n} \cdot \frac{1}{\sqrt{n\pi}} \sum_{i=1}^{n-1} \sqrt{\frac{n-i}{i}} & \text{for } n > 340 \end{cases} \quad (\text{Eq. 28})$$

gives further improvements for small n (Weron 2002).

Estimating H from the Power Spectrum

The method (Weron 2002) utilizes that (by Eq. 11) an fBn with exponent H has a power spectrum $G(f) \propto f^{-2H+1} = f^{-\beta}$ (Eq. 29) with $\beta = 2H - 1$ (Eq. 30) (see Korvin 1992: 341, Eq. 5.1.2.2), f is frequency. If $P(f)$ is power spectrum of a fractal process, its Hurst exponent H is obtained from the $\log P(f)$ versus $\log f$ plot. A straight line with a negative slope ($-\beta$) is obtained for a *self-affine* process, and the Hurst exponent is found as $H = (\beta + 1)/2$ (Eq. 31).

A nice example is Hewett's study (1986; Korvin 1992: 333, 335) who analyzed the *porosity values* derived from a *gamma-gamma* (“density”) log through 1000 feet of sandstone (Fig. 2a). R/S analysis of the porosity log resulted in $H = 0.85$. As an independent check, Hewett computed the power spectrum of the log, established a fall-off with frequency $\propto f^{-\beta}$, with $\beta = 0.71 \approx 2H - 1$. Note that the relation (Eq. 31) had been rigorously proved for fBn processes only (Mandelbrot and Van Ness 1968), and is not necessarily valid for an arbitrary time series.



Rescaled Range Analysis, Fig. 2 (a) Density-derived porosity log through a sandstone formation, depth in feet; (b) its R/S analysis, dashed line shows the $H = 2$ case. (From Korvin (1992: 333, Fig. 5.2 d1 and d2, after Hewett 1986))

Periodogram Regression

This is another spectral method which estimates the so-called *fractional differencing parameter* d from the *periodogram*, and from the relation $H = d + 0.5$ (Eq. 32) established for a fractional Gaussian noise with Hurst exponent H , gets H (Weron 2002). For the data $\{X_1, X_2, \dots, X_L\}$ the *periodogram* (analogue of the *spectral density* for sampled data) is computed as $P_L(\omega_k) = \frac{1}{L} \left| \sum_{t=1}^L x_t \cdot e^{-2\pi j(t-1)\omega_k} \right|^2$; $j = \sqrt{-1}$; $\omega_k = \frac{k}{L}$; $k = 1, \dots, L/2$ (Eq. 33), then we run a linear regression $\log[P_L(\omega_k)] \approx a - d \cdot \log\{4\sin^2(\omega_k/2)\}$ (Eq. 34) at sufficiently low frequencies ω_k , $k = 1, 2, \dots, K \leq L/2$ (Eq. 35) for which the logarithmic slope is approximately linear. The choice of K in Eq. (35) is crucial, in practice $L^{0.2} \leq K \leq L^{0.5}$. Apparently, this is the only method with known *asymptotic properties* for finding H (Weron 2002: 289).

Practical Issues

There is a consensus among practitioners (Bassingthwaight and Raymond 1994; Gilmore et al. 2002; Hamed 2007; Křišťoufek 2010) that if the data is of short duration, contains gaps or bursts of noises, or it is non-stationary, then the calculation of the H by R/S analysis is not reliable. When tested on computer-generated fBn signals of known H it has been found that R/S gives biased estimates of H : too low for $H > 0.72$, and too high for $H < 0.72$, and the method has poor convergence properties, it requires about 2000 points for 5% accuracy and 200 points for 10% accuracy (Bassingthwaight and Raymond 1994). According to Gilmore et al. (2002), the R/S method accurately determines

H in plasma fluctuation data provided a long enough record is used, but lags must be greater than five to ten times the *autocorrelation time* τ_L . They claim the *structure function* only requires lags $\tau \geq 1 \times \tau_L$. (For a random process $X(t)$ the *autocorrelation time* is the lag where the autocorrelation function falls off to $(1/e)$ -times its peak value, $|R_{XX}(\tau)| \leq \frac{1}{e} R_{XX}(0)$ for $\forall \tau \geq \tau_L$)

Is it important to remove *local trends* in R/S analysis? Hurst (1951) himself applied trend correction. Bassingthwaight and Raymond (1994), who tested the performance of R/S analysis on *fractional Brownian noise* of known H , reported that the locally trend-corrected method gives better estimates of H when $H \geq 0.5$, that is, for persistent signals with positive correlations between neighboring values.

Concluding Remarks

Since introduced by Hurst (1951), *Rescaled Range (R/S) Analysis* has been used to find long-term dependence in hydrologic and geophysical records and other real-world data. R/S Analysis provides the *Hurst exponent* which characterizes long-term correlations in stochastic processes. There are persistent ($H > 0.5$) and antipersistent ($H < 0.5$) processes, for the white noise $H = 0.5$. Natural processes with Hurst exponent H can be modeled by *fractional Brownian noise (fBn)* of Hurst exponent H . There are many variants of R/S Analysis, here we discussed (i) *Traditional R/S analysis*; (ii) *Detrended Fluctuation Analysis DFA*; (iii) a *modified scaling law*; (iv) the *Anis-Lloyd corrected R/S Hurst exponent*; (v) *Estimating H from the power spectrum*; (vi) *Periodogram regression*. To find their relative merits,

speed up their convergence, and determine their asymptotic properties need future research.

Cross-References

- ▶ Allometric Power Laws
- ▶ Autocorrelation
- ▶ Computational Geoscience
- ▶ Constrained Optimization
- ▶ Exploratory Data Analysis
- ▶ Geocomputing
- ▶ Geomathematics
- ▶ Geostatistics
- ▶ Hurst Exponent
- ▶ Interpolation
- ▶ Kriging
- ▶ Mandelbrot, Benoit B.
- ▶ Mathematical Geosciences
- ▶ Porosity
- ▶ Scaling and Scale Invariance
- ▶ Signal Analysis
- ▶ Stationarity
- ▶ Time Series Analysis
- ▶ Time Series Analysis in the Geosciences

Acknowledgments Dedicated to the memory of our Master, Benoit Mandelbrot (1924–2010) whose book has changed my life.

Bibliography

- Anis AA, Lloyd EH (1976) The expected value of the adjusted rescaled range of independent normal summands. *Biometrika* 63(1):111–116
- Bassingthwaite JB, Raymond GM (1994) Evaluating rescaled range analysis for time series. *Ann Biomed Eng* 22:432–444
- Gilmore M, Yu CX, Rhodes TL, Peebles WA (2002) Investigation of rescaled range analysis, the Hurst exponent, and long-time correlations in plasma turbulence. *Phys Plasmas* 9(4):1312–1317
- Hamed KH (2007) Improved finite-sample Hurst exponent estimates using rescaled range analysis. *Water Resour Res* 43:W04413
- Hewett TA (1986) Fractal distributions of reservoir heterogeneity and their influence on fluid transport. In: SPE annual technical conference and exhibition, New Orleans, October 1986. Paper number: SPE-15386
- Hurst HE (1951) Long-term storage capacity of reservoirs. *Trans Am Soc Civ Eng* 116:770–799
- Korvin G (1992) *Fractal models in the Earth sciences*. Elsevier, Amsterdam
- Křišťoufek L (2010) Rescaled range analysis and detrended fluctuation analysis: finite sample properties and confidence intervals. *Czech Econ Rev* 4:236–250
- Mandelbrot BB, Van Ness JW (1968) Fractional Brownian motions, fractional noises and applications. *SIAM Rev* 10(4):422–437
- Weron R (2002) Estimating long-range dependence: finite sample properties and confidence intervals. *Phys A* 312:285–299

Reyment, Richard A.

Peter Bengtson

Institute of Earth Sciences, Heidelberg University,
Heidelberg, Germany



Fig. 1 Richard A. Reyment (Courtesy of Professor Reyment's daughter Britt-Louise Kinnefors)

Biography

Richard Arthur Reyment (1926–2016) was born in Coburg, a suburb of Melbourne, Australia, in a family with roots in England, Ireland, Sweden, and Spain. After obtaining his B.Sc. from the University of Melbourne in 1948, he emigrated to Sweden. He started his scientific career as a geologist at the Geological Survey of Nigeria (1950–1956), followed by a senior lectureship at Stockholm University, Sweden (1956–1962). He was then appointed Professor of Petroleum Geology at the University of Ibadan, Nigeria (1963–1965), Associate Professor of Biometry at Stockholm University (1965–1967), and finally Professor and Chair of Historical Geology and Palaeontology at Uppsala University, Sweden (1967–1991). After retirement, he continued his research as Emeritus Professor at the Swedish Museum of Natural History in Stockholm.

In 1955 Reyment obtained his M.Sc. from the University of Melbourne on the subject of Cretaceous mollusks of Nigeria and shortly thereafter his Ph.D. from Stockholm University on the stratigraphy and palaeontology of the Cretaceous of Nigeria and the Cameroons. During his time as senior

lecturer at Stockholm University, his research focused mainly on micropalaeontology. In 1958/59 the Finnish palaeontologist Björn Kurtén visited Stockholm and introduced him to the use of statistical methods. This marked the beginning of what was to become his major field of research, biometric applications in palaeontology and related sciences. In 1960/61 he spent a period as research associate in the Soviet Union, where he was influenced by Andrei Borisovich Vistelius, generally regarded as the founder of the science of mathematical geology. In 1966/67 he spent a period at the Kansas Geological Survey (University of Kansas), USA, an institution with a strong record in quantitative geology, for example, through Robert R. Sokal, Daniel F. Merriam, F. James Rohlf, and Roger L. Kaesler.

In the mid-1960s Reymont headed the Subcommittee for the Application of Biometrics in Paleontology of the International Paleontological Union, and from 1968 to 1974 he was a member of the IUGS Committee on Storage, Automatic Processing and Retrieval of Geological Data (COGEODATA). At the 1968 International Geological Congress in Prague, the International Association for Mathematical Geology (IAMG) was born, with Reymont as prime mover. He was the first Secretary General of the Association 1968–1972 and President 1972–1976. Together with Daniel F. Merriam he founded the journal *Computers & Geosciences*, launched in 1975.

In 1967 Reymont was appointed to the Chair of Historical Geology and Palaeontology at Uppsala University. Besides teaching and supervising students, he continued his research and carried out extensive fieldwork. Important achievements during his Uppsala years were the IGCP Project “Mid-Cretaceous Events” and the launch of the journal *Cretaceous Research*.

Reymont was a highly versatile scientist. He published articles and books on several fossil groups (mainly ammonites, ostracods, and foraminifers), regional geology, stratigraphy, biogeography, dispersal of organisms, palaeoecology, evolution, sexual dimorphism, actuopalaeontology, sedimentology, random events in time (volcanic eruptions and earthquakes), geomagnetism, geochemistry, and history of geology. The majority of his publications concerned the Jurassic, Cretaceous, or Tertiary. Other publications dealt with mathematical methods and topics outside the geosciences, such as biology, biochemistry, genetics, serological analysis, linguistics, Romanis, and Spanish Moors. Remarkably enough, he also found time for his many hobbies and interests, such as music (clarinet and oboe), squash and tennis, family genealogy, history, and, not least, languages. He spoke fluent English, Swedish, French, German, and Spanish and had a good reading knowledge of Latin, Portuguese, Russian, and Afrikaans. He also studied Japanese, Yoruba, Finnish, and Romani.

Reymont was an influential pioneer in the application of mathematical methods in the geosciences. His impact is reflected in the numerous citations in the *Dictionary of Mathematical Geosciences* (Howarth 2017), found under the topics *asymmetry analysis*, *biostratigraphic zonation*, *biplot*, *canonical correlation*, *cross-validation*, *discriminant analysis*, *eigenanalysis*, *heteroscedacity*, *kurtosis*, *latent root*, *morphometrics*, *multivariate normal distribution*, *pattern classification*, *peakedness*, *point event*, *point process*, *positive definite matrix*, *principal component analysis*, *quantitative paleoecology*, *robust estimate*, *scores*, *serial correlation coefficient*, *singular value*, *slotting*, *statistical zap*, *survival function*, and *zap*. For additional information about Richard Reymont's activities and achievements, see Birks (1985), Merriam (2004), Merriam and Howarth (2004), Howarth (2017), and Sagar et al. (2018), and Bengtson (2018) for a general biography including a list of his publications.

Bibliography

- Bengtson P (2018) Richard A. Reymont (1926–2016) – Ammonitologist *sensu latissimo* and founder of *Cretaceous Research*. *Cretaceous Res* 88:5–35
- Birks HJB (1985) Recent and possible future mathematical developments in quantitative palaeoecology. *Palaeogeogr Palaeoclimatol Palaeoecol* 50:107–147
- Howarth RJ (2017) *Dictionary of mathematical geosciences*. Springer, Cham. xvi + 893 pp
- Merriam DF (2004) Richard Arthur Reymont: father of the International Association for Mathematical Geology. *Earth Sci Hist* 23:365–373
- Merriam DF, Howarth RJ (2004) Pioneers in mathematical geology. *Earth Sci Hist* 23:314–324
- Sagar BSD, Cheng Q, Agterberg F (eds) (2018) *Handbook of mathematical geosciences*. Springer, Cham. xxviii + 914 pp

R-Mode Factor Analysis

Norman MacLeod

Department of Earth Sciences and Engineering, Nanjing University, Nanjing, Jiangsu, China

Definition

R-Mode Factor Analysis – a statistical modeling procedure whose purpose is to find a small set of latent linear variables that estimate the influence of the factors controlling the inter-variable covariance or correlation structure in a multivariate dataset and are optimized to preserve as much of that structure as appropriate given an estimate of the number of causal influences.

R-Mode Factor Analysis, Table 1 Correlation matrix of academic test scores that exhibit a classic factor structure. (From Manly 1994)

	Classics	French	English	Mathematics	Music
Classics	1.00	0.83	0.78	0.70	0.63
French	0.83	1.00	0.67	0.67	0.57
English	0.78	0.67	1.00	0.64	0.51
Mathematics	0.70	0.67	0.64	1.00	0.51
Music	0.63	0.57	0.51	0.51	1.00

Introduction

Factor analysis (FA) is often regarded as a simple variant of principal components analysis (PCA). Indeed, in some quarters, FA has a somewhat nefarious reputation due to the nature of its statistical model and due to its overenthusiastic use in the field of psychology, especially the branch that deals with intelligence testing (see Gould 1981).

Historical Background

Around 1900 Charles Spearman noticed that scores on different academic subject tests often exhibited an intriguing regularity. Consider the following correlation matrix of test scores (Table 1).

If the diagonal values are ignored, in many cases the ratio between paired variable values approximates a constant. Thus, for English and Mathematics . . .

$0.78/0.70 = 1.114$	$0.67/0.67 = 1.000$	$0.51/0.51 = 1.000$
---------------------	---------------------	---------------------

. . . and for French and Music . . .

$0.83/0.63 = 1.317$	$0.67/0.51 = 1.314$	$0.67/0.51 = 1.314$
---------------------	---------------------	---------------------

Spearman (1904) suggested the most appropriate model for such data had the following generalized form.

$$X_j = a_{j,1}F_1 + e_j \quad (1)$$

In this expression, a_j was a standardized score the j^{th} test, F was a “factor” value for the individual taking the battery of tests, and e_j was the part of X_j specific to each test. Because Spearman was dealing with the matrix of correlations, both X and F were standardized, with means of zero and standard deviations of one. This fact was taken advantage of in Spearman’s other FA insight, that a_j^2 is the proportion of the variance accounted for by the factor.

Using these relations Spearman proposed that mental abilities could be deduced from the scores on a battery of different tests and were composed of two parts, one (F) that was common to all tests and so reflected the innate abilities – or “general intelligence” – and one (e) that was specific to each test and so could be interpreted as evidence of subject-specific

aptitude, culture-based bias in the test, or some combination of the two.¹ Generalizing the relations presented above, the model underlying FA has the following form.

$$\begin{aligned} X_1 &= a_{1,1}F_1 + a_{1,2}F_2 \cdots + a_{1,m}F_m + e_1 \\ X_2 &= a_{2,1}F_1 + a_{2,2}F_2 \cdots + a_{2,m}F_m + e_2 \\ X_3 &= a_{3,1}F_1 + a_{3,2}F_2 \cdots + a_{3,m}F_m + e_3 \\ &\vdots \\ X_j &= a_{j,1}F_1 + a_{j,2}F_2 \cdots + a_{j,m}F_m + e_j \end{aligned} \quad (2)$$

In this expression, a_j is a standardized value the j^{th} variable, F is the “factor” value for that variable across all measured or observed cases, and e_j is the part of the correlation not accounted for by the factor structure. *R*-mode factor analysis attempts to find a set of m linear factors (F , where $m < j$) that underly the expression of the sample’s covariance or correlation structure; hopefully one that provides a good estimate of the larger population’s structure from which the sample was drawn.

Note the similarity between this model and the multiple regression analysis model. The creators of factor analysis viewed it originally as a type of multiple regression analysis. But unlike a standard regression problem in which a set of observations (x) is related to an underlying general model (y) by specifying a rate of change (the slope β), one cannot observe any aspect of the factor(s) (F) directly. As a result, their structure must be inferred from the raw data, which represents an unknown mixture of common and unique variances.

Standard PCA also cannot resolve this problem. The PCA model is simply a transformation that re-expresses the observed data in a form that maximizes variance on the various eigenvectors and ensures their orientational independence from one another. This approach focuses on accounting for variance and keeps extracting vectors until the remaining unaccounted – or residual – variance is either exhausted or drops below some arbitrarily determined threshold. Factor analysis does not try to account for the variable variance

¹This seemingly straight-forward interpretation of Spearman’s work later became highly controversial when estimates of a person’s general intelligence were used to direct or limit their life opportunities and when the culturally biased character of many standard ‘intelligence’ tests was recognized (see Gould 1981 for a review).

values exhibited by a sample, but instead focuses on recovering the structure of the covariances or correlations among the variables. In other words, PCA attempts to reduce the residual values of the diagonal or trace of the covariance or correlation matrix. Factor analysis attempts to model the covariances or correlations in such a way as to minimize the residual values of the matrix's off-diagonal elements.

Among the many myths about FA is the idea that it only differs from PCA in the sense that the number of factors extracted from the covariance or correlation matrix in FA is less than the number of variables, whereas in PCA, the number of components extracted equals the number of variables. Nothing could be further from the truth. The principal use of PCA is to reduce the dimensionality of a dataset from p variables to m components, where $p < m$. Factor analysis attempts to infer both the number and the effects of the influences responsible for the variance structure. The erroneous view of equivalence in the relation between PCA and FA has been perpetuated by the innumerable over-generalized and/or superficial descriptions of both techniques, especially in the user's guides of commercial multivariate data-analysis software packages in which these methods are often discussed together owing to the use of eigenanalysis (or singular value decomposition) in their calculation.

Choosing and Extracting the Factors

There are a number of ways to perform a classic *R*-mode FA. The method presented below would not be accepted by many as "true" FA; especially as it is practiced in the social sciences. However, I will recount a basic version of the component-based procedure referred to as FA by most earth scientists (e.g., Jöreskog et al. 1976; Reymont and Jöreskog 1993; Manly 1994; Davis 2002).

Factor analysis begins with a PCA. In addition to a set of data, FA assumes the researcher has a level of understanding of the processes responsible for the data's variance structure sufficient to estimate how many independent factors are involved. This number is required as a controlling parameter.

In Spearman's case, he suspected a one-factor solution, the one-factor model quantifying his "general intelligence" index with the remaining test-specific variation being subsumed into the error term. For morphological data (see MacLeod

2005 for an example) where distance measurements have been collected from a set of fossil morphologies, the generalized factors might refer to "size" and "shape." Regardless of the number of factors specified, all applications of FA assume the existence of both "generalized effect" factors and a specific-effect deviation.

It is often the case that one has little idea beforehand how many factors are either present, or of interest, in the data. Fortunately, FA, like PCA, can be used in an exploratory mode. There are a number of FA rules-or-thumb that can be of assistance in making decisions regarding how many factors axes to specify.

If you are working with a correlation matrix, one of the most popular and easiest to remember is to set the number of factors to equal the number of PC eigenvectors that contain more information than the raw variables. This makes sense in that these axes are easy to identify (all will have eigenvalues >1.0) and there seems little logical reason to consider composite variables that contain less information than their raw counterparts. Another popular rule-of thumb is the so-called scree test in which the magnitudes of the eigenvalues are plotted in rank order and the decision made on the basis of where the slope of the resulting curve breaks from being dominantly vertical to dominantly horizontal. The other quick and easy test to apply when considering how many factors to extract, and indeed, whether factor analysis is appropriate, is Spearman's constant proportion test (see above). If a large number of the ratios between the non-diagonal row values of the correlation matrix do approximate a constant, the data exhibit the classic structure assumed by the factor analysis model and the appropriate number of factors will be the number of approximately constant values. Regardless, it is important to remember that, if there is uncertainty regarding the number of factors controlling the dataset's variance structure, FA may proceed but should be interpreted with caution.

Unlike classical PCA, factor analysis scales the loadings of the retained components by their eigenvalues, according to the following expression.

$$a_{j,m} = \sqrt[2]{\lambda_m} b_{j,m}. \quad (3)$$

Here, λ_m is the eigenvalue associated with the m^{th} principal component with m being set to the number of retained factors

R-Mode Factor Analysis, Table 2 Principal component loadings, factor loadings, and factor communalities for the example trilobite morphometric variables

Parameter/variable	Principal components			Factors		
	1	2	3	1	2	Communality
Eigenvalue	2.776 (92.54%)	0.142 (4.72%)	0.082 (2.74%)	2.776 (92.54%)	0.142 (4.72%)	
Body length	0.573	0.757	-0.314	0.954	0.285	0.992
Glabella length	0.583	-0.108	0.805	0.972	-0.041	0.947
Glabella width	0.576	-0.644	-0.504	0.959	-0.242	0.979

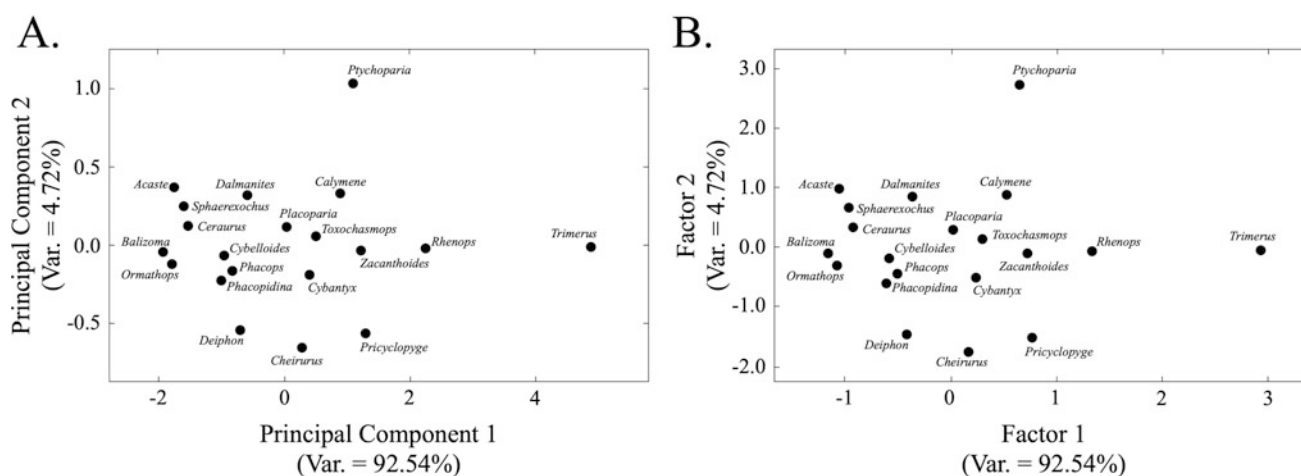
and b is the eigenvector loading of the j^{th} variable on the m^{th} principal component. This scaling, along with the reduction in the number of components (= factors) being considered, represents a basic computational difference between PCA and FA.

Next, the “communality” values of the m scaled factors are calculated as the sum of squares of the factor loading values. This summarizes the proportion of variance provided by those variables to each factor. The quantity “1- the sum of communality values for each variable” then expresses the proportion of the variable’s variance attributable to the error term (e) of the factor model. If the correct number of factors has been chosen, all the summed, squared factor values should exhibit a high communality on each variable with a low residual error.

Once the factor equations have been determined, the original data can be projected into the factor space. The resulting

factor scores ($\hat{X}$) may then be tabulated, plotted, inspected, and interpreted in the manner normal for PCA ordinations. Table 2 and Fig. 1 compare and contrast results of a PCA and two-factor FA for the trilobite data listed in MacLeod (2005).

For this simple dataset, both the first principal component and first factor would be interpreted as being consistent with generalized (allometric) size variation owing to their uniformly positive loading coefficients of varying magnitudes. Similarly, for both analyses, the second component/factor would be interpreted as a localized shape factor owing to the contrast between body length and the glabella variables. While the relative positions of the projected data points are also similar, and discontinuities in the gross form distributions are evident in both plots, the scaling of the PCA and FA axes differs reflecting the additional eigenvalue-based scaling that is part of component-based FA. Nevertheless, an important change has taken place in the ability of the FA vectors to



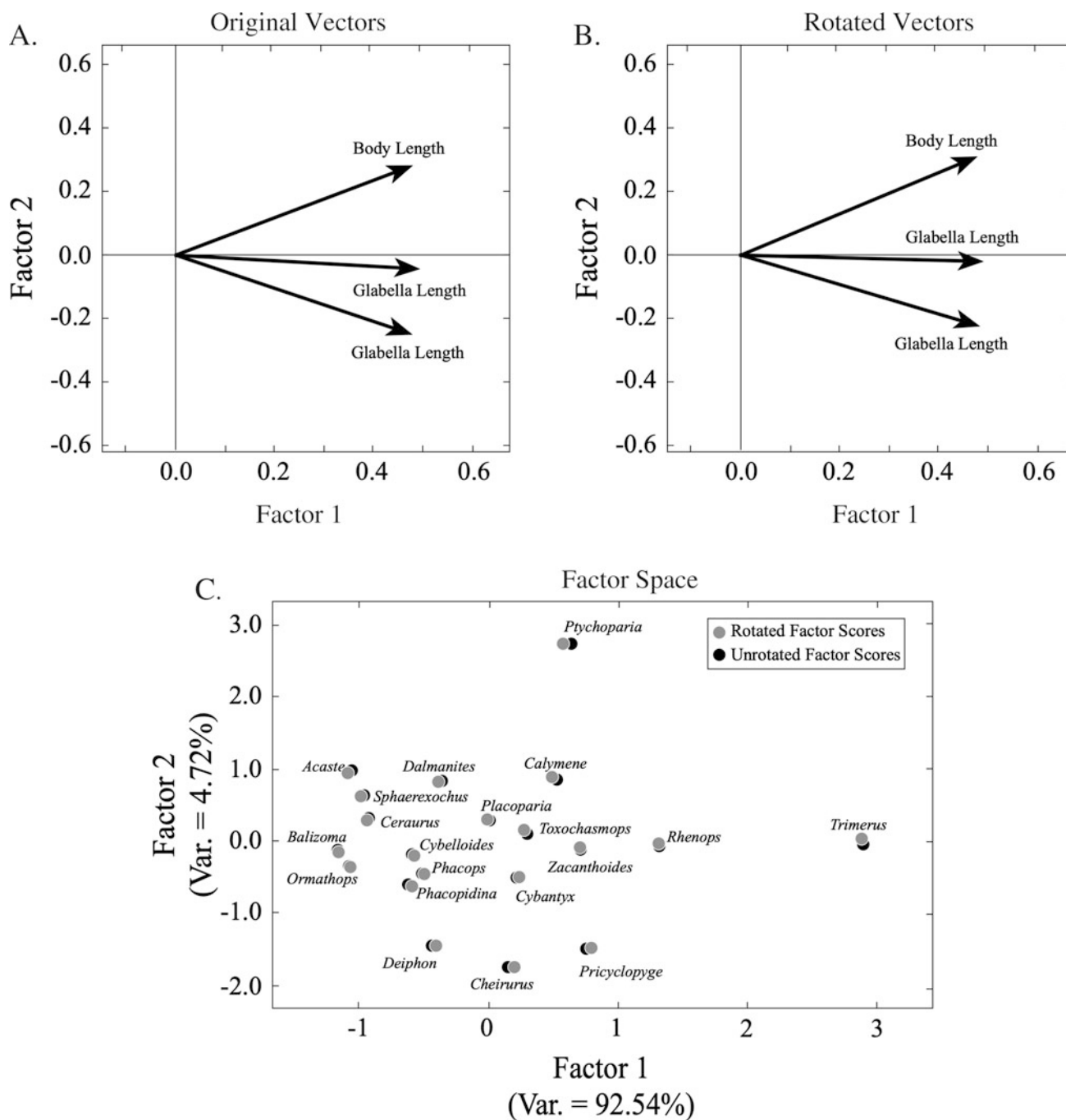
R-Mode Factor Analysis, Fig. 1 Ordination spaces formed by the first two principal components (a) and a two-factor extraction (b) from the three-variable trilobite dataset. Note the similarity of these results in

terms of the relative positions of forms projected into the PCA and factor spaces and the difference in terms of the axis scales

R-Mode Factor Analysis, Table 3 Original correlation matrix, correlation matrix reproduced on the basis of the first to principal components, and correlation matrix reproduced by the two-factor FA solution. Note

the bottom two matrices are based on the component/loading values shown in Table 2

Original correlation matrix			
	Body length	Glabella length	Glabella width
Body length	1.000	0.895	0.859
Glabella length	0.895	1.000	0.909
Glabella width	0.859	0.909	1.000
Reproduced correlation matrix (PCA)			
Body length	0.902	0.253	−0.158
Glabella length	0.253	0.352	0.405
Glabella width	−0.158	0.405	0.746
Reproduced correlation matrix (FA)			
Body length	0.992	0.916	0.847
Glabella length	0.916	0.947	0.943
Glabella width	0.847	0.943	0.979



R-Mode Factor Analysis, Fig. 2 Original (a) and rotated (b) variable-axis orientations with respect to the fixed, orthogonal factor axis orientation for the three-variable trilobite data. The rotation of these variables is centered on Factor 1 owing to their high mutual correlations.

A summary of the (minor) projected data displacements arising as a result of varimax axis rotation (c). More complex datasets will typically exhibit larger variable-vector rotations

represent the structure of the original correlation matrix (Table 3).

Here it is evident that the two-factor PCA result focuses on reproducing the values along the trace of the original covariance/correlation matrix, whereas the FA result focuses on reproducing the entire structure of the covariance/correlation

matrix. If all three principal components had been used in these calculations, only the trace of the original correlation matrix would be reproduced exactly, whereas if a three-factor solution had been calculated, the original correlation matrix would have been reproduced exactly (except for rounding error). In this sense then, FA represents a more information-

rich analysis than PCA, especially if a dramatic reduction in dimensionality is required. The validity of applying the FA model to a dataset is predicated, however, on the dataset exhibiting a factor-based variance structure.

Factor Rotation

In addition to scaling of the factor loadings, the other aspect of a classic component-based FA that differs from standard PCA is factor rotation. As can be seen from Fig. 1, the factor-scaling calculation does not alter the orientation of the factor axes with respect to the original variable axes. However, since the goal of a FA is to “look beyond” the data to hand in order to discover the structure of the influences responsible for observed patterns of variation, strict geometric conformance between the original data and the factor model is regarded by many factor analysts as being beside the point. Thus, a justification is provided to alter the orientations of the factor axes relative to the original variable axes (or vice versa) in order to make them easier to interpret in terms of the original variables.

Two general approaches to factor rotation are popular: orthogonal (in which the orthogonality of the factor-axes system is maintained) and oblique (in which the factor axis-orthogonality constraint is relaxed). In both cases, the goal of factor rotation is to align the factor axes with the covariance/correlation structure of the variable set to the maximum extent possible. Orthogonal axis rotation (usually accomplished using Kaiser’s 1958 varimax procedure) performs this operation by iteratively changing the orientation of the factor-axis system two factors at a time until the squares of the variances are either maximized or minimized across all factors. Operationally this amounts to adjusting the set of factor loadings, rigidly in a geometric sense, until all variables exhibit loadings at or as close to 0.0 or ± 1.0 as possible. Figure 2 illustrates the unrotated and varimax rotated solutions for the example trilobite dataset.

Obviously, in this simple example, the variable vectors were already quite well aligned with factor 1. However, a small adjustment was necessary to achieve a fully optimized result. In higher-dimensional datasets, the angle of varimax-optimized factor rotation can be quite large (see Davis 2002 for an example).

A final word about calculation of the factor scores that project the original data into the space of the factor axes. The PCA method accomplishes this by postmultiplying the original data values (X) by the component loadings (A). The factor model differs from principal component model, however, in that the former subdivides the variance structure into common and unique factors (see above). Thus, true factor scores (X')

cannot be calculated in the same way as principal component scores because the data matrix contains contributions from both the common and unique factors. In order to obtain the true factor scores, the raw (or standardized) data must be normalized for the unique factor’s influence. Fortunately, this can be accomplished with minimal effort by postmultiplying the data matrix by the inverse of the covariance/correlation matrix (S^{-1}), prior to the factor loading postmultiplication.

$$\hat{X} = XS^{-1}F \quad (4)$$

Conclusion

R-mode factor analysis has a long and distinguished history in the earth sciences and has played a key data-analytic role in advancing many earth-science research programs (e.g., Jöreskog et al. 1976; Sepkoski 1981; Reyment and Jöreskog 1993). Recently, Bookstein (2017) has advocated its use in the context of general morphometric analysis as a more path-analytic approach to understanding of the influences that underly shape variation. It is to be hoped that this and (perhaps) other applications of FA in novel data-analytic contexts will lead to the wider use of FA in geological, oceanographic, biological, meteorological, and cryological contexts.

Cross-References

- [Correlation and Scaling](#)
- [Eigenvalues and Eigenvectors](#)
- [Morphometry](#)
- [Principal Component Analysis](#)
- [Regression](#)
- [Shape](#)
- [Variance](#)

References

- Bookstein FL (2017) A method of factor analysis for shape coordinates in physical anthropology. *J Phys Anthropol* 164:221–245
- Davis JC (2002) *Statistics and data analysis in geology*, 3rd edn. Wiley, New York
- Gould SJ (1981) *The mismeasure of man*. W. W. Norton, New York
- Jöreskog KG, Klován JE, Reyment RA (1976) *Geological factor analysis*. Elsevier, Amsterdam
- Kaiser HF (1958) The varimax criterion for analytic rotations in factor analysis. *Psychometrika* 23:187–200
- MacLeod N (2005) Factor analysis. *Palaeontol Assoc Newsl* 60:38–51
- Manly BFJ (1994) *Multivariate statistical methods: a primer*. Chapman & Hall, Bury, St. Edmonds, Suffolk

- Reyment RA, Jöreskog KG (1993) Applied factor analysis in the natural sciences. Cambridge University Press, Cambridge
- Sepkoski JJ (1981) A factor analytic description of the Phanerozoic marine fossil record. *Paleobiology* 7:36–53
- Spearman C (1904) “General intelligence”, objectively determined and measured. *Am J Psychol* 15:201–293

Robust Statistics

Peter Filzmoser

Institute of Statistics & Mathematical Methods in Economics,
Vienna University of Technology, Vienna, Austria

Definition

Robust statistics is concerned with the development of statistical estimators that are robust against certain model deviations, caused, for example, by outliers.

Introduction

Data analysis and robust statistics have a strong historical link, because many questions regarding specific features in the data structure are connected to the outlier problem. Outliers have always been viewed as something atypical, not represented in the data majority, and for that reason they are usually of special interest to the analyst. Robust statistics, in general, is concerned with the development of statistical estimators that are robust against certain model deviations, such as deviations from normal distribution, which is frequently assumed in estimation theory. Robust estimators are supposed to still deliver reliable results in case of such deviations, e.g., caused by outliers, and various tools that allow to quantify the robustness of an estimator have been developed (Maronna et al. 2019). Probably the most prominent tool is the *breakdown point* (BP), which refers to the smallest fraction of observations that could be replaced by any arbitrary values in order to make the estimator completely meaningless. For example, for estimating the univariate location, the commonly used arithmetic mean has a BP of zero, because already replacing a single observation by an arbitrary value can drive the estimator beyond all bounds. The median, on the other hand, achieves the maximum BP of 0.5. Under normality, however, the median comes with a lower statistical efficiency, and thus often a compromise between (positive) BP and reasonable efficiency needs to be taken (e.g., trimmed mean or, better, M estimator – see below).

This entry aims to explain some important tools and concepts in robust statistics from a practical perspective. Robust regression and covariance estimation, as well as principal component analysis (PCA) will be in focus. The interested reader can find many more details and background to robust statistics in the book Maronna et al. (2019).

Regression

Many text books on robustness motivate the use of robust regression routines by presenting some x - and y -data which follow a linear trend – with the exception of some single outliers. Those outliers, appropriately placed, are able to completely spoil the least-squares (LS) regression line but do not affect the fit from a robust regression. While such examples may be inspiring, one would always ask why the outliers are not simply removed, as they are clearly visible, and a subsequent LS regression would be useful for predicting y from x . The problem is that in more complex situations, especially in higher dimension, outlier detection is not at all trivial. Fortunately, the outlier detection step is not necessary in practice, because robust regression methods will automatically downweight outliers, thus observations which are inconsistent with the linear model, and this also works in the higher-dimensional situation.

Example Data

In order to be more specific, a data set from the GEMAS project (Reimann et al. 2014), a geochemical mapping project covering most European countries, is considered in the following. Similar as in van den Boogaart et al. (2020), a regression of pH on the composition of the major oxides Al_2O_3 , CaO , Fe_2O_3 , K_2O , MgO , MnO , Na_2O , P_2O_5 , SiO_2 , TiO_2 , and LOI is of interest. Here, only the data from Poland are used, as they are relatively homogeneous concerning the soil type. In other countries, the pH values heavily depend on whether the soils are on karstic landscapes or on felsic rocks, for instance.

Regression Model

The response, here pH, is denoted as variable y , while the explanatory variables are x_1 to x_D . In total, there are $n = 129$ samples taken from Poland, which lead to the observations y_i , and $x_{i1}, \dots, x_{iD}$, for $i = 1, \dots, n$. Since the explanatory variables are compositional data, it is advisable to first transform them into the usual Euclidean geometry for which the traditional regression methods have been designed. One option is to use pivot coordinates (Filzmoser et al. 2018), which results in the variables $z_1, \dots, z_{D-1}$, with the corresponding observations $\mathbf{z}_i = (z_{i1}, \dots, z_{i,D-1})'$. The linear regression model is

$$y_i = b_0 + \sum_{j=1}^{D-1} z_{ij}b_j + \varepsilon_i, \quad (1)$$

with the error terms ε_i , which are supposed to be independent and $N(0, \sigma^2)$ distributed, where σ^2 is the error variance. For pivot coordinates, the particular interest is in the coefficient b_1 associated to variable z_1 , which contains all relative information of x_1 to the remaining parts in the composition. A reordering of the composition, with a different part on the first position, followed by computing pivot coordinates allows for an interpretation of the coefficient related to another (first) part (Filzmoser et al. 2018). For a given estimator $\hat{\mathbf{b}} = (\hat{b}_0, \hat{b}_1, \dots, \hat{b}_{D-1})'$, one can compute the fitted values $\hat{y}_i = (1, \mathbf{z}_i')\hat{\mathbf{b}}$, and the residuals $r_i = r_i(\hat{\mathbf{b}}) = y_i - \hat{y}_i$, for $i = 1, \dots, n$, which refer to the discrepancy of the observed responses to the fitted ones.

Least-Squares (LS) Estimator

The LS estimator minimizes the sum of the squared residuals and is thus defined as

$$\hat{\mathbf{b}}_{LS} = \underset{\mathbf{b}}{\operatorname{argmin}} \sum_{i=1}^n r_i^2(\mathbf{b}). \quad (2)$$

Outliers in the context of regression may refer to atypical values in the response, or to unusual observations in the space of the explanatory variables. The former are called vertical outliers, and the latter leverage points, where a distinction between good (values along the regression hyperplane) and

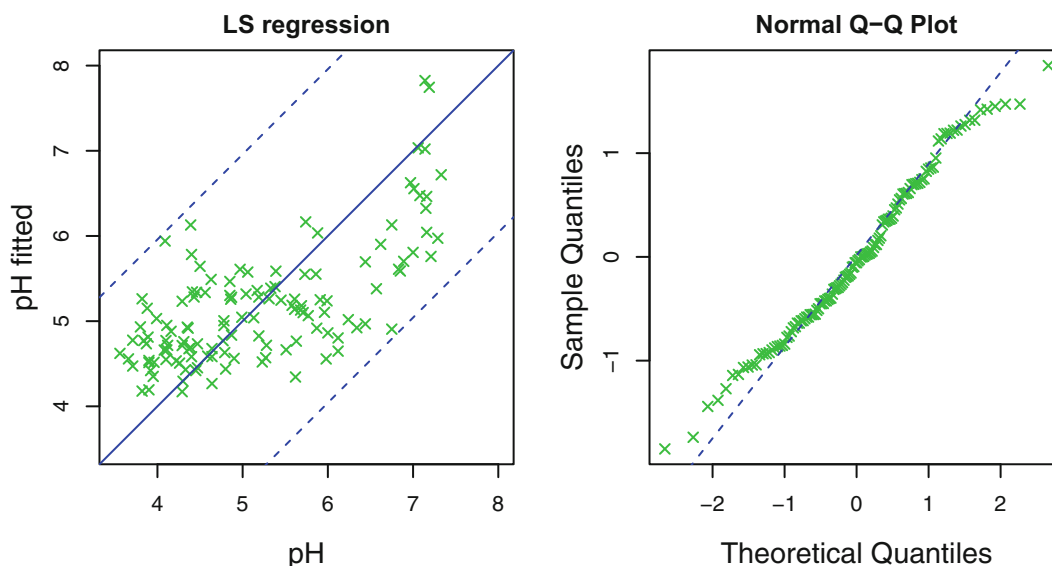
bad leverage points (values far away from the regression hyperplane) is made. Vertical outliers as well as bad leverage points lead to large residuals, and their square can heavily dominate the minimization problem (Eq. 2). For this reason, $\hat{\mathbf{b}}_{LS}$ is sensitive to outliers and has a BP of zero, since even a single observation could “attract” the regression hyperplane and thus make the estimator useless for the data majority.

LS Regression for the Example

For the considered example data set, a reliable outlier detection based on visual inspection is practically impossible. Visual diagnostics is thus done only after model fitting. Figure 1 (left) shows the response values y_i versus the fitted values $\hat{y}_i$ from LS regression. The solid line indicates equality of response and fitted values, and the dashed lines refer to the outlier cutoff values for the residuals, which are typically considered as $r_i(\hat{\mathbf{b}}_{LS})/\hat{\sigma}_{LS} = \pm 2.5$, where $\hat{\sigma}_{LS}^2$ is the estimated residual variance from the LS estimator. According to these cutoff values, there are no outlying residuals, and the QQ-plot (Fig. 1, right) even confirms approximate normality of the residuals. The model does not fit well, because for small values of pH it leads to an overestimation, and for bigger values to an underestimation. At this stage, however, it is unclear if a nonlinear model would be more appropriate, if outliers could have affected the LS estimator, or if the explanatory variables are simply not more useful for modeling the response.

Robust Regression

Robustification of the LS estimator are M estimators for regression, which minimize a function, say ρ of the residuals.



Robust Statistics, Fig. 1 LS regression of pH on the composition of the major oxides. Left plot: pH versus fitted values of pH; right plot: normal QQ-plot for the LS residuals

Since this would depend on the residual scale, the M estimator is defined as

$$\hat{\mathbf{b}}_M = \underset{\mathbf{b}}{\operatorname{argmin}} \sum_{i=1}^n \rho\left(\frac{r_i(\mathbf{b})}{\hat{\sigma}}\right), \quad (3)$$

where $\hat{\sigma}$ is a robust scale estimator of the residuals (Maronna et al. 2019). If ρ is the square function, thus $\rho(r) = r^2$, one would again obtain the LS estimator. In order to achieve robustness, the function ρ thus needs to be chosen adequately: For small (absolute) residuals, it should be approximately quadratic to obtain high efficiency, while for large (absolute) residuals it needs to increase more slowly than the square. Various proposals are available in the literature, and they also affect the robustness properties of the resulting estimator. It also matters how $\hat{\sigma}$ is derived. The so-called MM estimator combines a highly robust scale estimator with an M estimator, and it achieves a BP of 0.5, with a high (tunable) efficiency (Maronna et al. 2019).

Another robust regression estimator which is probably easier to understand and still widely applied is the LTS (least trimmed squares) estimator, defined as

$$\hat{\mathbf{b}}_{LTS} = \underset{\mathbf{b}}{\operatorname{argmin}} \sum_{i=1}^h r_{(i)}^2(\mathbf{b}), \quad (4)$$

where $r_{(1)}^2 \leq \dots, r_{(h)}^2 \leq \dots \leq r_{(n)}^2$ are the ordered values of the squared residuals and h is an integer between $n/2$ and n . The maximum BP of 0.5 is attained for $h = \lfloor \frac{n+D}{2} \rfloor$, where $\lfloor a \rfloor$ is the integer part of a (Maronna et al. 2019). Based on the

residuals from the LTS estimator, it is possible to robustly estimate the residual variance by

$$\hat{\sigma}_{LTS}^2 = c \cdot \frac{1}{h} \sum_{i=1}^h r_{(i)}^2(\hat{\mathbf{b}}_{LTS}), \quad (5)$$

where c is a constant for consistency under normality.

Similar as above, residuals would be considered as outlying if

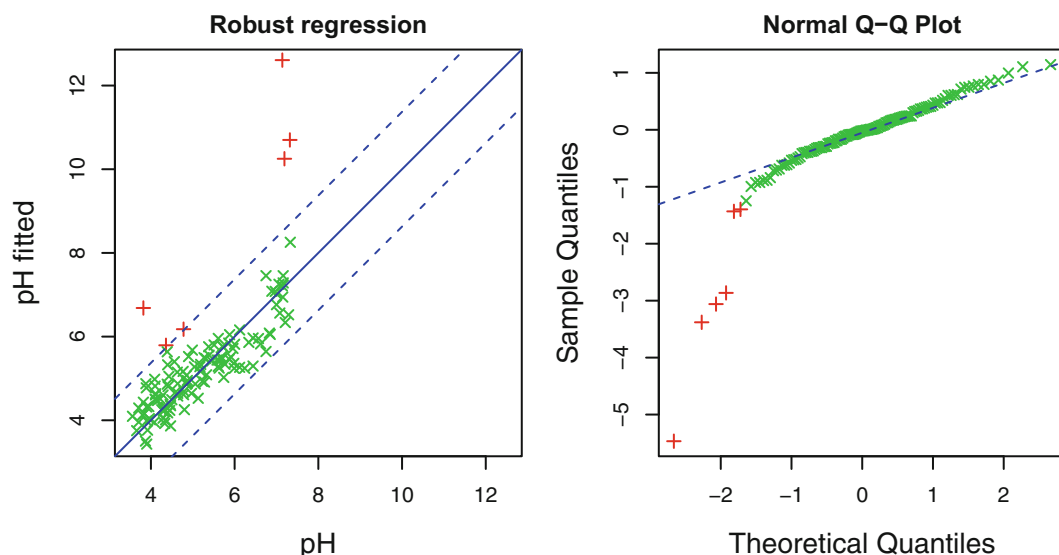
$$|r_i(\hat{\mathbf{b}}_{LTS}) / \hat{\sigma}_{LTS}| > 2.5. \quad (6)$$

In contrast to the outlier diagnostics for LS regression, both the residuals and the residual variance are estimated robustly within LTS regression, and thus, this diagnostic tool is reliable and useful in presence of outliers.

Since the statistical efficiency of the LTS estimator is low, it is common to add a reweighting step. This consists of a weighted LS estimator, with weights 0/1, where zero is assigned to residuals for which (Eq. 6) is valid, and one otherwise.

Robust Regression for the Example

Figure 2 presents the resulting plots for LTS regression, analogously as in Fig. 1 for LS regression. The dashed lines in the left plot correspond to the outlier cutoff values according to (Eq. 6), and here indeed, several points are identified as outliers (red symbols +). The model still has a problem with accommodating pH values in the range from about 6 to 7, but it fits much better for lower pH values



Robust Statistics, Fig. 2 LTS regression of pH on the composition of the major oxides. Left plot: pH versus fitted values of pH; right plot: normal QQ-plot for the LTS residuals. Red points with + are outliers

compared to the LS model. Also the QQ-plot confirms approximate normality of the (nonoutlying) residuals.

The superiority of the LTS model is also reflected in the MSE (mean-squared error), which is 0.22 (outliers not included). The MSE for the LS model is 0.61 (no outliers identified). A further interesting outcome is the inference table, which reveals the explanatory variables that are significant in the model. Accordingly, for LTS regression, the pivot coordinates for Al_2O_3 , CaO , Fe_2O_3 , K_2O , Na_2O , and LOI are significant. In contrast, for the LS model, only MnO and CaO resulted in significance, and thus the interpretation would be quite different.

In practice, there is often the desire to continue with removing the outliers, and then applying again LS regression. However, this has precisely been done with the reweighted LS step after LTS regression. Other robust regression estimators such as MM regression also determine weights for a weighted LS regression, but the weights can be in the whole interval $[0,1]$, corresponding to the outlyingness of the residuals.

Principal Component Analysis (PCA)

PCA is a dimension reduction method and widely used for data exploration. The projection onto the first two principal components (PCs) often allows to inspect the essential information contained in the data. However, in presence of outliers, classical PCs can be strongly determined by the outliers themselves, since PCs are defined as directions maximizing the (classical) variance. Usually, the PC directions are computed as the eigenvectors of the covariance matrix, which first needs to be estimated.

Covariance Estimation

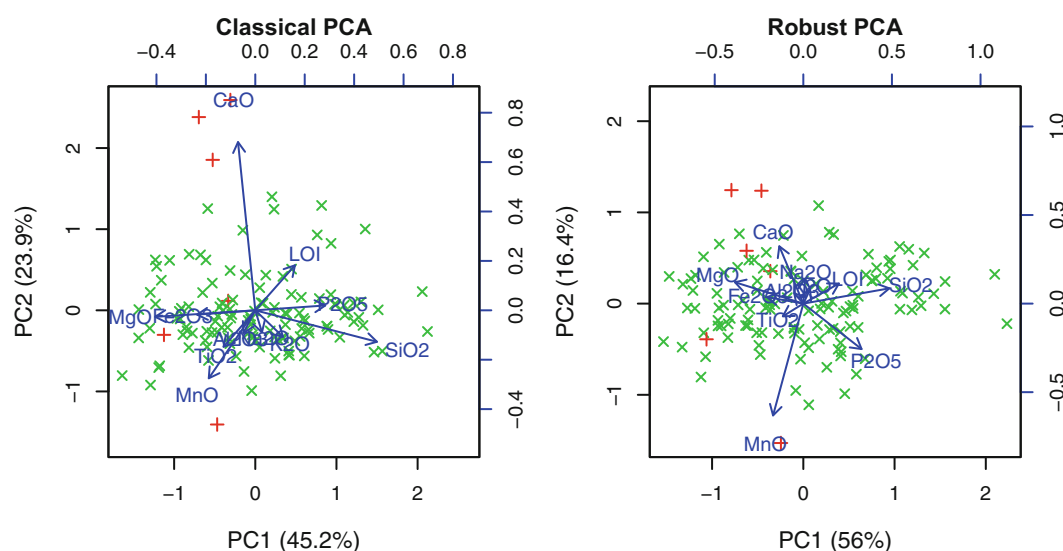
Assume $(D-1)$ -dimensional observations $\mathbf{z}_1, \dots, \mathbf{z}_n$. Then the classical (empirical) covariance matrix is

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})', \quad (7)$$

with the arithmetic mean $\bar{\mathbf{z}} = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i$. This covariance estimator is sensitive to outliers, and there are many options for a more robust estimation. One well-known estimator is the MCD (minimum covariance determinant) estimator, which is given by the mean and covariance of the subset of those h points (with $\frac{n}{2} \leq h \leq n$) which yield the smallest determinant of the empirical covariance matrix (Maronna et al. 2019). Similar to LTS regression, the choice of h determines the BP of the estimator. Higher efficiency of the estimator is achieved by a reweighting step (see Maronna et al. 2019).

PCA for the Example

Consider again the oxide composition from the GEMAS data set, which has been used above in a regression model to explain the response pH. With robust regression, several outliers in the residuals have been identified, and these are shown in the plots in Fig. 3 by red symbols +. These plots are so-called CoDa-biplots, and the compositional parts are transformed from pivot coordinates to centered logratio coefficients for a meaningful interpretation (Filzmoser et al. 2018). The left biplot is for classical PCA, based on an eigen-decomposition of the sample covariance matrix, and the right biplot for robust PCA, using the MCD estimator as a basis for the calculations.



Robust Statistics, Fig. 3 CoDa-biplots for classical (left) and robust (right) PCA of the oxide composition. Red symbols + are residual outliers from robust regression, see Fig. 2

The classical biplot reveals that the direction of the second component PC2 seems to be mainly determined by outliers, and these outliers appear in the part CaO. Note that the pivot coordinate for CaO was significant in the LS regression model, and this is very likely just an artifact of the outliers. The biplot for robust PCA is not affected by the outliers, and thus it better reflects the relationships in the oxide composition.

Summary

Robust statistical methods are supposed to give reliable results even if strict model assumptions that are required for the classical methods are violated to some extent. This means, however, that for the data majority the model assumptions need to be fulfilled. An example is multivariate outlier detection, where outlier cutoff values are usually derived from distributional assumptions valid for the data majority. There are also other approaches, in particular for outlier detection, which are not based on distributional assumptions, typically originating from the machine learning community. Methods from both disciplines can be very useful, but this also depends on the problem setting. For a recent overview in this context, see Zimek and Filzmoser (2018).

Nowadays, it is easy to compare the results of a classical and a robust statistical method, because software is freely available. For example, many robust methods are implemented in the package *robustbase* (Maechler et al. 2020) of the software environment R (R Development Core Team 2020), and the authors did their best to guarantee a similar input and output structure for the classical and robust methods. If the results of both analyses differ (clearly), one should go into detail to see why this happened.

The field of robust statistics is still very active with developing new methods and technologies. The “big data era” has not only led to bigger data sets, but also to more complex data structures, which brings in new challenges and requirements for the analysis methods. As an example, in the high-dimensional data setting it is desirable that not the complete information of an observation is discarded, but rather that only deviating entries of the observation are downweighted. For recent references to such approaches, see Filzmoser and Gregorich (2020).

Cross-References

- [Compositional Data](#)
- [Exploratory Data Analysis](#)
- [Least Median of Squares](#)
- [Least Squares](#)
- [Multivariate Analysis](#)

- [Ordinary Least Squares](#)
- [Principal Component Analysis](#)
- [Regression](#)
- [Statistical Outliers](#)

Bibliography

- Filzmoser P, Gregorich M (2020) Multivariate outlier detection in applied data analysis: global, local, compositional and cellwise outliers. *Math Geosci* 52(8):1049
- Filzmoser P, Hron K, Templ M (2018) Applied compositional data analysis: with worked examples in R. Springer, Cham
- Maechler M, Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibián-Barrera M, Verbeke T, Koller M, Conceicao ELT, Anna di Palma M (2020) *Robustbase: basic robust statistics*. R Package Version 0.93-6
- Maronna R, Martin R, Yohai V, Salibián-Barrera M (2019) *Robust statistics: theory and methods* (with R). John Wiley & Sons, New York
- R Development Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. ISBN 3900051-07-0
- Reimann C, Birke M, Demetriades A, Filzmoser P, O'Connor P (eds) (2014) *Chemistry of Europe's agricultural soils – part a: methodology and interpretation of the GEMAS data set*. Geologisches Jahrbuch (Reihe B 102). Schweizerbart, Hannover
- van den Boogaart K, Filzmoser P, Hron K, Templ M, Tolosana-Delgado R (2020) Classical and robust regression analysis with compositional data. *Math Geosci* 1–36
- Zimek A, Filzmoser P (2018) There and back again: outlier detection between statistical reasoning and data mining algorithms. *WIRES Data Min Knowl Discov* 8(6):e1280

Rock Fracture Pattern and Modeling

Katsuaki Koike¹ and Jin Wu²

¹Department of Urban Management, Graduate School of Engineering, Kyoto University, Kyoto, Japan

²China Railway Design Corporation, Tianjin, China

Definition

Fracture is a general term for any type of rock discontinuity (e.g., crack, fissure, joint, fault) and covers a wide size range (mm to km scale). Fractures are generated by the destruction of rock mass in a compressive or tensile stress field. Fractures are mechanically classified into the following categories: mode I, extension or opening; mode II, shear; and mode III, a mixture of modes I and II. The degree of fracture development is essential to engineering practices because it strongly affects the hydraulic and mechanical properties of rock mass including their magnitude and anisotropic behavior. Fractures are generally targeted as representative elements of a fracture pattern to characterize their development, length, orientation,

density, connectivity, aperture, and geometry (Fig. 1). Using these elements, a spatial pattern of the fracture distribution can be computationally modeled following several statistical assumptions, for which discrete fracture network (DFN) modeling is the most representative approach (e.g., Lei et al. 2017). Such computational modeling is indispensable, because the fracture distribution in rock mass is difficult to predict by 1D and 2D surveys using borehole and rock surface data, respectively. Constructed fracture models can be linked to fluid flow modeling, permeability assessments, and deformation simulations of rock mass.

Fracture Pattern Quantification Methods

Length Distribution

The true length of a fracture is impossible to measure because only a part of a fracture system appears on the rock mass surface. The trace length on the surfaces of an outcrop, tunnel, dam, or drift is therefore typically used to characterize the fracture length distribution. This distribution can be approximated using a negative power law, lognormal distribution, or negative exponential distribution (Bonnet et al. 2001). For example, the distribution of fracture length l following a negative power law with scaling exponent a and density constant α is formulated as:

$$n(l) = \alpha \cdot l^{-a} \quad (1)$$

where $n(l)dl$ is the number of fractures per unit volume in the length range $[l, l + dl]$. Because this law signifies the absence of a characteristic length, the definition of a total range

between the minimum and maximum lengths $[l_{\min}, l_{\max}]$ is necessary for practical use. Upon increasing the a value, the $n(l)$ inclination becomes steeper and the ratio of short to long fractures increases, which is demonstrated by a comparison of the simulated fracture distributions in a cubic domain of 100-m side lengths with $a = 2.5, 3.0$, and 3.5 and a length range of $[1 \text{ m}, 1000 \text{ m}]$ (Fig. 2a). Long and short fractures are dominantly distributed in models with $a = 2.5$ and 3.5 , respectively.

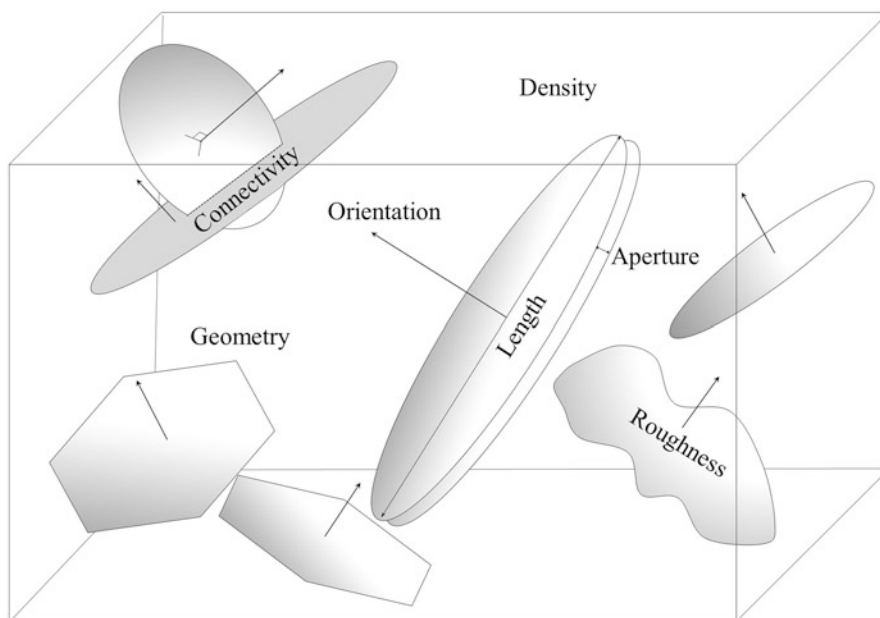
Orientation Distribution

Fracture orientation is defined by the strike and dip or dip and dip direction. The distribution of fracture orientation is typically approximated using a uniform, Gaussian, or Fisher distribution by considering the statistical property of the orientations, such as the dispersion and shape. Among them, the Fisher distribution (Fisher 1993), which is expressed by the following probability density distribution $f(\theta)$, has been widely adopted owing to its adequate fitness to many case studies (e.g., Hyman et al. 2016).

$$f(\theta) = \frac{\kappa \sin \theta e^{\kappa \cos \theta}}{e^{\kappa} - e^{-\kappa}} \quad (2)$$

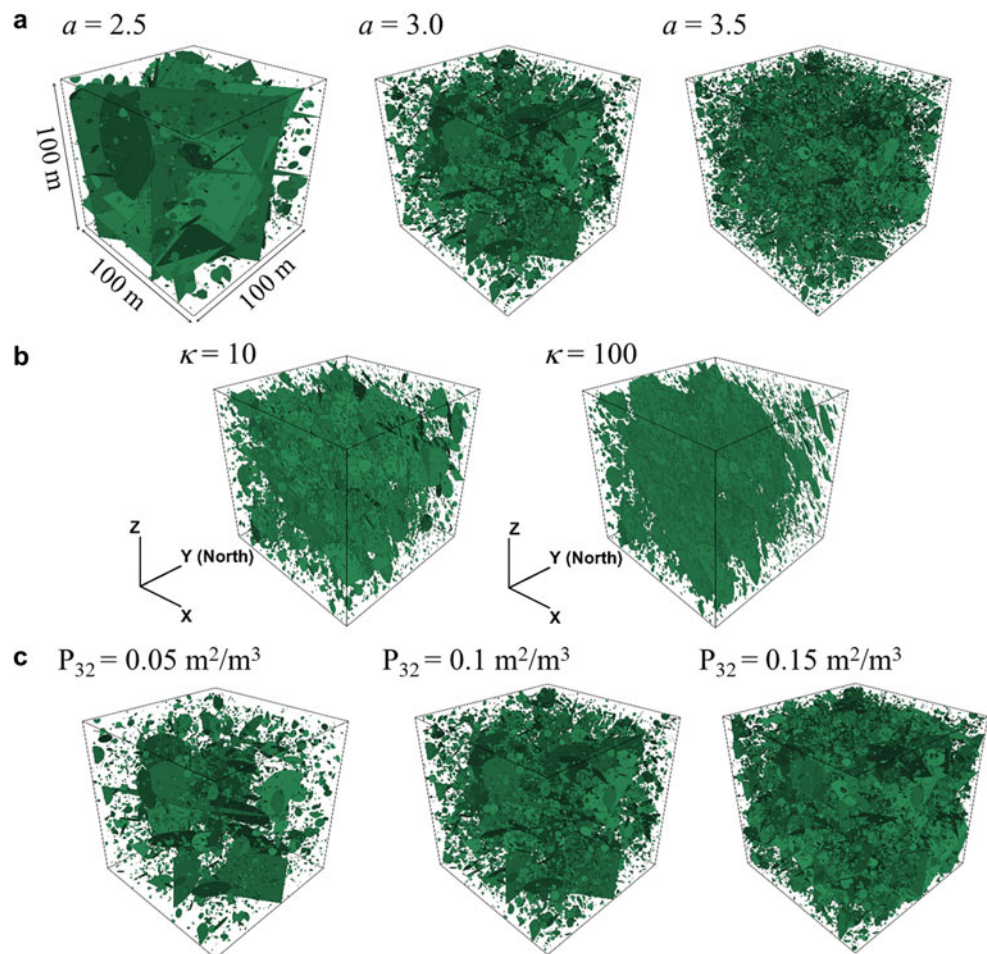
where θ ($^\circ$) represents the deviation angle from the mean vector of the dip direction and κ is the Fisher constant that expresses the dispersion of direction clusters. Fracture orientations are scattered with small κ values and become concentrated with increasing κ values, as demonstrated by comparing two DFN models with $\kappa = 10$ and 100 in Fig. 2b. The greatest advantage of the Fisher distribution is

Rock Fracture Pattern and Modeling, Fig. 1 Representative elements for characterizing rock-fracture patterns



Rock Fracture Pattern and Modeling, Fig. 2

DFN models with different distributions of fracture length and orientation and densities in a calculation domain of $100 \times 100 \times 100$ m. (a) DFN models following a power law of length distribution with three scaling exponents a , length range [1 m, 1000 m], and $P_{32} = 0.1 \text{ m}^2/\text{m}^3$. (b) DFN models with two dispersion parameters κ with a mean dip angle of 60° and azimuth angle of 60° clockwise from the north. (c) DFN models with three fracture densities $P_{32} = 0.05, 0.1$, and $0.2 \text{ m}^2/\text{m}^3$.



the correct modeling of preferential orientation upon suitably setting the angle dispersion around the mean orientation.

Density

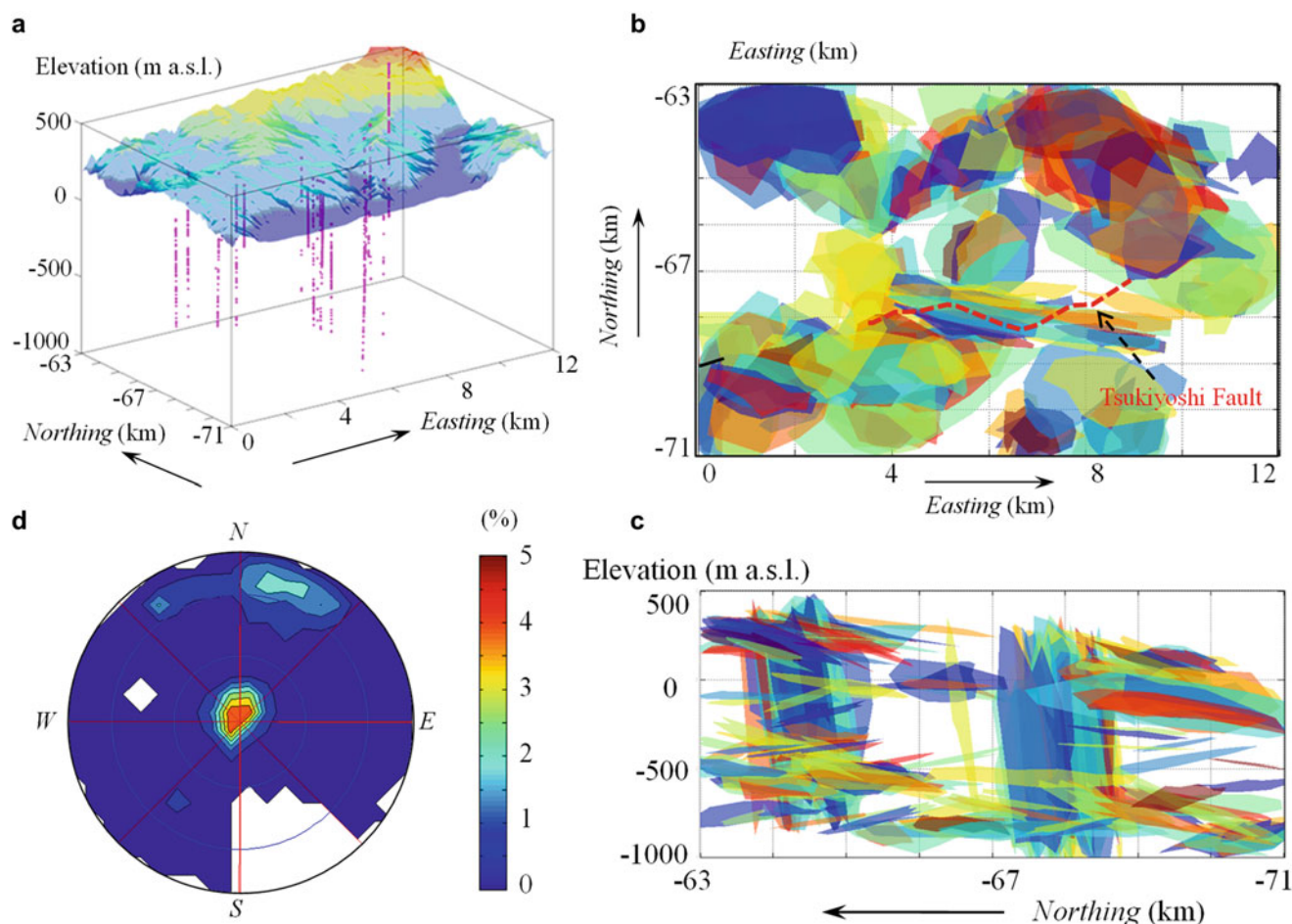
Fracture density is typically quantified using scale-independent indices, P_{10} , P_{21} , and P_{32} (Dershowitz and Herda 1992). $P_{10} (\text{m}^{-1})$ is the number of fractures per unit length intersected by a 1D investigation line using a scanline or borehole. $P_{21} (\text{m}/\text{m}^2)$ is the total length of fracture traces per unit surface area by 2D investigation using an outcrop or tunnel wall. $P_{32} (\text{m}^2/\text{m}^3)$ is the volumetric fracture intensity for a 3D density value expressed by the total surface area of fractures per unit volume. P_{32} is experimentally derivable from P_{10} or P_{21} because it cannot be directly measured owing to the lack of access to the interior sections of rock mass. For example, a 3D DFN model can be generated with a P_{10} value along a borehole or different P_{10} values along several boreholes by setting the fracture size and orientation distribution. The total fracture area over the calculation domain is computed, and P_{32} can therefore be quantified. More fractures are distributed and naturally connected with increasing P_{32} values, as shown in Fig. 2c.

Connectivity

The connectivity of fractures strongly affects the mechanical and hydraulic properties of rock mass. The most common method for connectivity measurements is percolation theory, which quantifies connectivity by the number of intersections per fracture using a percolation parameter p (Lei and Wang 2016). Connectivity increases with increasing p . For a 3D domain, p can be defined as the sum of $\pi^2 \times (\text{each fracture area} / \pi)^{1.5}$ per unit volume (Itasca 2019).

Aperture

The aperture is the opening width of a fracture, which is measured along the perpendicular direction of the fracture surface. The real aperture distribution is difficult to specify because natural fractures are irregularly shaped with variable apertures rather than simple plane and aperture data, which can be obtained from outcropped fractures and borehole TV observations of a rock wall. Gaussian, log-normal, negative power law, and uniform distributions have therefore been assumed for the aperture distribution (Lei et al. 2017), and one must be selected depending on the statistical properties of the measured data and rock-mass conditions.



Rock Fracture Pattern and Modeling, Fig. 3 Simulated fractures using GEOFRAC for a granitic area (modified from Koike et al. (2015)). (a) Perspective view of the locations of 2309 observed fractures from 26 deep boreholes, perspective views showing the distribution pattern of

the simulated continuous fractures viewed from (b) above, (c) the west, and (d) the lower hemisphere projection of Schmidt's net of a pole density distribution of the fractures. Each fracture is randomly and semi-transparently colored to discriminate the different fractures

The hydraulic conductivity (or permeability) of rock mass is strongly controlled by the fracture aperture because open fractures act as essential pathways of fluid flow. An open fracture is partly filled with minerals (e.g., quartz, calcite), which causes a rough fracture surface, impermeable portions in the pathway, and a highly variable aperture. The hydraulic aperture, which is substantially considered for characterizing hydraulic properties and fluid flow, is therefore considerably smaller than the aperture in general.

Geometry

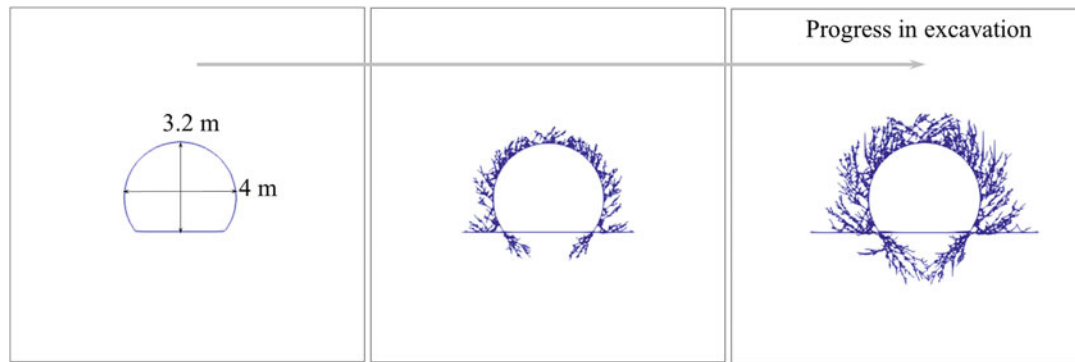
Because real fracture shapes in a rock mass are impossible to specify, they must be approximated as simple regular shapes, such as planar polygons or disks (Jing and Stephansson 2007). However, complicated shapes can form by connecting several disks that are located close to one another with similar orientations and enveloping them (Koike et al. 2012).

Fractures are modeled as a smooth parallel plate for the simplest laminar fluid. For this simplification, the cubic law is

adopted in which the flow rate through the fracture is proportional to the cube of the hydraulic aperture (Witherspoon et al. 1980). For actual rock conditions, the fracture-surface roughness must be considered because roughness significantly affects the hydromechanical behavior of the fractured rock mass (Brown 1987). The joint roughness coefficient (JRC) is a representative index for the roughness measure, which can be related to the nonlinear shear strength criterion of fracture with normal stress in engineering (Barton 1976). Because JRC is known to vary with scale, self-affine or self-similar fractal models have been used for a more accurate expression of the fracture surface (e.g., Dauskardt et al. 1990; Maximo and de Lacer 2012).

Fracture Modeling

The rock fracture pattern includes the size, orientation, distribution density, aperture in some cases, and intersection, and is typically modeled using a DFN. Numerous application software programs are available for DFN modeling (e.g.,



Rock Fracture Pattern and Modeling, Fig. 4 Fracture propagation in rock mass around a tunnel with the excavation progress simulated by the hybrid FEMDEM

FLAC3D, PFC3D), as demonstrated by the example in Fig. 2 generated by FLAC3D using different geometric parameters. These applications conventionally define the probability distributions of the fracture geometric parameters and shapes as polygons or disks for modeling. These parameters are randomly selected from the probability distributions and assigned to each fracture. Geostatistical methods have also been developed for DFN modeling to honor the location and orientation data of measured fractures (e.g., Dowd et al. 2007). GEOFRAC (GEOstatistical FRACTure simulation method; Koike et al. 2015) is one such method that considers the spatial correlations of fracture orientations and densities and forms an arbitrary fracture shape by connecting several disks, as described above (Koike et al. 2012). Regional fracture patterns can be revealed using GEOFRAC from a borehole fracture dataset that show clear dominant orientations, continuous fracture distributions, and concentrated zones of parallel and intersecting fractures (Fig. 3).

Continuum- and discontinuum-based methods have been developed to model fracture initiation and propagation under tectonic stress conditions (Lisjak and Grasselli 2014; Wolper et al. 2019). The former methods are based on the mechanics of plasticity and continuum damage by assuming continuous displacement over the calculation domain, and typically use the finite element method (FEM) for the simulation. In contrast, the latter methods assume that fractures are represented by the discontinuous displacement field, and typically use the discrete element method (DEM), discontinuous deformation analysis (DDA) method, and hybrid finite-DEM (FEMDEM) for the simulation. Figure 4 shows an example of fracture propagation around a tunnel during the excavation progress, as simulated by the hybrid FEMDEM. The change in permeability and extension of the damage zone is predictable with the fracture development. Machine learning has recently been adopted to predict fracture propagation by considering the fracture geometry and mechanics (e.g., Hsu et al. 2020).

Summary

The rock fracture pattern (i.e., fracture distribution configuration) is a control factor of the hydraulic and mechanical properties of rock mass, and is typically characterized by the fracture length, orientation, density, connectivity, aperture, and geometry. The length, orientation, and aperture distributions are approximated using the most suitable statistical and power-law distributions of the fracture survey data. The volumetric fracture intensity, P_{32} (m^2/m^3), is indispensable for the 3D density to realistically model the fracture distribution in rock mass. Although this density cannot be directly measured, it is experimentally derivable from the 1D intensity P_{10} or 2D intensity P_{21} . Percolation theory is most commonly used for connectivity measurements. The fracture shape is simply approximated as a planar polygon or disk, and self-affine or self-similar fractal models are reasonable to accurately express the fracture surface.

Considering the fracture pattern, the spatial fracture distribution is typically modeled by a DFN that uses a conventional probability distribution of the geometric parameters and the fracture shape as a polygon or disk. DFN modeling is a type of nonconditional simulation. In contrast, geostatistical methods have also been developed to honor the fracture location and orientation data obtained by measurements, which includes the incorporation of spatial correlations of fracture orientations and densities into the DFN model and the formation of an arbitrary fracture shape by connecting several disks. In addition to the above static fracture modeling, dynamic fracture modeling using continuum- and discontinuum-based methods is essential for modeling fracture initiation and propagation in a tectonic stress field. Changes in the permeability and extension of the damage zone with the fracture development are predictable by dynamic modeling using a method such as hybrid FEMDEM.

Cross-References

- [Geostatistics](#)
- [Kriging](#)
- [Machine Learning](#)
- [Statistical Rock Physics](#)
- [Stochastic Geometry in the Geosciences](#)

Bibliography

- Barton N (1976) The shear strength of rock and rock joints. *Int J Rock Mech Min Sci Geomech Abstr* 13:255–279. [https://doi.org/10.1016/0148-9062\(76\)90003-6](https://doi.org/10.1016/0148-9062(76)90003-6)
- Bonnet E, Bour O, Odling NE, Davy P, Main I, Cowie P, Berkowitz B (2001) Scaling of fracture systems in geological media. *Rev Geophys* 39:347–383. <https://doi.org/10.1029/1999RG000074>
- Brown SR (1987) Fluid flow through rock joints: the effect of surface roughness. *J Geophys Res* 92:1337–1347. <https://doi.org/10.1029/JB092IB02P01337>
- Dauskardt RH, Haubensak F, Ritchie RO (1990) On the interpretation of the fractal character of fracture surfaces. *Acta Metall Mater* 38: 143–159. [https://doi.org/10.1016/0956-7151\(90\)90043-G](https://doi.org/10.1016/0956-7151(90)90043-G)
- Dershowitz WS, Herda HH (1992) Interpretation of fracture spacing and intensity. Paper presented at the 33rd U.S. Symposium on Rock Mechanics (USRMS), Santa Fe, New Mexico, June 1992
- Dowd PA, Xu C, Mardia KV, Fowell RJ (2007) A comparison of methods for the stochastic simulation of rock fractures. *Math Geol* 39:698–714. <https://doi.org/10.1007/s11004-007-9116-6>
- Fisher NI (1993) Statistical analysis of circular data. Cambridge University Press. <https://doi.org/10.1017/CBO9780511564345>
- Hsu YC, Yu CH, Buehler MJ (2020) Using deep learning to predict fracture patterns in crystalline solids. *Matter* 3:197–211. <https://doi.org/10.1016/J.MATT.2020.04.019>
- Hyman JD, Aldrich G, Viswanathan H, Makedonska N, Karra S (2016) Fracture size and transmissivity correlations: implications for transport simulations in sparse three-dimensional discrete fracture networks following a truncated power law distribution of fracture size. *Water Resour Res* 52:6472–6489. <https://doi.org/10.1002/2016WR018806>
- Itasca Consulting Group Inc (2019) FLAC3D user manual, Version 6.0
- Jing L, Stephansson O (2007) The basics of fracture system characterization – field mapping and stochastic simulations. *Dev Geotech Eng* 85:147–177. [https://doi.org/10.1016/S0165-1250\(07\)85005-X](https://doi.org/10.1016/S0165-1250(07)85005-X)
- Koike K, Liu C, Sanga T (2012) Incorporation of fracture directions into 3D geostatistical methods for a rock fracture system. *Environ Earth Sci* 66:1403–1414. <https://doi.org/10.1007/s12665-011-1350-z>
- Koike K, Kubo T, Liu C, Masoud A, Amano K, Kurihara A, Matsuoka T, Lanyon B (2015) 3D geostatistical modeling of fracture system in a granitic massif to characterize hydraulic properties and fracture distribution. *Tectonophysics* 660:1–16. <https://doi.org/10.1016/J.TECTO.2015.06.008>
- Lei Q, Wang X (2016) Tectonic interpretation of the connectivity of a multiscale fracture system in limestone. *Geophys Res Lett* 43: 1551–1558. <https://doi.org/10.1002/2015GL067277>
- Lei Q, Latham JP, Tsang CF (2017) The use of discrete fracture networks for modelling coupled geomechanical and hydrological behaviour of fractured rocks. *Comput Geotech* 85:151–176. <https://doi.org/10.1016/J.COMPGE0.2016.12.024>
- Lisjak A, Grasselli G (2014) A review of discrete modeling techniques for fracturing processes in discontinuous rock masses. *J Rock Mech*

Geotech Eng 6:301–314. <https://doi.org/10.1016/J.JRMGE.2013.12.007>

Maximo L, de Lacer LA (2012) Fractal fracture mechanics applied to materials engineering. In: *Applied fracture mechanics*. <https://doi.org/10.5772/52511>

Witherspoon PA, Wang JSY, Iwai K, Gale JE (1980) Validity of cubic law for fluid flow in a deformable rock fracture. *Water Resour Res* 16: 1016–1024. <https://doi.org/10.1029/WR016I006P01016>

Wolper J, Fang Y, Li M, Lu J, Gao M, Jiang C (2019) CD-MPM: continuum damage material point methods for dynamic fracture animation. *ACM Trans Graph* 38:1–15. <https://doi.org/10.1145/3306346.3322949>

Rodionov, Dmitriy Alekseevich

Hannes Thiergärtner

Department of Geosciences, Free University of Berlin, Berlin, Germany



Fig. 1 D. A. Rodionov. (Photo: Thiergärtner 1984)

Biography

Dmitriy Alekseevich Rodionov was the leading Russian scientist in applications of mathematical geosciences between 1970 and 1990 and one of the 20 founding members of the International Association for Mathematical Geology (IAMG).

Born in Saratov (Russia) on October 9, 1931, he studied geochemistry at the State University in Saratov before he started a career as exploratory geologist. Participating in a

field campaign, he befriended the mathematician Prof. Yu. V. Prochorov who inspired many young geologists about using mathematics. Rodionov obtained his PhD doctorate in the field of statistical treatment of geochemical and mineralogical data. He developed quickly into the leading expert for applied mathematical geosciences in the former Soviet Union when he worked in the Moscow Institute for Mineralogy, Geochemistry, and Crystal Chemistry of Rare Elements (IMGRE) of the Academy of Science of the USSR. On the occasion of the 23rd International Geological Congress 1968 in Prague Prof. Rodionov was one of the founding members of the IAMG and was elected as one of its Councillors. Because of the former political situation, he could not really take part in IAMG activities until 1990 but he always was in close contact with colleagues from other eastern European countries. He moved to the Moscow All-Union Institute for the Economy of Mineral Resources (VIEMS) in 1969 where he worked until 1981 as the leading specialist for the development of mathematical techniques and data processing in the field of economical geology. Under his leadership, numerous mathematical approaches were introduced for geological exploration and documented in several monographs and other publications in the Russian language. He was also one of the Russian scientists who regularly participated in the biennial scientific meetings at Příbram (then Czechoslovakia) where scientists from both political blocks could meet and exchange scientific ideas. Between 1981 and 1992, Rodionov was the director of the Laboratory for Mathematical Geology of the Institute for Ore Deposit Geology, Mineralogy, Petrology, and Geochemistry (IGEM) of the USSR Academy of Sciences in Moscow. The last few years before his demise on October 2, 1994 he worked again at VIEMS.

The IAMG honored his lifetime contributions in 1994 with the William Christian Krumbein medal.

Former friends remember Dmitriy Alekseevich as a wonderful personality. All his life he remained an agile geoscientist, always connected with the science, helpful towards colleagues and open-minded toward any new development. Last but not least, Dmitriy Alekseevich inspired everybody he worked with.

Cross-References

► [Object Boundary Analysis](#)

Bibliography

- Kogan RI, Belov Yu, Rodionov DA (1983) Statistical rank criteria in geology. Nedra, Moscow. (in Russian)
- Rodionov DA (1964) Distribution functions of the content of chemical elements and minerals in eruptive rocks. Nauka, Moscow. (in Russian)

- Rodionov DA (1968) Statistical methods to mark geological objects by a complex of attributes. Nedra, Moscow. (in Russian)
- Rodionov DA (1981) Statistical solutions in geology. Nedra, Moscow. (in Russian)
- Rodionov DA, Zabelina TM, Rodionova MK (1973) Semi-quantitative analysis in biostratigraphy and palaeoecology. Nedra, Moscow. (in Russian)
- Rodionov DA, Kogan RI, Belov Yu (1979) Statistical methods classifying geological objects. VIEMS publishers, Moscow. (in Russian)
- Rodionov DA, Kogan RI, Golubeva VA (1987) Handbook for mathematical methods applied in geology. Nedra, Moscow. (in Russian)

Rodriguez-Iturbe, Ignacio

B. S. Daya Sagar

Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, Karnataka, India



Fig. 1 Professor Ignacio Rodriguez-Iturbe (1942–2022) (Courtesy of Ignacio Rodriguez-Iturbe)

Biography

Ignacio Rodriguez-Iturbe is Distinguished University Professor and TEES Eminent Professor at the Departments of Ocean Engineering, Civil Engineering, and Biological and Agricultural Engineering, Texas A&M University. He is also James S. McDonnell Distinguished University Professor Emeritus at Princeton University.

Born on 8 March 1942 in Caracas, Venezuela, and educated as a civil engineer at Universidad del Zulia in Maracaibo, he finished his Ph.D. in Colorado State University (1967) under Professor Vujica Yevjevich. After teaching in Venezuela and the United States (MIT, Iowa, Texas A&M) he was in Princeton during 2000–2018 as the James

S. McDonnell Distinguished University Professor (now Emeritus) and Professor of Civil and Environmental Engineering. He is a member of numerous academies including the National Academy of Sciences, the National Academy of Engineering, the American Academy of Arts and Sciences, the Spanish Royal Society of Sciences, the Vatican Academy of Sciences, the Venezuelan National Academy of Engineering, the Mexican Academy of Engineering, the Water Academy (Sweden), the Third World Academy of Sciences, and several others.

He is the recipient of the Stockholm Water Prize (2002), the Mexico International Prize for Science and Technology (1994), the Venezuela National Science Prize (1991), the Prince Sultan Bin Abdulaziz International Water Prize (2010), the Venezuela National Engineering Research Prize (1998), the E.O. Wilson Biotechnology Pioneer Award (2009), and several other international prizes and awards. In the United States, he received from the American Geophysical Union, the Bowie Medal (2009), the Horton Medal (1998), the Macelwane Medal (1977), the Hydrologic Sciences Award (1974), and the Langbein Lecture Award (2000). He is a Fellow of the AGU and the American Meteorological Society where he received the Horton Lecture Award (1995). From the American Society of Civil Engineers he received the Huber Research Prize (1975) and the V.T.Chow Award (2001). He has Honorary Doctor's Degrees from the University of Genova (Italy, 1992), Universidad del Zulia (Venezuela, 2003), and Universidad de Cantabria (Spain, 2011) as well as Teaching Awards in the MIT Civil and Environmental Engineering Department (1974) and Universidad del Zulia (1970).

Among many other awards, he has received the Hydrology Days Award (2002) and Distinguished Alumnus Award (1994) from Colorado State University; Peter S.Eagleson Lecturer (Consortium of Universities for Hydrologic Sciences, 2010); Enrico Marchi Memorial Lecture (Italy, 2011); Chester Kiesel Memorial Lecture (Arizona, 1991); Landolt Lecture (EPFL, Switzerland, 2009); Moore Lecture (University of Virginia, 1999); Evnin Lecture (Princeton, 2000); Smallwood Distinguished Lecture (University of Florida, 2010); Distinguished Lecturer (College of Engineering, Texas A&M, 2003); Miller Research Professor (University of California, Berkeley, 2004). He is also the namesake of the "Ignacio Rodríguez-Iturbe Prize" awarded annually for the best paper in the journal "Ecohydrology."

He is the author of four books and editor of several others. His book on "Fractal River Basins: Chance and Self-Organization" coauthored with Andrea Rinaldo is a mathematically fascinating book and presents several cases supporting the self-organized nature of river basins and related terrestrial phenomena and processes (Rodríguez-Iturbe and Rinaldo 2001). He has contributed near 300 papers in international journals and many more in Proceedings and Chapters in

books. His research interests are broadly in Surface Water Hydrology with special emphasis in Ecohydrology, Hydrogeomorphology, Hydrometeorology, and the International Trade of Virtual Water.

Bibliography

Rodríguez-Iturbe I, Rinaldo A (2001) Fractal river basins: chance and self-organization. Cambridge University Press, Cambridge, p 570

Root Mean Square

Tianfeng Chai

Cooperative Institute for Satellite Earth System Studies (CISESS), University of Maryland, College Park, MD, USA
NOAA/Air Resources Laboratory (ARL), College Park, MD, USA

Definition

Root mean square (RMS), also called the quadratic mean, is the square root of the mean square of a set of numbers.

Introduction

Unlike arithmetic mean which allows positive and negative numbers offset each other, RMS has a non-negative contribution from each. While RMS has many uses, such as RMS voltage in electrical engineering and RMS speed for gas molecules in statistical mechanics, in geoscience it is almost exclusively associated with root mean square error (RMSE) or root mean square deviation (RMSD).

Model simulation results, estimates with empirical formulas, and some measurements in question often need to be evaluated using more reliable measurements or theoretical predictions. Those quantities that need to be evaluated are first paired with the assumed "truth," that is, more accurate measurements or predictions. Then the pairwise differences are obtained as a set of deviation or error samples, d_i , $i = 1, \dots, n$. RMSD (or RMSE) is then defined as:

$$\text{RMSD} = \sqrt{\frac{1}{n} \sum_{i=1}^n d_i^2} \quad (1)$$

While the terms of RMSD and RMSE are interchangeable in practice, RMSE appears more frequently than RMSD in geoscience literature. Hereafter we will only use the term RMSE. Note that RMSE is scale dependent, and normalized

RMSE is sometimes employed to avoid the dependence on units.

RMSE and MAE

While RMSE is often used to measure the overall magnitude of error samples, it is often compared with mean absolute error (MAE), which is a simple arithmetic mean of the absolute values of errors. Clearly, RMSE and MAE measure the magnitude of error samples differently. The larger errors contribute to RMSE much more than the smaller errors due to the squaring operation. RMSE is suitable to describe a set of error samples that follow a normal (or Gaussian) distribution with a zero mean. If the error samples have a non-zero mean, the mean error (or the mean bias) is often removed before the standard deviation of the errors is calculated. In such a case, it is the standard error (SE) that should be used. Note that the degree of freedom is reduced by one when calculating the mean bias-corrected sample variance. In geoscience literature, unbiased RMSE is often used without correcting the degree of freedom change. However, the difference between unbiased RMSE and SE is minimal for typical large count of error samples in model evaluation studies. For error samples from a Gaussian distribution population, calculating mean and SE of the errors will allow one to reconstruct the actual distribution if there are enough samples. Table 1 shows RMSEs calculated for randomly generated pseudo-errors which follow a Gaussian distribution with a zero mean and unit standard deviation, similarly as those shown in Chai and Draxler (2014). As the sample size reaches 100 or above, using the calculated RMSEs one can reconstruct the error distribution close to its “truth” or “exact solution,” with its standard deviation within 5% to its truth (i.e., 1) as shown in Table 1. For errors following a normal distribution with non-zero mean, the error in the estimate of the mean is proportional to $1/\sqrt{n}$. With enough observations

for error samples, the error distribution can be reliably constructed using the estimated mean and SE. When there are not enough error samples, presenting the values of the errors themselves (e.g., in tables) is probably more appropriate than calculating RMSE or any other statistics.

Although sometimes error samples do not exactly follow a Gaussian distribution, the error distributions in most applications are often close to be normal distributions. It has been observed that many physical quantities that are affected by many independent processes typically follow Gaussian distributions. This phenomenon is commonly explained by the central limit theorem, which states that the properly normalized sum of independent random variables tends toward a normal distribution even if the original variables themselves are not normally distributed. There are also other explanations for the ubiquity of normal distributions (e.g., Lyon 2014). Note that the model evaluation studies where error samples are calculated as the differences between predictions and observations are focused here. Unlike some physical quantities which are often non-negative and may not directly follow a normal distribution without a mathematical transformation, the error samples in model evaluation studies usually closely resemble normal distributions.

In model evaluation studies, both the root mean square error (RMSE) and the mean absolute error (MAE) are regularly employed to measure the overall error magnitude. In the literature, RMSE is sometimes considered to be ambiguous as a metric for average model performance (e.g., Willmott and Matsuura 2005; Willmott et al. 2009). However, such suggestion is probably misleading. Chai and Draxler (2014) show that some of the arguments to avoid RMSE are indeed mistaken. For instance, RMSE actually satisfies the triangle inequality requirement for a distance metric, contrary to what was claimed by Willmott et al. (2009). Furthermore, the use of absolute value operation in calculating MAE makes the gradient or sensitivity of MAE with respect to certain model parameters difficult if not impossible. In fact, it is not wise to avoid either RMSE or MAE without considering the specifics of the application. The best statistics metric should be based on the actual distribution of the errors in order to provide a better performance measure. The MAE is suitable to describe uniformly distributed errors while RMSE is good to measure errors following a normal distribution. Also note that when both RMSE and MAE are calculated, RMSE is by definition never smaller than MAE. As any single statistical metric condenses a set of error values into a single number, it provides only one projection of samples and, therefore, emphasizes only a certain aspect of the error characteristics. Thus, a combination of metrics, including but certainly not limited to RMSE and MAE, is often required to better assess model performance. For errors sampled from an unknown distribution, providing more statistical moments of the

Root Mean Square, Table 1 RMSEs of randomly generated pseudo-errors which follow a Gaussian distribution with a zero mean and unit variance. For seven different sample sizes, $n = 4, 10, 100, 1000, 10,000, 100,000$, and $1,000,000$, five sets of errors are generated with different random seeds

n (error sample size)	RMSE				
	Set #1	Set #2	Set #3	Set #4	Set #5
4	0.9185	0.6543	1.4805	1.0253	0.7906
10	0.8047	1.1016	0.8329	0.9506	1.0099
100	1.0511	1.0293	1.0305	0.9961	1.0401
1000	1.0362	0.9772	1.0107	1.0004	1.0029
10,000	0.9974	0.9841	1.0142	0.9956	1.0023
100,000	1.0034	0.9998	1.0038	1.0028	1.0015
1,000,000	0.9996	1.0007	1.0003	0.9993	0.9996

model errors, such as mean, variance, skewness, and flatness, will help depict a better picture of the error variation.

Usage of RMSE

It is well known that RMSE is sensitive to outliers. In fact, the existence of outliers and their probability of occurrence are well described by the normal distribution underlying the use of the RMSE. Chai and Draxler (2014) showed that with a large amount of error samples that include some outliers, one can still closely reconstruct the original normal distribution. In addition, sensitivity to larger errors makes RMSE more discriminating as a model evaluation metric. This aspect is often more desirable when evaluating models. In data assimilation field, the sum of squared errors is often formed as the main component of the cost function to be minimized by adjusting model parameters. In such applications, penalizing the outliers associated with large discrepancies between model and observations through the defined least-square terms proves to be very effective in improving model performance. In practice, it might be necessary to throw out the outliers that are several orders larger than the other samples when calculating the RMSE, especially when the number of samples is limited.

Summary

Root mean square errors (RMSEs) are often used for model evaluation studies in geoscience. The arguments to choose MAE over RMSE in literature are mistaken. Compared with MAE, RMSE is suitable to measure errors which follow a normal distribution. As a more discriminating metric because of its sensitivity to large errors, RMSE is probably more desirable in model evaluations. In addition, RMS is closely related to cost function formulation in data assimilation field.

References

- Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE)? – arguments against avoiding RMSE in the literature. *Geosci Model Dev* 7:1247–1250. <https://doi.org/10.5194/gmd-7-1247-2014>
- Lyon A (2014) Why are normal distributions normal? *Br J Philos Sci* 65(3):621–649. <https://doi.org/10.1093/bjps/axs046>
- Willmott C, Matsuura K (2005) Advantages of the Mean Absolute Error (MAE) over the Root Mean Square error (RMSE) in assessing average model performance. *Clim Res* 30:79–82
- Willmott CJ, Matsuura K, Robeson SM (2009) Ambiguities inherent in sums-of-squares-based error statistics. *Atmos Environ* 43:749–752

Sampling Importance: Resampling Algorithms

Anindita Dasgupta and Uttam Kumar
Spatial Computing Laboratory, Center for Data Sciences,
International Institute of Information Technology Bangalore
(IIITB), Bangalore, Karnataka, India

Definition

Sampling is a technique from which information about the entire population can be inferred. In case of remote sensing (RS) and geographic information system (GIS), training and test data are created during the early part of the project using random sampling while post-classification stratified random sampling is employed for different land-cover classes behaving as strata. RS images may have distortions that may be corrected using georectification. Georectification or georeferencing is the process of establishing mathematical relationship between the addresses of pixels in an image with the corresponding coordinates of those pixels on another image or map or ground. Georeferenced imagery is used to extract distance, polygon area, and direction information accurately. Georeferencing indicates that resampling of an image requires matching not only to a reference image but also to reference points that correspond to specific known locations on the ground. Resampling is related to but distinct from georeferencing. It is the application of interpolation to bring an image into registration with another image or a planimetrically correct map. The three methods used for resampling are nearest-neighbor, bilinear interpolation and cubic convolution. The nearest-neighbor approach uses the value of the closest input pixel for the output pixel value. Bilinear interpolation uses the weighted average of the nearest four pixels to the output pixel. Cubic convolution approach uses the weighted average of the nearest 16 pixels to the output pixel.

Introduction

Studying an entire population is both time-consuming and resource-intensive. Therefore, it is advisable to take a sample from the population of interest, and on the basis of this sample, information about the entire population is inferred. A sample is a portion, piece, or segment that is representative of a whole population. Conclusions about an entire population are drawn based on the sample information through statistical inference. A sample generally provides an accurate picture of the population from which it is drawn. It is random; each individual in the population has an equal chance of being selected (Groves et al. 2011). Different sampling schemes are explained next.

Simple random sampling: In this technique, units are independently selected one at a time until the desired sample size is achieved. Each study unit in the finite population has an equal chance of being included in the sample. This method of sampling has the advantage of representativeness, freedom from bias, and ease of sampling and analysis. The disadvantages of this method are errors in sampling, and additional time and labor requirements.

Systematic sampling: This method has evenly distributed samples over the entire population. When N units exist in the population and n units are to be selected, then sampling interval $R = N/n$. The advantage is that it is spatially well distributed with small standard errors. The disadvantages of this method are that it is less flexible to increase or decrease the sampling size and is not applicable for fragmented strata.

Stratified sampling: It involves grouping the population of interest into strata to estimate the characteristics of each stratum and improve the precision of an estimate for the entire population (Cochran 2007). Stratified random sampling is useful for data collection if the population is heterogeneous, as the entire population is divided into a number of homogeneous groups known as strata. The sampling error depends on the population variance within stratum but not between the strata (Singh and Masuku 2014). The advantage is that it

allows specifying the sample size within each stratum and different sampling design for each stratum; stratification may increase precision. The major drawbacks are that it yields large standard error if the sample size selected is not appropriate. Moreover, it is not effective if all the variables are equally important.

Cluster sampling: It involves grouping of the spatial units or objects sampled into clusters. All observations in the selected clusters are included in the sample. This method can reduce the time and expense of sampling by reducing travel distance. The disadvantage is that clustering can yield higher sampling error. Further, it can be difficult to select representative clusters.

Multistage random sampling: In this sampling, the region is separated into different subsets that are randomly selected (first stage), and then the selected subsets are randomly sampled (second stage). This is similar to stratified random sampling, except that with stratified random sampling, each stratum is sampled. However, there is stronger clustering here than simple random sampling.

Sampling in Geospatial Data Analysis

Let's understand sampling in the context of geospatial data analysis in RS applications. In case of land-cover classification from remotely sensed data, training and test reference data are randomly sampled. When the sample size is sufficient, simple random sampling provides adequate estimate of the population. While using stratified sampling, the caution is not to over-estimate the population parameter. Stratified sampling may be preferred as minimum number of samples is taken from each stratum. For example, if we consider land-cover classes such as built-up and agriculture, then samples have to be taken from both the classes. Test sample can be collected using Global Positioning System (GPS) to identify ground control points (GCPs). During early part of the project, random sampling can be employed to collect data followed by post-classification random sampling within the strata.

Sample Size

Sample size is an important part of any investigation or research project where the study seeks to make inferences about the population from a sample. When the population is heterogeneous, stratified sampling can be used and different sample sizes are collected for each population. Census method is a technique of statistical enumeration where all members of the population are studied and the sample size is equal to the population size.

Three criteria determine the appropriate sample size: level of precision, level of confidence or risk, and degree of

variability in the attributes being measured (Miaoulis and Michener 1976).

- **Level of precision:** Sampling error, also known as the level of precision, is the range in which the true value of the population is estimated (Singh, 2014). This range is expressed in percentage (e.g., $\pm 5\%$).
- **Confidence interval:** In a normal distribution, approximately 95% of the sample values are within two standard deviations of the true population value. This confidence interval is also known as the risk of error in statistical hypothesis testing.
- **Degree of variability:** When the variables exist with more homogeneous population, the sample size required is smaller. If the population is heterogeneous, larger sample size is required to obtain a given level of precision (Singh, 2014).

In case of RS data, Fitzpatrick Lins (1981) suggested that sample size N to be used to access accuracy of land-cover classified map should be determined from the formula of binomial probability theory:

$$N = \frac{Z^2(p)(q)}{E^2} \quad (1)$$

where p is the expected percent accuracy, $q = (100 - p)$, E is the allowable error, and $Z = 2$ from standard normal deviation of 1.96 from 95% two-side confidence interval. For example, if the expected accuracy is 85% and allowable error is 5%, then

$$N = \frac{2^2(85)(15)}{5^2} = 204$$

Ground Control Points (GCPs)

Remotely sensed imagery exhibits internal and external geometric error, and these random and systematic distortions are corrected by analyzing well-distributed Ground Control Points (GCPs) occurring in an image. Hence, remotely sensed data should be preprocessed to remove geometric distortion so that individual picture elements (pixels) are in their proper planimetric (x, y) map locations. GCP is a location on the surface of the Earth (e.g., a road intersection, rail-road crossings, towers, buildings, etc.) that can be identified on the remotely sensed imagery and located accurately on a map (Jenson, 1986). The locations of the output pixels are derived from locational information provided by GCPs, places on the input image that can be located with precision on the ground and on planimetrically correct maps. The locations of these points establish the geometry of the output image and its

relationship to the input image. Thus, this first step establishes the framework of pixel positions for the output image using the GCPs.

Georectification or Georeferencing

Georectification allows RS-derived information to be related to the other thematic information in GIS or spatial decision support systems (SDSS). It is the process of establishing mathematical relationship between the addresses of pixels in an image with the corresponding coordinates of those pixels on another image or map or ground. In the correction process, numerous GCPs are located in terms of their image coordinates (column and row numbers) on the distorted image and in terms of their ground coordinates. This involves fitting the coordinate system of an image to that of the second image of the same area. The transformation of remotely sensed images so that it has a scale and projections of a map is called geometric correction (Lillesand, 2015). Geometrically corrected imagery can be used to extract distance, polygon area, and direction information accurately.

In other words, rectification involves identification of geometric relationship between input pixel location (column and row) and associated map coordinates of the same point (x,y). During rectification, GCPs are selected and polynomial equations are fitted using least-squares technique. Least-squares regression is used to determine coefficients for two coordinate transformation equations that can be used to interrelate the geometrically correct (map) coordinates and the distorted image coordinates (Jenson, 1986). The correct image coordinates for any map position can be precisely estimated with the help of the coefficients from these equations:

$$x = f_1(X, Y) \quad y = f_2(X, Y)$$

where (x, y) is the correct map coordinates (column, row), (X, Y) is the distorted image coordinates, and f_1 and f_2 are the transformation functions.

GCPs should be uniformly spread across the image and present in a minimum number, depending on the type of transformation to be used. Conformal, affine, projective, and polynomial transformations are frequently used in RS image rectification. Higher-order transformation may be required such as projective and polynomial transformations to geocode an image and make it fit with another image. Depending upon the distortions in the imagery, the number of GCPs used, and their locations relative to one another, complex polynomial equations are also used. The degree of complexity of the polynomial is expressed as the order of the polynomial. The order is the highest exponent used in the polynomial.

$$x' = a_0 + a_1X + a_2Y \quad (2)$$

$$y' = b_0 + b_1X + b_2Y \quad (3)$$

where x' y' are the rectified positions (in output image), and x , y are the positions in the input distorted image.

When the six coefficients a_0 , a_1 , a_2 , b_0 , b_1 and b_2 are known, then it is possible to transfer the pixel values from the original to the rectified output image. However, before applying rectification to the entire dataset, it is important to determine how well the six coefficients derived from the least-squares regression of the initial GCPs account for the geometric distortion in the input image. The method used involves computing root mean square error (RMSE) for each of the GCPs. RMSE is the distance between the input (source or measured) location of a GCP and the retransformed (or computed) location for the same GCP (Jenson, 1986). RMS error is computed with a Euclidean distance measure, and it is a way to measure distortion for each control point.

$$\text{RMS error} = \left((x' - x_{\text{orig}})^2 + (y' - y_{\text{orig}})^2 \right)^{1/2} \quad (4)$$

where x_{orig} and y_{orig} are the original column and row coordinates of the GCP in the image, and x' and y' are the computed coordinate in the distorted image. The square root of the standard deviation represents measure of accuracy. After each computation of a transformation, the total RMSE reveals that a given set of control points exceeds the threshold. The GCP that has the highest individual error is deleted from the analysis, and the coefficients and RMSE for all the points are recomputed.

Image Resampling

It is the application of interpolation to bring an image into registration with another image or a planimetrically correct map. Resampling is related but distinct from georeferencing. Georeferencing indicates that resampling of an image requires matching not only to a reference image but also to reference points that correspond to specific known locations on the ground (Campbell and Wynne 2011).

Interpolation is applied to bring an image into registration with another image or a planimetrically correct map. Pixel brightness values (BVs) must be determined. There is no direct one-to-one relationship between the movement of input pixel values to output pixel locations. A pixel in the rectified output image often requires a value from the input pixel grid that does not fall neatly on a row-and-column coordinate. When this occurs, there must be some mechanism for determining the BV to be assigned to the output-rectified

pixel. This process is called intensity interpolation. Various popular resampling techniques used in RS applications are mentioned below.

- **Nearest-Neighbor**

The nearest-neighbor approach uses the value of the closest input pixel for the output pixel value. The pixel value occupying the closest image file coordinate to the estimated coordinate will be used for the output pixel value in the georeferenced image.

Let's try to understand this with the help of Fig. 1. The green grid is considered to be the output image to be created. To determine the value of central pixel (highlighted in light green box), in the nearest-neighbor approach, the value of the nearest original pixel is assigned, i.e., the value of black pixel in this example. In other words, the value of each output pixel is assigned on the basis of digital number closest to the pixel in the input matrix. This method has the advantage of simplicity and ability to preserve original input values as the output values. While resampling methods tend to average surrounding values, this may be an important consideration when discriminating between vegetation types or locating boundaries. This is easily computed and faster than other techniques. The disadvantage is that image has a rough appearance relative to the original unrectified data due to "stair-stepped" effect (see Fig. 2) (Bakx et al. 2012).

- **Bilinear Interpolation**

This method uses the weighted average of the nearest four pixels to the output pixel. Weighted mean is calculated for the four nearest pixels in the original image (here, it is represented as dark gray and black pixels) (see Fig. 1).

$$BV_{wt} = \frac{\sum_{k=1}^4 \frac{Z_k}{D_k^2}}{\sum_{k=1}^4 \frac{1}{D_k^2}} \quad (5)$$

where Z_k are the surrounding 4 data point values, and D_k^2 are the distances squared from the point in question (x', y') to the data points.

The advantage of this technique is that the stair-step effect caused by the nearest-neighbor approach is reduced and this image looks smooth. The disadvantage is that original data is altered and contrast is reduced by averaging neighboring values together. It is computationally more expensive than nearest-neighbor.

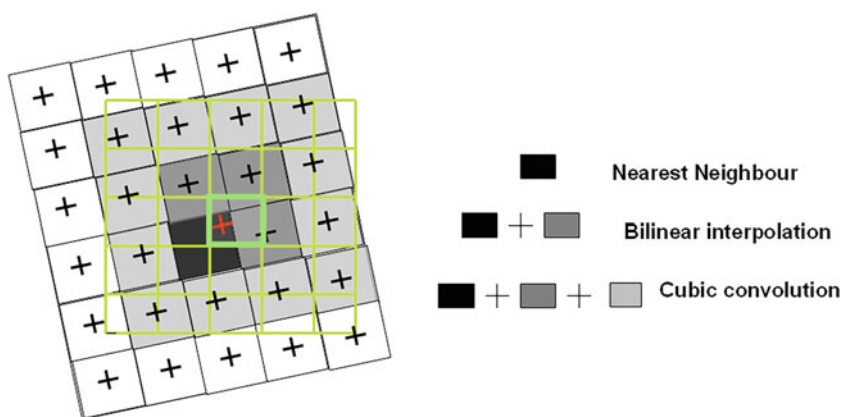
- **Cubic Convolution**

Cubic convolution approach uses the weighted average of the nearest 16 pixels to the output pixel. In Fig. 1, it is represented as a black and gray pixel in the input image.

Sampling Importance:

Resampling Algorithms,

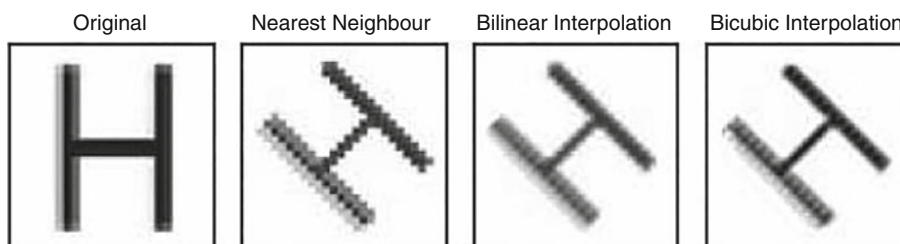
Fig. 1 Principle of resampling using nearest-neighbor, bilinear interpolation, and cubic convolution. (Bakx et al. 2012)



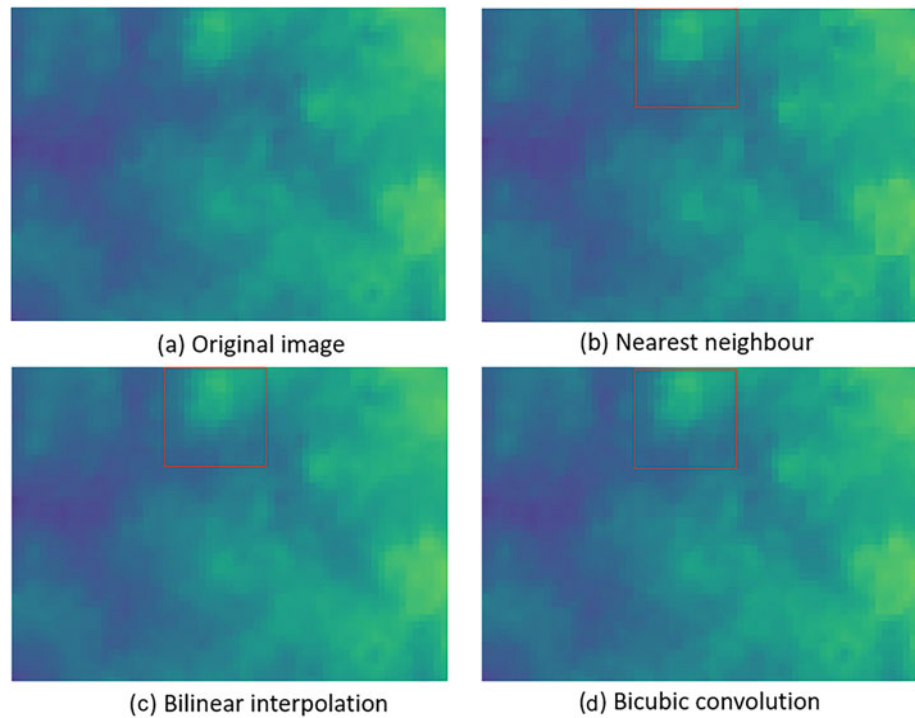
Sampling Importance:

Resampling Algorithms,

Fig. 2 Nearest-neighbor, bilinear interpolation, and bicubic convolution. (Bakx et al. 2012)



**Sampling Importance:
Resampling Algorithms,
Fig. 3** Resampled images



The output is similar to bilinear interpolation, but the smoothing effect caused by averaging of surrounding input pixel values is more dramatic.

stepped effect is reduced and images are subsequently smoothed with bilinear and bicubic interpolation.

$$BV_{wt} = \frac{\sum_{k=1}^{16} \frac{Z_k}{D_k^2}}{\sum_{k=1}^{16} \frac{1}{D_k^2}}$$

where Z_k are the surrounding 16 data point values, and D_k^2 are the distances squared from the point in question ($x' y'$) to the data points.

The advantage is that the stair-step effect caused by the nearest-neighbor approach is reduced and the image looks much smoother (Fig. 2). The disadvantages of this method are same as that of the bilinear interpolation. It is computationally more expensive than nearest-neighbor or bilinear interpolation.

Case Study of Resampling in Remote Sensing Applications

In the following experiment, digital elevation model (DEM) of a part of Bangalore City, India, at a spatial resolution of 150 m is considered. The stair-stepped effect of nearest-neighbor is visible in Fig. 3b. It is distinctly seen within the red rectangular bonding box. Similarly, in Fig. 3c, d, the stair-

(6) Conclusions

Resampling is the application of interpolation technique to bring an image into registration with another image or a planimetrically correct map. Resampling techniques used for geo-registration are nearest-neighbor, bilinear interpolation and cubic convolution. After applying these techniques, the result obtained was observed. Nearest-neighbor was easily computed and faster than other techniques. However, the stair-stepped effect of nearest-neighbor was prominent, while the same effect reduced and images were subsequently smoothed in bilinear and bicubic interpolation approaches. However, in bilinear interpolation and cubic convolution, the original image values got altered and the contrast reduced by averaging neighboring values. These two techniques are computationally more expensive than nearest-neighbor.

Cross-References

- [Cluster Analysis and Classification.](#)
- [Geographical Information Science.](#)
- [Remote Sensing.](#)

Bibliography

- Bakx JPG, Janssen L, Schetselaar EM, Tempfli K, Tolpekin VA (2012) Image analysis. In: The core of GIScience: a systems-based approach. University of Twente, Faculty of Geo-Information Science and Earth Observation (ITC), pp 205–226
- Campbell JB, Wynne RH (2011) Introduction to remote sensing. Guilford Press
- Cochran WG (2007) Sampling techniques. Wiley
- Fitzpatrick-Lins K (1981) Comparison of sampling procedures and data analysis for a land-use and land-cover map. Photogram Eng Remote Sensing 47(3):343–351
- Groves RM, Fowler FJ Jr, Couper MP, Lepkowski JM, Singer E, Tourangeau R (2011) Survey methodology, vol 561. Wiley, New York
- Jensen JR (1986) Introductory digital image processing: a remote sensing perspective. Univ. of South Carolina, Columbus
- Lillesand T, Kiefer RW, Chipman J (2015) Remote sensing and image interpretation. Wiley, New York
- Miaoulis G, Michener RD (1976) An introduction to sampling. Iowa: endall/Hunt Publishing Company, Dubuque
- Singh AS, Masuku MB (2014) Sampling techniques & determination of sample size in applied statistics research: an overview. Int J Econ Comm Manage 2(11):1–22

Scaling and Scale Invariance

S. Lovejoy

Physics, McGill University, Montreal, Canada

Definition

Scaling is a scale symmetry that relates small and large scales of a system in scale-free, power law manner: without a characteristic scale. Because it is a symmetry, in systems with structures/fluctuations over wide ranges of scale, it is the simplest such relationship.

In one dimension, the notion of scale is usually taken as the length of an interval, and in two or higher dimensions, the Euclidean distance is often used. However, the latter is isotropic: It is restricted to “self-similar” fractals, multifractals. Most geosystems are stratified in the vertical and have other anisotropies so that the scale notion must be broadened. “Generalized Scale Invariance” does this with two elements: a definition of the unit scale (all the unit vectors) defining the scales of the nonunit vectors with an operator that changes scale by a given scale ratio. This scale-changing operator forms a group whose generator is a scale-invariant exponent.

Systems that are symmetric with respect to these scale changes may be geometric sets of points (fractals) or fields (multifractals). The relationship between scales is generally (but not necessarily) statistical and involves additional scaling relationships and corresponding scale-invariant exponents. These are often fractal dimensions, or codimensions (sets),

or corresponding dimension, codimension functions (multifractals).

Scaling Sets: Fractal Dimensions and Codimensions

Geosystems typically display structures spanning huge ranges of scale in space, in time, and in space-time. The relationship between small and large, fast and slow processes is fundamental for characterizing the dynamical regimes – for example, weather, macroweather, and climate – as well as for developing corresponding mathematical and numerical models. Since scaling can be formulated as a symmetry principle, the simplest assumption about such relationships is that they are connected in a scaling manner governed by scale-invariant exponents.

In general terms, a system is scaling if there exists a “scale free” (power law) relationship (possibly deterministic, but usually statistical) between fast and slow (time) or small and large (space, space-time). If the system is a geometric set of points – such as the set of meteorological measuring stations (Lovejoy et al. 1986) – then the set is a fractal set and the number of points n in a circle radius L varies as:

$$n(L) \propto L^D \quad (1)$$

where the (scale invariant) exponent D is the fractal dimension (Mandelbrot 1977). Consider instead, the density of points $\rho(L)$:

$$\rho(L) \propto n(L)/L^d \approx L^{-c}; \quad c = d - D \quad (2)$$

where d is the dimension of space (in this example, stations on the surface of the Earth, $d = 2$ and empirically, $D \approx 1.75$). The exponent of ρ characterizes the sparseness of the set; its exponent c is its fractal *codimension*. For geometric sets of points, it is the difference between the dimension of the embedding space and the fractal dimension of the set. More generally, codimensions characterize probabilities and therefore statistics; they are usually more useful than fractal dimensions (see below). The distinction is important for considering stochastic processes that are typically defined on infinite dimensional probability spaces so that $d, D \rightarrow \infty$ even though c remains finite. Notice that whereas $n(L)$ and $\rho(L)$ are scaling, their exponents D and c are scale invariant; the two notions are closely linked.

Scaling Functions, Fluctuations, and the Fluctuation Exponent H

Geophysically interesting systems are typically not sets of points but rather scaling fields such as the rock density or temperature $f(\underline{r}, t)$ at the space-time point $(\underline{r}, t)$. (We will generally use the term “scaling fields,” but for multifractals, the more precise notion is of singular densities of multifractal

measures). In such a system, some aspect – most often a suitably defined fluctuation – Δf – has statistics whose small and large scales are related by a scale-changing operation that involves only the scale ratios: Over the scaling range, the system has no characteristic size.

In one dimension – temporal series $f(t)$ or spatial transects, $f(x)$ – scaling can be expressed as:

$$\Delta f(\Delta t) \stackrel{d}{=} \varphi_{\Delta t} \Delta t^H \quad (3)$$

where Δf is a fluctuation, Δt is the time interval over which the fluctuation occurs, H is the fluctuation exponent, and $\varphi_{\Delta t}$ is a random variable that itself generally depends on the scale (for transects, replace Δt by the spatial interval Δx). The sign $\stackrel{d}{=}$ is in the sense of random variables; this means that the random fluctuation $\Delta f(\Delta t)$ has the same probability distribution as the random variable $\varphi_{\Delta t} \Delta t^H$. In one dimension, the notion of scale can be taken to be the absolute temporal or spatial interval Δt , Δx ; for processes in two or higher dimensional spaces, the definition of scale itself is nontrivial and we need Generalized Scale Invariance (GSI) discussed later.

Although the general framework for defining fluctuations is wavelets, the most familiar fluctuations are either $\Delta f(\Delta t)$ taken as a scale Δt difference: $\Delta f(\Delta t) = f(t) - f(t - \Delta t)$, or a scale Δt anomaly $f'_{\Delta t}$, which is the average over the interval Δt of the series $f' = f - \bar{f}$ whose overall mean $\bar{f}$ has been removed. We suppress the t dependence of Δf since we consider the case where the statistics are independent of time or space: They are respectively statistically stationary or statistically homogeneous. Physically, this is the assumption that the underlying physical processes are the same at all times and everywhere in space. Difference fluctuations are typically useful when average fluctuations increase with scale ($1 > H > 0$), whereas anomaly fluctuations are useful when they decrease with scale ($-1 < H < 0$). A simple type of fluctuation that covers both ranges (i. e., $-1 < H < 1$) is the Haar fluctuation (Lovejoy and Schertzer 2012) that appears to be adequate for almost all geoprocesses.

To see that fluctuations obey the scaling property eq. 3, consider the simplest case where φ is a random variable independent of resolution. This “simple scaling” (Lamperti 1962) is respected, for example, by regular Brownian motion in which case $H = 1/2$, $\Delta f(\Delta t)$ is a difference and φ is a Gaussian random variable. The extensions to cases $0 < H < 1$ are called “fractional Brownian motions” (fBm, Kolmogorov 1940) and to the first differences (increments) of fBm, “fractional Gaussian noises” (fGn) with $-1 < H < 0$ (Mandelbrot and Van Ness 1968). In fGn, $\Delta f(\Delta t)$ must be taken as an anomaly fluctuation $\bar{f}'_{\Delta t}$, Haar fluctuation, or other appropriate fluctuation definition.

Although Gaussian statistics are commonly assumed, in scaling processes, they are in fact exceptional. In the more usual case, at fixed scales, the “tails” (extremes) of the

probability distribution of φ is also a power law. Scaling in probability space means that for large enough thresholds s , the probability of a random φ exceeding s is $Pr(\varphi > s) \approx s^{-q_D}$. q_D is a critical exponent since statistical moments $\langle \varphi^q \rangle$ of order $q > q_D$ diverge (the notation “Pr” indicates probability, “ $\langle \rangle$ ” indicates statistical averaging, and the “ D ” subscript is because the value of the critical exponent generally depends on the dimension of space over which the process is averaged). Power law probabilities (with possibly any $q_D > 0$) are a generic property of multifractal processes; more classically, they are also a property of stable Levy variables with Levy index $0 < \alpha < 2$ (for which $q_D = \alpha$; the exceptional $\alpha = 2$ case is the Gaussian, all of whose moments converge so that $q_D = \infty$). Probability distributions with power law extremes can give rise to events that are far larger than those predicted from Gaussian models, indeed they can be so strong that they have sometimes been termed “black swans” (Adapted from Taleb 2010 who originally termed such extreme events “grey swans”). Empirical values of q_D in the range 3–7 for geofields ranging from the wind, precipitation, temperature, and seismicity have been reported in dozens of papers, (many are reviewed in ch.5 of Lovejoy and Schertzer 2013 that also includes the theory).

To see why Eq. 3 has the scaling property, consider the simplest case where the fluctuations in a temporal series $f(t)$ follow eq. 1, but with φ a random variable independent of resolution: $\varphi_{\Delta t} \stackrel{d}{=} \varphi$. If we denote $\lambda > 1$ as a scale ratio, it is easy to see that the statistics of large-scale ($\Delta f(\lambda \Delta t)$) and small-scale ($\Delta f(\Delta t)$) fluctuations are related by a power law:

$$\Delta f(\lambda \Delta t) \stackrel{d}{=} \lambda^H \Delta f(\Delta t) \quad (4)$$

Equation 4 directly shows the scaling property relating the fluctuations differing by a scale ratio λ . Since the scaling exponent H is the same at all scales, it is scale invariant.

Here and below, we treat time series and spatial sections (transects) without distinction, their scales are simply absolute differences in time Δt , or in space Δx . This ignores the sometimes important difference that – due to causality – the sign of Δt (i.e., whether the interval is forward or backward in time) can be important; whereas in space one usually assumes left-right symmetry (the sign of Δx is not important), here we treat both as absolute intervals (see Marsan et al. 1996).

Equation 4 relates the probabilities of small and large fluctuations; it is usually easier to deal with the deterministic equalities that follow by taking q^{th} order statistical moments of Eq. 3 and then averaging over a statistical ensemble (indicated by “ $\langle \rangle$ ”):

$$\langle (\Delta f(\Delta t))^q \rangle = \langle \varphi_{\Delta t}^q \rangle \Delta t^{qH} \quad (5)$$

The equality is now a usual, deterministic one. Eq. 5 is the general case where the resolution Δt is important for the

statistics of ϕ . In general, $\phi_{\Delta t}$ is a random function or field averaged at resolution Δt ; if it is scaling, its statistics will follow:

$$\langle \phi_{\lambda}^q \rangle = \lambda^{K(q)}; \quad \lambda = \tau/\Delta t \geq 1 \quad (6)$$

where τ is the “outer” (largest) scale of the scaling regime satisfied by the equation, and λ is a convenient dimensionless scale ratio. $K(q)$ is a convex ($K''(q) \geq 0$) exponent function; since the mean fluctuation is independent of scale ($\langle \phi_{\lambda} \rangle = \lambda^{K(1)} = \text{const}$), we have $K(1) = 0$.

Equation 6 describes the statistical behavior of cascade processes; these are the generic multifractal processes. Since large q moments are dominated by large, extreme values, and small q moments by common, typical values, $K(q) \neq 0$ implies a different scaling exponent for each statistical moment q : “multiscaling.” Fluctuations of various amplitudes change scale with different exponents, and each realization of the process is a multifractal: The values that exceed fixed thresholds are fractal sets. Increasing the threshold defines sparser and sparser exceedance sets (see the “spikes” in Figs. 1a and 2); the fractal dimensions of these sets decrease as the threshold is increased; this is discussed further in the next section. In general, $K(q) \neq 0$ is associated with intermittency, a topic we treat in more detail in Sect. 2.

Structure Functions and Spectra

Returning to the characterization of scaling series and transects via their fluctuations, we can combine eqs. 5, 6, to obtain:

$$S_q(\Delta t) = \langle (\Delta f(\Delta t))^q \rangle = \langle \phi_{\Delta t}^q \rangle \Delta t^{qH} \propto \Delta t^{\xi(q)};$$

$$\xi(q) = qH - K(q) \quad (7)$$

where S_q is the q^{th} order (“generalized”) structure function and $\xi(q)$ is its exponent. The structure functions are scaling since the small and large scales are related by a power law:

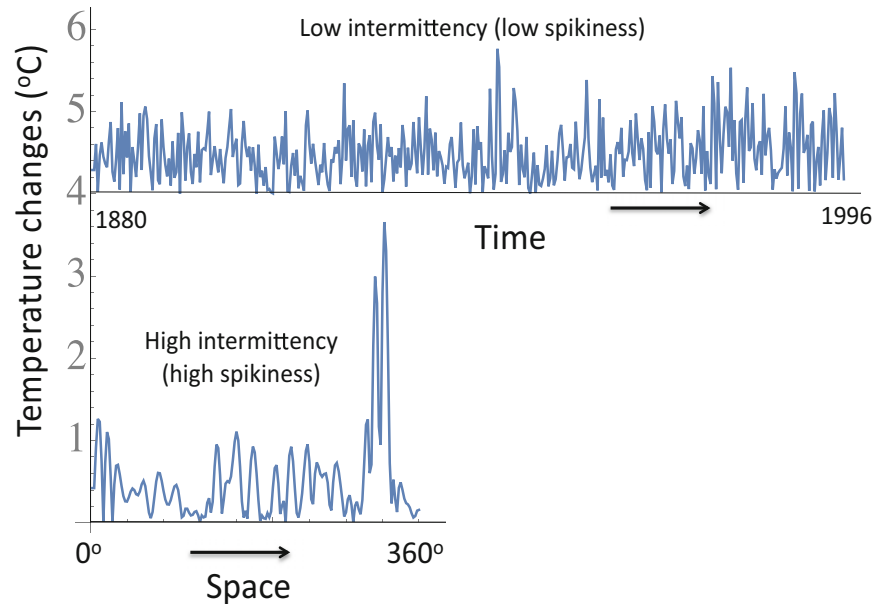
$$S_q(\lambda \Delta t) = \lambda^{\xi(q)} S_q(\Delta t) \quad (8)$$

When $K(q) \neq 0$, different moments change differently with scale so that, for example, the root mean square fluctuation has statistical moment $\xi(2)/2$ which can readily be smaller than the exponent of the first-order moment: $\xi(1) - \xi(2)/2 = K(2)/2 > 0$ (since $K(1) = 0$ and for all q , $K''(q) > 0$). This is shown graphically in Fig. 1a, b. A useful way of quantifying this deviation is the derivative at the mean: $C_1 = K'(1)$ where $C_1 \geq 0$ is the codimension of the mean fluctuation (see below). Typical values of C_1 are ≈ 0.05 – 0.15 for turbulent quantities (wind, temperature, humidity, and pressure (Lovejoy and Schertzer 2013) as well as topography, susceptibility, and rock density (Lovejoy and Schertzer 2007). It is significant that the Martian atmosphere has nearly identical exponents (H , C_1) to Earth (wind, temperature, and pressure, (Chen et al. 2016), and Martian topography is similarly close to Earth’s (Landais et al. 2019).

Precipitation and seismicity are notably much more extreme with $C_1 \approx 0.4$ (Lovejoy et al. 2012), ≈ 1.3 (Hooze et al. 1994). We could note that whereas the moments are “scaling,” the exponent functions such as $K(q)$, $\xi(q)$, etc. are “scale invariant.”

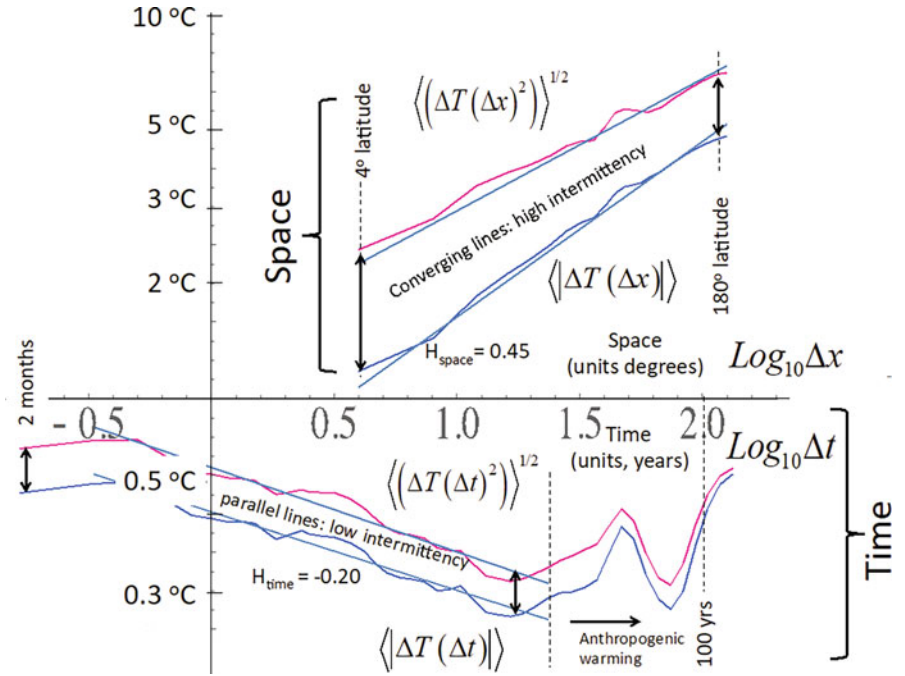
Scaling and Scale Invariance,

Fig. 1a A comparison of temporal and spatial macroweather series at 2° resolution. The top are the absolute first differences of a temperature time series at monthly resolution (from 80° E, 10° N, 1880–1996, annual cycle removed, displaced by 4°C for clarity), and the bottom is the series of absolute first differences (“spikes”) of a spatial latitudinal transect (annually averaged, 1990 from 60° N) as a function of longitude. One can see that while the top is noisy, it is not very “spikey”



Scaling and Scale Invariance,

Fig. 1b The first order and RMS Haar fluctuations of the series transect in Fig. 1a. One can see that in the spikey transect (space, top), the first order and RMS statistics converge at large lags (Δx), and the rate of the converge is quantified by the intermittency parameter C_1 . The time series (bottom) is less spikey, and it converges very little and has low C_1 (see Fig. 1a, top). The break in the scaling at ≈ 20 years is due to the dominance of anthropogenic effects at longer time scales. Quantitatively, the intermittency parameters near the mean are $C_1 \approx 0.12$ (space), $C_1 \approx 0.01$ (time)



Equations 7 and 8 are the statistics of (real space) fluctuations; however, it is common to analyze data with Fourier techniques. These yield the power spectrum:

$$E(\omega) \propto \langle |\tilde{f}(\omega)|^2 \rangle \quad (9)$$

where $\tilde{f}(\omega)$ is the Fourier Transform of $f(t)$, and ω is the frequency. Due to “Tauberian theorems” (e.g., Feller 1971), power laws in real space are transformed into power laws in Fourier space, hence for scaling processes:

$$E(\omega) \approx \omega^{-\beta} \quad (10)$$

where β is the spectral exponent. Due to the Wiener-Khinchin theorem, the spectrum is the Fourier transform of the autocorrelation function, itself closely related to $S_2(\Delta t)$. We therefore obtain the general relation:

$$\beta = 1 + \xi(2) = 1 + 2H - K(2) \quad (11)$$

(Technical note: Scaling is generally only followed over a finite range of scales, and unless there are cut-offs, there are generally low- or high-frequency divergences. However, if the scaling holds over wide enough ranges, then scaling in real space does imply scaling of the spectrum and vice versa, and if the scaling holds over a wide enough range, then eq. 11 relates their exponents).

Returning now to the case of simple scaling, we find $\langle \varphi^q \rangle = B_q$ where B_q is a constant independent of scale (Δt); hence we have $K(q) = 0$ and $S_q(\Delta t) \propto \Delta t^{qH}$ so that:

$$\xi(q) = qH \quad (12)$$

i.e., $\xi(q)$ is a linear function of q . This “linear scaling” with $\beta = 1 + 2H$ is also sometimes called “simple scaling.” Linear scaling arises from scaling linear transformations of noises; the general linear scaling transformation is a power law filter (multiplication of the Fourier Transform by ω^{-H}) which is equivalently a fractional integral ($H > 0$) or fractional derivative ($H < 0$). Fractional integrals of order $H + 1/2$ of Gaussian white noises yield fBm ($1 > H > 0$) and fGn ($-1 < H < 0$). The analogous Levy motions and noises are obtained by the filtering of independent Levy noises (in this case, $\xi(q)$ is only linear for $q < \alpha < 2$; for $q > \alpha$, the moments diverge so that both $\xi(q)$ and $S_q \rightarrow \infty$).

The more general “nonlinear scaling” case where $K(q)$ is nonzero is associated with fractional integrals or derivatives of order H of scaling multiplicative (not additive) random processes (cascades, multifractals). These fractional integrals ($H > 0$) or derivatives ($H < 0$) filter the Fourier Transform of φ by ω^{-H} ; this adds the extra term qH in the structure function exponent: $\xi(q) = qH - K(q)$.

In the literature, the notation “ H ” is not used consistently. It was introduced in honor of Edwin Hurst a pioneer of long memory processes sometimes called “Hurst phenomena” (Hurst 1951). Hurst introduced the rescaled range exponent notably in the study of Nile river streamflow records. To

explain Hurst's findings, Mandelbrot and Van Ness [1968] developed Gaussian scaling models (fGn, fBm) and introduced the symbol " H ." At first, this represented a "Hurst exponent," and they showed that for fGn processes, it was equal to Hurst's exponent. However, by the time of the landmark "Fractal Geometry of Nature" (Mandelbrot 1982), the usage was shifting from a scaling exponent to a model specification: the "Hurst parameter." In this new usage, the symbol H was used for both fGn and its integral fBm, even though the fBm-scaling exponent is larger by one. To avoid confusion, we will call it H_M . Subsequently, a mathematical literature has developed using H_M with $0 < H_M < 1$ to parametrize both the process (fGn) and its increments (fGn). However, also in the early 1980s (Grassberger and Procaccia 1983; Hentschel and Procaccia 1983; Schertzer and Lovejoy 1983), much more general scaling processes with an infinite hierarchy of exponents – multifractals – were discovered clearly showing that a single exponent was not enough. Schertzer and Lovejoy (1987) showed that it was nevertheless possible to keep H in the role of a mean fluctuation exponent (originally termed a cascade "conservation exponent"). This is the sense of the H exponent discussed here. As described above, using appropriate definitions of fluctuations (i.e., by the use of an appropriate wavelet), H can take on any real value. When the definition is applied to fBm, it yields the standard fBm value $H = H_M$, yet when applied to fGn, it yields $H = H_M - 1$.

Intermittency

Spikes

Often, scaling systems are modeled by linear stochastic processes, leading to linear exponent $\xi(q) = qH$, i.e., $K(q)$ is neglected; the processes are sometimes called "monofractal processes," the most common examples being fBm, fGn. Nonzero $K(q)$ is associated with the physical phenomenon of "spikiness" or *intermittency*: Compare the spatial spiky and temporal nonspiky series in Fig. 1a and their structure functions in Fig. 1b.

Classically, intermittency was first identified in laboratory flows as "spottiness" (Batchelor and Townsend 1949), in the atmosphere by the concentration of most of atmospheric fluxes in tiny, sparse (fractal) regions. In solid Earth geophysics, fields such as concentrations of ore are similarly sparse and scaling, a fact that notably prompted (de Wijs 1951) to independently develop a multiplicative (cascade) model for the distribution of ore. In the 1980s, de Wijs' model was rediscovered in statistical physics as the multifractal " p model" (see also Agterberg 2005).

Early quantitative intermittency definitions were developed originally for fields (space). These are of the "on-off" type: When the temperature, wind, or other field exceeds a threshold, then it is "on," i.e., in a special state – perhaps of

strong/violent activity. At a specific measurement resolution, the on-off intermittency can be defined as the fraction of space that the field is "on" (where it exceeds the threshold). In a scaling system, for any threshold the "on" region will be a fractal set (sparseness characterized by c , eq. 2) and threshold by exponent γ . The resulting function $c(\gamma)$ describes the intermittency over all scales and all intensities (thresholds) and is related to $K(q)$ as described below. In scaling time series, the same intermittency definition applies (note that other definitions are sometimes used in deterministic chaos).

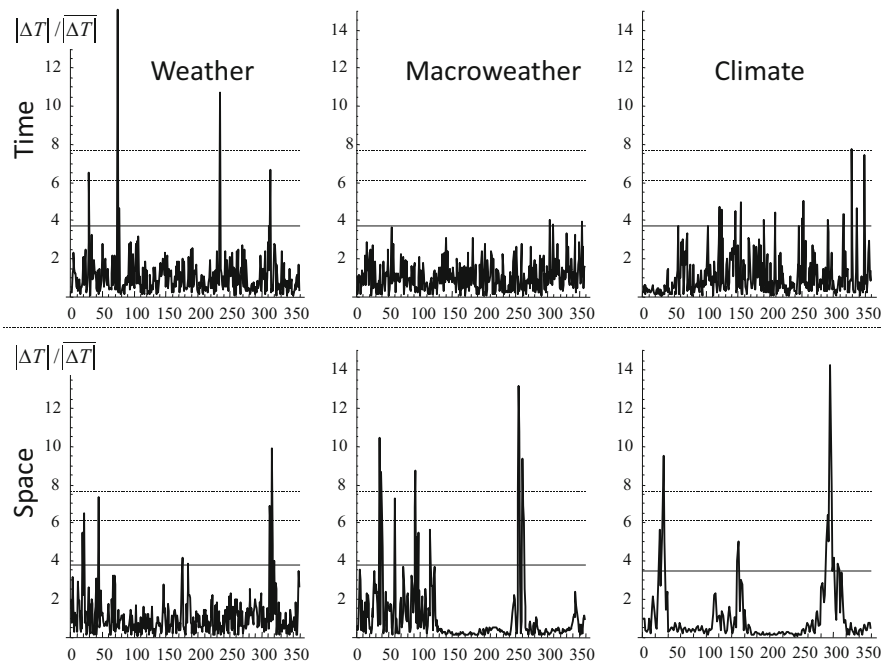
Sometimes, the intermittency is obvious, but sometimes it is hidden and underestimated or overlooked; let us discuss some examples. In order to vividly display the nonclassical intermittency, it suffices to introduce a seemingly trivial tool – a "spike plot." A spike plot is the series (or transect) of the absolute first differences Δf normalized by their means, $\Delta f / \overline{\Delta f}$ (for a series f). Fig. 2 shows examples from the main atmospheric scaling regimes (temperatures, the middle figure corresponds to the macroweather data analyzed in Fig. 1a, 1b). Fig. 3 shows examples from solid earth geophysics. The resolutions have been chosen to be within the corresponding scaling regimes. In Fig. 2, one immediately notices that with a single exception – macroweather in time – all of the regimes are highly "spiky," exceeding the maximum expected for Gaussian processes by a large margin (the solid horizontal line). Indeed, for the five other plots, the maxima corresponds to (Gaussian) probabilities $p < 10^{-9}$ (the top dashed line), and four of the six to $p < 10^{-20}$. The exception – macroweather in time – is the only process that is close to Gaussian behavior, but even macroweather is highly non-Gaussian in space (bottom, middle). In Fig. 3, we show corresponding examples from the KTB borehole of density and susceptibility of rocks. For the susceptibility, the intermittency is so extreme as to be noticeable in the original series (upper right), but less so for the rock density (upper left). In both cases, the spike plots (bottom row) are strongly non-Gaussian with extremes corresponding to Gaussian probabilities of 10^{-15} , 10^{-162} , respectively.

Codimensions and Singularities

In order to understand the spike plots, recall that if a process is scaling, we have eq. 1 where $\varphi_{\Delta t}$ is the (normalized) flux driving the process. The normalized spikes $\Delta T / \overline{\Delta T}$ can thus be interpreted as estimates of the nondimensional, normalized driving fluxes:

$$\Delta T(\Delta t) / \overline{\Delta T(\Delta t)} = \varphi_{\Delta t, un} / \overline{\varphi_{\Delta t, un}} = \varphi_{\Delta t} \quad (13)$$

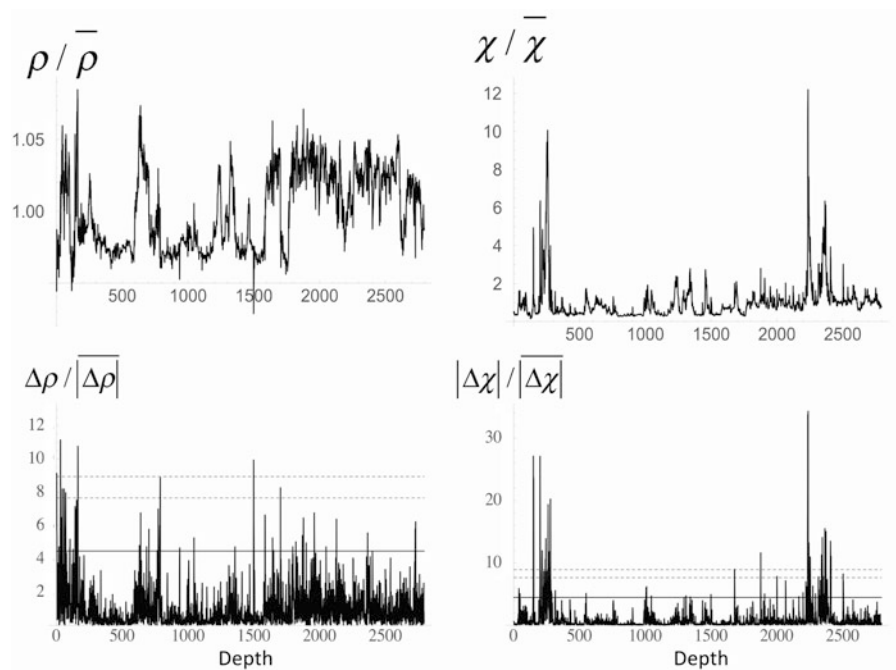
(where $\varphi_{\Delta t, un}$ is the raw, unnormalized flux, the overbar indicates an average over the series). The interpretation in terms of fluxes comes from turbulence theory and is routinely used to quantify turbulence. In the weather regime in



Scaling and Scale Invariance, Fig. 2 Temperature spike plots for weather, macroweather, climate (left to right), and time and space (top and bottom rows). The solid horizontal black line indicates the expected maximum for a Gaussian process with the same number of points (360 for each with the exception of the lower right which had only 180 points); the dashed lines are the corresponding probability levels $p = 10^{-6}$, $p = 10^{-9}$ for Gaussian processes; and two of the spikes exceed

14; $p < 10^{-77}$. The upper left is Montreal at 1 hour resolution; upper middle Montreal at 4 month resolution; upper right, paleotemperatures from Greenland ice cores (GRIP) at 240 year resolution; lower left aircraft at 280 m resolution; lower middle one monthly resolution temperatures at 1o resolution in space; lower right 140 year resolution in time, 2o in space at 45oN. Reproduced from [Lovejoy, 2018]

Scaling and Scale Invariance, Fig. 3 Density (left), magnetic susceptibility (right), nondimensionalized by their means (top), and corresponding spike plots (normalized absolute first differences, bottom). The data are from the first 2795 points of the KTB borehole, each point at a 2 m interval; the horizontal axis indicates the number of points from the surface. In the spike plots, the solid line indicates the maximum expected for a Gaussian process, the dashed lines corresponding to (Gaussian) probability levels of 10^{-9} , 10^{-12} . The extreme spikes correspond to Gaussian probabilities of $\approx 10^{-15}$, 10^{-162} , respectively



respectively time and space, the squares and cubes of the wind spikes are estimates of the turbulent energy fluxes. This spikiness is because most of the dynamically important events are sparse, hierarchically clustered, occurring mostly in storms and the center of storms.

The spikes visually display the intermittent nature of the process. As long as $H < 1$ (true for nearly all geo-processes), the differencing that yields the spikes acts as a high pass filter, and the spikes are dominated by the high frequencies. The Gaussian fGn, fBm processes – or nonscaling processes such as autoregressive and moving average processes (and the hybrid fractional versions of these) – will have spikes that look like the macroweather spikes: In Figs. 2 and 3, they will be roughly bounded by the (Gaussian) solid horizontal lines.

To go beyond Gaussians to the general multifractal scaling case, each spike is considered to be a singularity of order γ :

$$\lambda^\gamma = |\Delta f|/|\overline{\Delta f}| \quad (14)$$

λ is the scale ratio: $\lambda = (\text{the length of the series}) / (\text{the resolution of the series}) = \text{the number of pixels}$; in Fig. 3, $\lambda = 2795$. The most extreme spike ($|\Delta f|/|\overline{\Delta f}|$) therefore has a probability $\approx 1/2795$. For Gaussian processes, spikes with this probability have $|\Delta f|/|\overline{\Delta f}| = 4.47$; this is shown by the solid lines in Fig. 3; the line therefore corresponds to $\gamma = \log(|\Delta f|/|\overline{\Delta f}|) / \log \lambda \approx \log 4.47 / \log 2795 \approx 0.19$. For comparison, the actual maxima from the spike plots are $\gamma_{\max} = \log(34.5)/\log(2795) = 0.45$ and $\log(11.6)/\log(2795) = 0.30$ (susceptibility, density, respectively). These values are close to those predicted by multifractal models for these processes (for more details, see Lovejoy 2018).

To understand the spike probabilities, recall that the statistics of φ_λ were defined above by the moments, and $K(q)$. However, this is equivalent to specifying them via the corresponding (multiscaling) probability distributions:

$$\Pr(\varphi_\lambda > s) \approx \lambda^{-c(\gamma)}; \quad \gamma = \frac{\log s}{\log \lambda} \quad (15)$$

where “ $\approx$ ” indicates equality to within an unimportant prefactor and $c(\gamma)$ is the codimension function that specifies how the probabilities change with scale (Eq. 2 with $\lambda \propto L^{-1}$ and $\Pr \propto \rho$). In scaling systems, the relationship between probabilities and moments is between the corresponding scaling exponents $c(\gamma)$, $K(q)$. It turns out that it is given by the simple Legendre transformation:

$$\begin{aligned} c(\gamma) &= \max_q (q\gamma - K(q)) \\ K(q) &= \max_\gamma (q\gamma - c(\gamma)) \end{aligned} \quad (16)$$

(Parisi and Frisch 1985). The Legendre relations are generally well behaved since K and c are convex, although

discontinuities in the first- and second-order derivatives can occur notably due to the divergence of high-order moments $q > q_D$ (“multifractal phase transitions”). The Legendre transformation also allows us to interpret $c(\gamma)$ as the fractal codimension of the set of points characterized by the singular behavior λ^γ .

Returning to the spike plots, using Eq. 15 and the estimate Eq. 14 for the singularities, we can write the probability distribution of the spikes as:

$$\Pr(|\Delta f|/|\overline{\Delta f}| > s) \approx \lambda^{-c(\gamma)}; \quad \gamma = \frac{\log s}{\log \lambda} \quad (17)$$

$\Pr(|\Delta f|/|\overline{\Delta f}| > s)$ is the probability that a randomly chosen spike $|\Delta f|/|\overline{\Delta f}|$ exceeds a fixed threshold s (it is equal to one minus the more usual cumulative distribution function). $c(\gamma)$ characterizes sparseness because it quantifies how the probabilities of spikes of different amplitudes change with resolution λ (for example, when they are smoothed out by averaging). The larger $c(\gamma)$, the sparser the set of spikes that exceed the threshold $s = \lambda^\gamma$. A series is intermittent whenever it has spikes with $c > 0$. Gaussian series are not intermittent since $c(\gamma) = 0$ for all the spikes.

If there were no constraints on the system beyond scaling, an empirical specification would require the (unmanageable) determination of the entire scale-invariant functions $c(\gamma)$ or $K(q)$: the equivalent of an infinite number of parameters. Fortunately, it turns out that stable, attractive “universal” multifractal processes exist that require only two parameters:

$$K(q) = \frac{C_1}{(\alpha - 1)} (q^\alpha - q) \quad (18)$$

where $0 \leq \alpha \leq 2$ is the Levy index that characterizes the degree of multifractality (Schertzer and Lovejoy 1987). Notice that for any α , the intermittency parameter $C_1 = K'(1)$ and in addition $\alpha = K''(1)/K'(1)$. Together, α and C_1 thus characterize the tangent and curvature of K near the mean ($q = 1$). For the universal multifractal probability exponent, the Legendre transformation of eq. 18 yields:

$$c(\gamma) = C_1 \left(\frac{\gamma}{C_1 \alpha'} + \frac{1}{\alpha} \right)^{\alpha'} \quad (19)$$

where the auxiliary variable α' is defined by: $1/\alpha' + 1/\alpha = 1$. This means that processes, whose moments have exponents $K(q)$ given by eq. 7, have probabilities with exponents $c(\gamma)$ given by eq. 15. Empirically, most geofields have α in the range 1.5–1.9, not far from the “log-normal” multifractal ($\alpha = 2$).

The $c(\gamma)$, $K(q)$ codimension multifractal formalism is appropriate for stochastic multifractals (Schertzer and

Lovejoy 1987). Another commonly used multifractal formalism (often used in solid earth geophysics (Cheng and Agterberg 1996) is the dimension formalism of (Halsey et al. 1986) that was developed for characterizing phase spaces in deterministic chaos with notation $f_d(\alpha_d)$, α_d , $\tau_d(q)$. In a finite dimensional space of dimension d , $f_d(\alpha_d) = d - c(\gamma)$ where $\alpha_d = d - \gamma$, and $\tau_d(q) = (q-1)d - K(q)$ where the subscript “ d ” (the dimension of the space) has been added to emphasize that unlike c , γ , K , in the dimension formalism, the basic quantities depend on both the statistics as well as d . For example, whereas c , γ , and K are the same for 1-D transects, 2-D sections, or 3-D spaces, f_d , α_d , and τ_d are different for each subspace.

Scaling in Two or Higher Dimensional Spaces: Generalized Scale Invariance

Scale Functions and Scale-Changing Operators: From Self-Similarity to Self-Affinity

If we only consider scaling in 1-D (series and transects), the notion of scale itself can be taken simply as an interval (space, Δx) or lag (time, Δt), and large scales are simply obtained from small ones by multiplying by their scale ratio λ . More general geoscience series and transects are only 1-D subspaces of (x, t) space-time geoprocesses. In both the atmosphere and the solid earth, an obvious issue arises due to vertical/horizontal stratification: We do not expect the scaling relationship between small and large to be the same in horizontal and in vertical directions. Unsurprisingly, the corresponding transects generally have different exponents ($\xi(q)$, H , $K(q)$, $c(\gamma)$). In general, the degree of stratification of structures systematically changes with scale. To deal with this fundamental issue, we need an anisotropic definition of the notion of scale itself.

The simplest scaling stratification is called “self-affinity”: The squashing is along orthogonal directions whose directions are the same everywhere in space, for example, along the x and z axes in an x - z space, e.g., a vertical section of the atmosphere or solid earth. More generally, even horizontal sections will not be self-similar: As the scale changes, structures will be both squashed and rotated with scale. A final complication is that the anisotropy can depend not only on scale but also on position. Both cases can be dealt with by using the formalism of Generalized Scale Invariance (GSI; Schertzer and Lovejoy 1985b), corresponding respectively to linear (scale only) and nonlinear GSI (scale and position) (Lovejoy and Schertzer 2013, ch. 7; Lovejoy 2019, ch. 3).

The problem is to define the notion of scale in a system where there is no characteristic size. Often, the simplest (but usually unrealistic) “self-similar” system is simply assumed without question: The notion of scale is taken to be isotropic. In this case, it is sufficient to define the scale of a

vector $\underline{r}$ by the usual vector norm (the length of the vector $\underline{r}$ denoted by $|\underline{r}| = (x^2 + z^2)^{1/2}$. $|\underline{r}|$ satisfies the following elementary scaling rule:

$$|\lambda^{-1}\underline{r}| = \lambda^{-1}|\underline{r}| \quad (20)$$

where again, λ is a scale ratio. When $\lambda > 1$, this equation says that the scale (here, length) of the reduced, shrunken vector $\lambda^{-1}\underline{r}$ is simply reduced by the factor λ^{-1} , a statement that holds for any orientation of $\underline{r}$.

To generalize this, we introduce a more general scale function $\|\underline{r}\|$ as well as a more general scale-changing operator T_λ ; together they satisfy the analogous equation:

$$\|T_\lambda \underline{r}\| = \lambda^{-1} \|\underline{r}\| \quad (21)$$

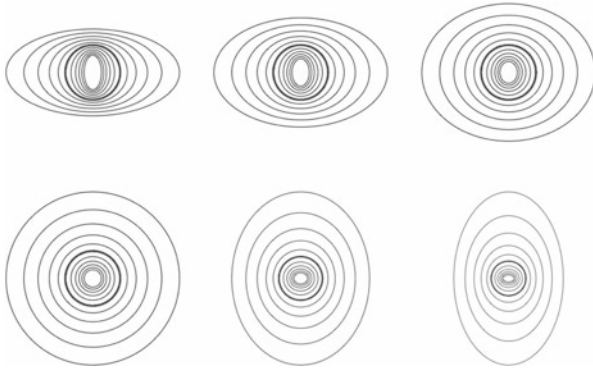
For the system to be scaling, a reduction by scale ratio λ_1 followed by a reduction λ_2 should be equal to first reduction by λ_2 and then by λ_1 , and both should be equivalent to a single reduction by factor $\lambda = \lambda_1 \lambda_2$. The scale-changing operator therefore satisfies group properties, so T_λ is a one-parameter Lie group with generator G :

$$T_\lambda = \lambda^{-G} \quad (22)$$

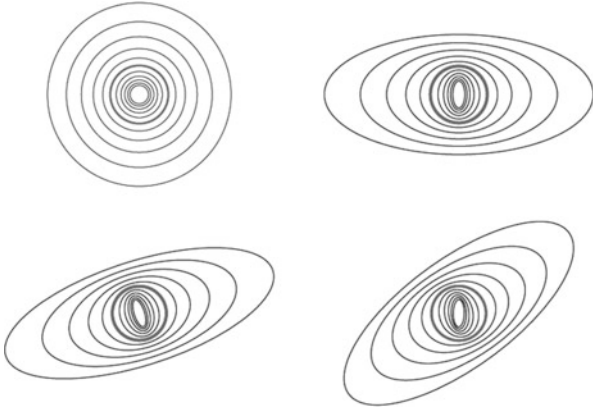
When G is the identity operator (I), then $T_\lambda \underline{r} = \lambda^{-I} \underline{r} = \lambda^{-1} \underline{r}$ so that the scale reduction is the same in all directions (an isotropic reduction): $\|\lambda^{-1} \underline{r}\| = \lambda^{-1} \|\underline{r}\|$. However, a scale function that is symmetric with respect to such isotropic changes is not necessarily equal to the usual norm $|\underline{r}|$ since the vectors with unit scale (i.e., those that satisfy $\|\underline{r}\| = 1$) may be any (nonconstant, hence anisotropic) function of the polar angle – they are not necessarily circles (2D) or spheres (3D). Indeed, in order to complete the scale function definition, we must specify all the vectors whose scale is unity – the “unit ball.” If in addition to $G = I$, the unit scale is a circle (sphere), then the two conditions imply $\|\underline{r}\| = |\underline{r}|$ and we recover Eq. 20. In the more general – but still linear case where G is a linear operator (a matrix) – T_λ depends on scale but is independent of location; more generally – nonlinear GSI – G also depends on location and scale; Figs. 4, 5, 6a, 6b and 7 give some examples of scale functions, and Figs. 8a, 8b, 9 and 10 show some of the corresponding multifractal cloud, magnetization, and topography simulations.

Scaling Stratification in the Earth and Atmosphere

GSI is exploited in modeling and analyzing the earth’s density and magnetic susceptibility fields (the review Lovejoy and Schertzer 2007), and in many atmospheric fields (wind, temperature, humidity, precipitation, cloud density, and aerosol concentrations; see the monograph Lovejoy and Schertzer 2013).



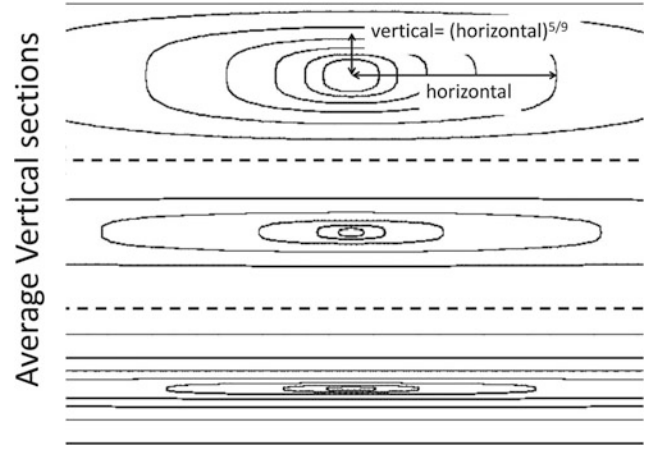
Scaling and Scale Invariance, Fig. 4 A series of ellipses each separated by a factor of 1.26 in scale, red indicating the unit scale (here, a circle, thick lines). Upper left to lower right, H_z increasing from 2/5, 3/5, 4/5 (top), 1, 6/5, 7/5 (bottom, left to right). Note that when $H_z > 1$, the stratification at large scales is in the vertical rather than the horizontal direction (this is required for modeling the earth's geological strata). Reproduced from Lovejoy (2019)



Scaling and Scale Invariance, Fig. 5 Blowups and reductions by factors of 1.26 starting at circles (thick lines). The upper left shows the isotropic case, the upper right shows the self-affine (pure stratification case), the lower left example is stratified but along oblique directions, and the lower right example has structures that rotate continuously with scale while becoming increasingly stratified. The matrices used are: $\begin{pmatrix} 1.35 & 0 \\ 0 & 0.65 \end{pmatrix}$, $\begin{pmatrix} 1.35 & 0 \\ 0 & 0.65 \end{pmatrix}$, $\begin{pmatrix} 1.35 & 0.25 \\ 0.25 & 0.65 \end{pmatrix}$, and $\begin{pmatrix} 1.35 & -0.45 \\ 0.85 & 0.65 \end{pmatrix}$ (upper left to lower right). Reproduced from Lovejoy (2019)

To give the idea, we can define the “canonical” scale function for the simplest stratified system representing a vertical (x, z) section in the atmosphere or solid earth:

$$\begin{aligned} \|(x, z)\| &= l_s \left| \left(\frac{x}{l_s} \right), \text{sign}(z) \left| \frac{z}{l_s} \right| \right|^{1/H_z} \\ &= l_s \left[\left(\frac{x}{l_s} \right)^2 + \left| \frac{z}{l_s} \right|^{2/H_z} \right]^{1/2} \end{aligned} \quad (23)$$



Scaling and Scale Invariance, Fig. 6a The theoretical shapes of average vertical cross sections using the empirical parameters estimated from CloudSat – derived mean parameters: $H_z = 5/9$, with sphero-scales 1 km (top), 100 m (middle), 10 m (bottom), roughly corresponding to the geometric mean and one-standard-deviation fluctuations. In each of the three, the distance from left to right horizontally is 100 km, from top to bottom vertically is 20 km. It uses the canonical scale function. The top figure in particular shows that structures 100 km wide will be about 10 km thick whenever the sphero-scale is somewhat larger than average (Lovejoy et al. 2009)

H_z characterizes the degree of stratification (see below), and l_s is the “sphero-scale,” so-called because it defines the scale at which horizontal and vertical extents of structures are equal (although they are generally not exactly circular):

$$\|(l_s, 0)\| = \|(0, l_s)\| = l_s \quad (24)$$

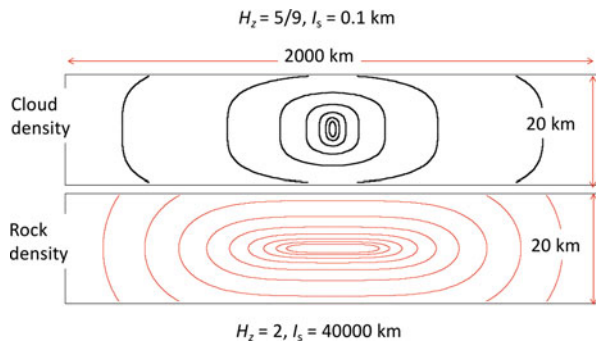
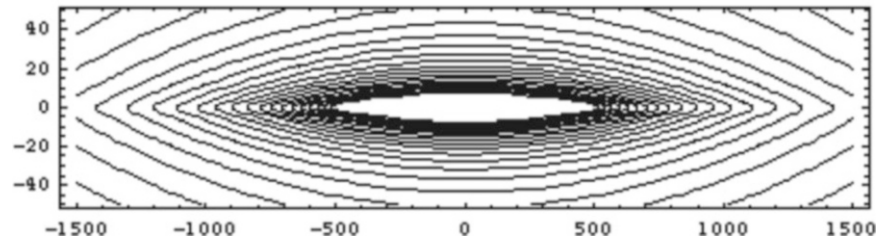
It can be seen by inspection that $\|(x, z)\|$ satisfies:

$$\|T_\lambda(x, z)\| = \lambda^{-1} \|(x, z)\|; \quad T_\lambda = \lambda^{-G}; \quad G = \begin{pmatrix} 1 & 0 \\ 0 & H_z \end{pmatrix} \quad (25)$$

(note that matrix exponentiation is simple only for diagonal matrices – here $T_\lambda = \begin{pmatrix} \lambda^{-1} & 0 \\ 0 & \lambda^{-H_z} \end{pmatrix}$ – but when G is not diagonal, it can be calculated by expanding the series: $\lambda^{-G} = e^{-G \log \lambda} = 1 - G \log \lambda + (G \log \lambda)^2/2 - \dots$). Notice that in this case, the ratios of the horizontal/vertical statistical exponents (i.e., $\xi(q)$, H , $K(q)$, and $c(\gamma)$) are equal to H_z . We could also note that linear transects taken any direction other than horizontal or vertical will have two scaling regimes (with a break near the sphero-scale). However, the break is spurious; it is a consequence of using the wrong notion of scale.

Equipped with a scale function, the general anisotropic generalization of the 1-D scaling law (Eq. 1) may now be expressed by using the scale $\|\Delta r\|$:

Scaling and Scale Invariance, Fig. 6b Vertical cross-section of the magnetization scale function assuming $H_z = 2$ and a sphero-scale of 40,000 km. The scale is in kilometers and the aspect ratio is $\frac{1}{4}$ (Lovejoy et al. 2001)



Scaling and Scale Invariance, Fig. 7 The unity of geoscience illustrated via a comparison of typical (average) vertical sections in the atmosphere (top), and in the solid earth (bottom), an aspect ratio of 1/5 was used. The key points are as follows: $H_z < 1$, l_s small (atmosphere, stratification increasing at larger scales) and $H_z > 1$, l_s large (solid earth, stratification increasing at smaller scales). The sphero-scale (l_s) varies in space and in time (see Fig. 6a, 6b)

$$\Delta f(\Delta r) \stackrel{d}{=} \varphi_{\|\Delta r\|} \|\Delta r\|^H \quad (26)$$

This shows that the full scaling model or full characterization of scaling requires the specification of the notion of scale via the scale-invariant generator G and unit ball (hence the scale function), the fluctuation exponent H , as well the statistics of $\varphi_{\|\Delta r\|}$ specified via $K(q)$, $c(\gamma)$ or – for universal multifractals, C_1 , α . In many practical cases – e.g., vertical stratification – the direction of the anisotropy is fairly obvious, but in horizontal sections, where there can easily be significant rotation of structures with scale, the empirical determination of G and the scale function is a difficult, generally unsolved problem.

In the atmosphere, it appears that the dynamics are dominated by Kolmogorov scaling in the horizontal ($H_h = 1/3$) and Bolgiano-Obukhov scaling in the vertical ($H_v = 3/5$) so that $H_z = H_h/H_v = 5/9 = 0.555\dots$ Assuming that the horizontal directions have the same scaling, then typical structures of size $L \times L$ in the horizontal have vertical extents of L^{H_z} , hence their volumes are $L^{D_{el}}$ with “elliptical dimension” $D_{el} = 2 + H_z = 2.555\dots$; the “23/9D model” (Schertzer and Lovejoy 1985a). This model is very close to the empirical data, and it contradicts the “standard model” that is based on isotropic symmetries and that attempts to combine a small-

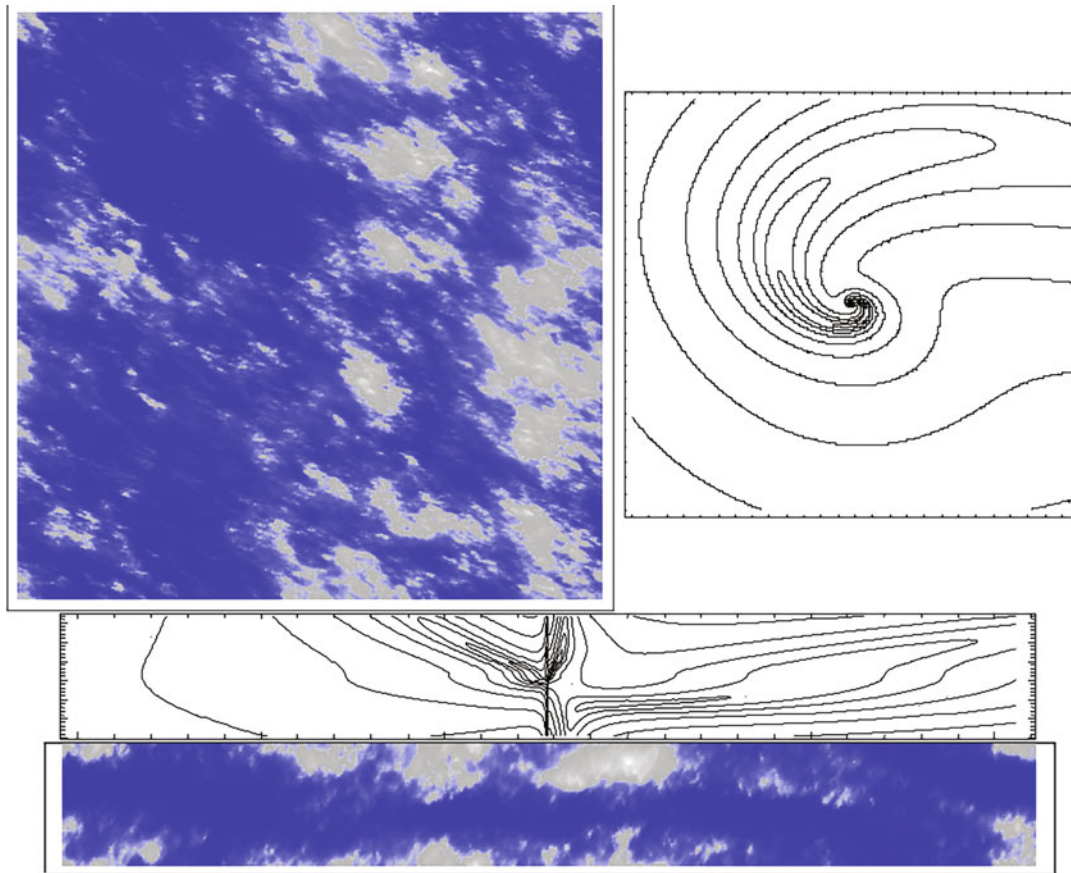
scale isotropic 3D regime and a large-scale isotropic (flat) 2D regime with a transition supposedly near the atmospheric scale height of 10 km. The requisite transition has never been observed, and claims of large-scale 2D “geostrophic” turbulence have been shown to be spurious (reviewed in ch. 2 of Lovejoy and Schertzer 2013).

Since $H_z < 1$, the atmosphere becomes increasingly stratified at large scales. In solid earth applications, it was found that $H_z \approx 2-3$ (rock density, susceptibility, and hydraulic conductivity (Lovejoy and Schertzer 2007)) i.e., $H_z > 1$ implying on the contrary that the earth’s strata are typically more horizontally stratified at the smallest scales (Figs. 6a, 6b and 7).

Conclusions

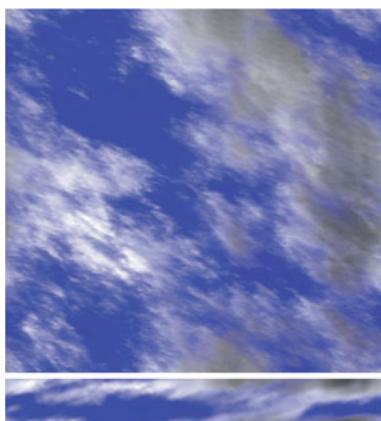
Geosystems typically involve structures spanning wide ranges of scale: from the size of the planet down to millimetric dissipations scales (the atmosphere) or micrometric scales (solid earth). Classical approaches stem from the outdated belief that complex behavior requires complicated models. The result is a complicated “scalebound” paradigm involving hierarchies of phenomenological models / mechanisms each spanning small ranges of scale. For example, conventionally, the atmosphere was already divided into synoptic, meso- and microscales when Orlanski (1975) proposed further divisions implying new mechanisms every factor of two or so in scale. Ironically, this still popular paradigm arose at the same time that scale-free numerical weather and climate models were being developed that were based on the scaling dynamical equations and now known to display scaling meteorological fields.

At the same time as the scale-bound paradigm was ossifying, the “Fractal Geometry of Nature” (Mandelbrot 1977, 1982) proposed an opposite approach based on sets that respected an isotropic scaling symmetry: self-similar fractals. These have scale-free power law number-size relationships whose exponents are scale-invariant geo-applications already included topography and clouds. However, geosystems are rarely geometric sets of points but rather fields (i.e., with values, e.g., temperature, rock density) varying in space and

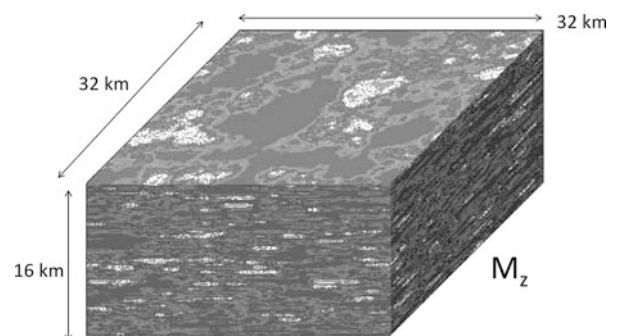


Scaling and Scale Invariance, Fig. 8a Simulations of the liquid water density field. The top is a horizontal section, to the right the corresponding central horizontal cross section of the scale function. The stratification is determined by $H_z = 0.555$ with $l_s = 32$. The bottom

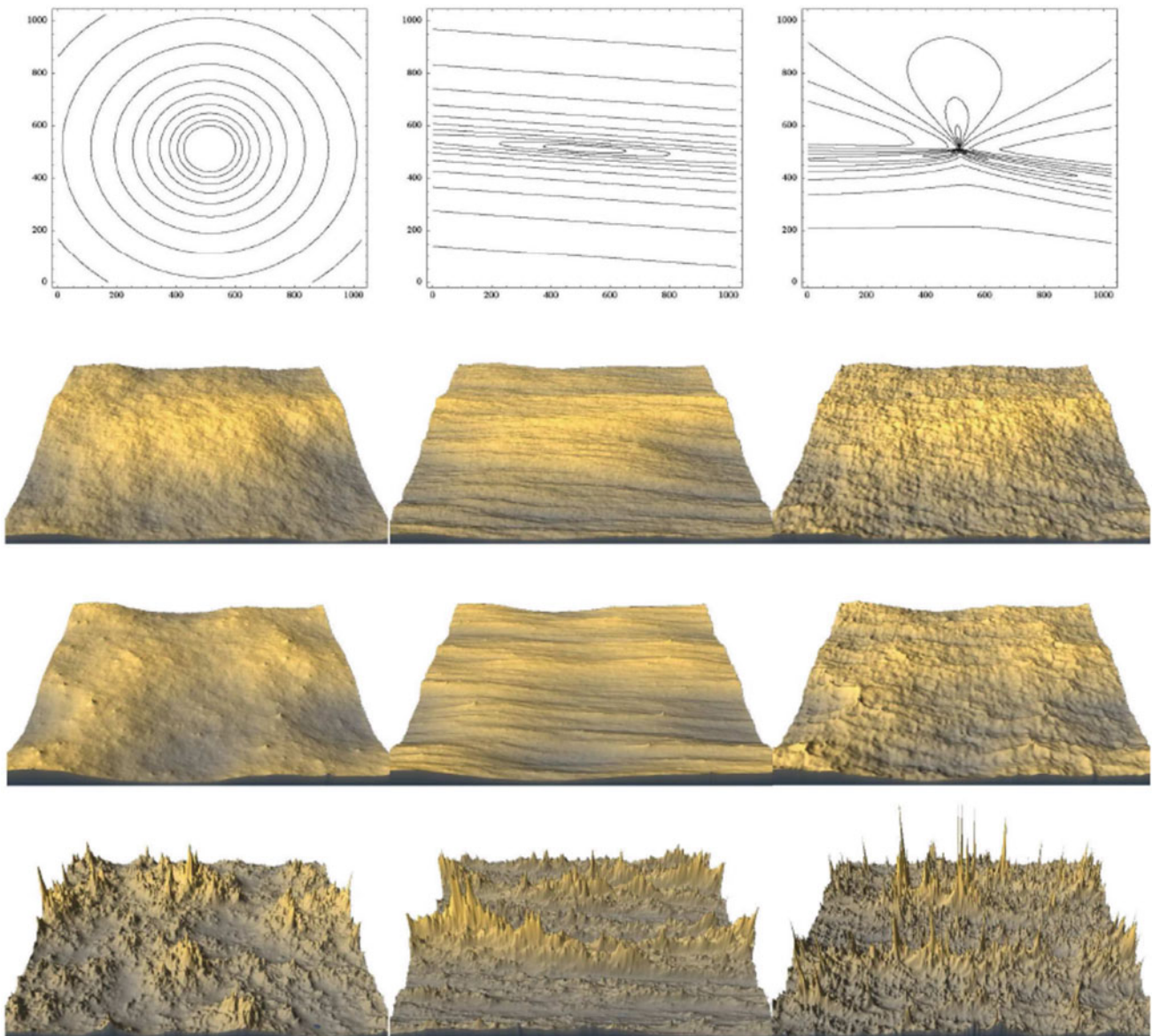
shows side views. There is also horizontal anisotropy with rotation. Statistical (multifractal) parameters: $\alpha = 1.8$, $C_1 = 0.1$, and $H = 0.333$, on a $512 \times 512 \times 64$ grid (Lovejoy and Schertzer 2013)



Scaling and Scale Invariance, Fig. 8b The top is the visible radiation field (corresponding to Fig. 8a), looking up (sun at 45° from the right); the bottom is a side radiation field (one of the 512×64 pixel sides) (Lovejoy and Schertzer 2013)



Scaling and Scale Invariance, Fig. 9 Numerical simulation of a multifractal rock magnetization field (vertical, component M_z) with parameters deduced from rock samples and near surface magnetic anomalies (Pecknold et al. 2001). The simulation was horizontally isotropic with vertical stratification specified by $H_z = 1.7$, $l_s = 2500$ km; the region is $32 \times 32 \times 16$ km, resolution 0.25 km. The statistical parameters specifying the horizontal multifractal statistics are $H = 0.2$, $\alpha = 1.98$, and $C_1 = 0.08$. This is a reasonably realistic crustal section, although the spheroscale was taken to be a bit too small in order that strata may be easily visible. Notice that the structures become more stratified at smaller scales. The direction of M is assumed to be fixed in the z direction



Scaling and Scale Invariance, Fig. 10 Comparison of isotropic versus anisotropic simulations for three different scaling models. Top row shows the scale functions. From left to right, we change the type of anisotropy: The left column is self-similar (isotropic) while the middle and right columns are anisotropic and symmetric with respect to $G = \begin{pmatrix} 0.8 & -0.5 \\ 0.05 & 1.2 \end{pmatrix}$. The middle column has unit ball circular at 1 pixel, while for the right one the unit ball is also anisotropic. The second, third, and fourth rows show the corresponding fBm (with $H = 0.7$), the analogous fractional Levy motion (fLm $\alpha = 1.8$, $H = 0.7$), and multifractal ($\alpha = 1.8$, $C1 = 0.12$, $H = 0.7$) simulations. We note that in the case of fBm, one mainly

perceives textures; there are no very extreme mountains or other morphologies evident. One can see that the fLm is too extreme; the shape of the singularity (particularly visible in the far right) is quite visible in the highest mountain shapes. The multifractal simulations are more realistic in that there is a more subtle hierarchy of mountains. When the contour lines of the scale functions are close, the scale $\|x\|$ changes rapidly over short distances. For a given order of singularity γ , λ^γ will therefore be larger. This explains the strong variability depending on direction (middle bottom row) and on shape of unit ball (right bottom row). Indeed, spectral exponents will be different along the different eigenvectors of G (Lovejoy and Schertzer 2007).

in time. In addition, these fields are typically highly anisotropic notably with highly stratified vertical sections. In the 1980s, the necessary generalizations to scaling fields (multifractals) and to anisotropic scaling (Generalized Scale Invariance, GSI) were achieved.

GSI clarifies the significance of scaling in geoscience since it shows that scaling is a rather general symmetry principle: It is thus the simplest relation between scales. Just as the classical symmetries (temporal, spatial invariance, and directional invariance) are equivalent (Noether's theorem) to conservation laws (energy, momentum, and angular momentum), the

(nonclassical) scaling symmetry conserves the scaling functions G , K , and c . Symmetries are fundamental since they embody the simplest possible assumption, model: Under a system change, there is an invariant. In physics, initial assumptions about a system are that it respects symmetries. Symmetry breaking is only introduced on the basis of strong evidence or theoretical justification: in the case of scale symmetries, they are broken by the introduction of characteristic space or time scales.

Theoretically, scaling is a unifying geophysical principle that has already shown its realism in numerous geoapplications both on Earth and Mars proving the pertinence of anisotropic scaling models and analyses. It is unfortunate that in geoscience, scale-bound models and analyses continue to be justified on superficial phenomenological grounds. Geoapplications of scaling are therefore still in their infancy.

References

- Agterberg F (2005) New applications of the model of de Wijs in regional geochemistry. *Math Geol.* <https://doi.org/10.1007/s11004-006-96063-7>
- Batchelor GK, Townsend AA (1949) The nature of turbulent motion at large wavenumbers. *Proc R Soc Lond A* 199:238
- Chen W, Lovejoy S, Muller JP (2016) Mars' atmosphere: the sister planet, our statistical twin. *J Geophys Res Atmos* 121. <https://doi.org/10.1002/2016JD025211>
- Cheng Q, Agterberg F (1996) Multifractal modelling and spatial statistics. *Math Geol* 28:1–16
- de Wijs HJ (1951) Statistics of ore distribution, part I. *Geol Mijnb* 13: 365–375
- Feller W (1971) An introduction to probability theory and its applications, vol 2. Wiley, p 669
- Grassberger P, Procaccia I (1983) Measuring the strangeness of strange attractors. *Physica* 9D:189–208
- Halsey TC, Jensen MH, Kadanoff LP, Procaccia I, Shraiman B (1986) Fractal measures and their singularities: the characterization of strange sets. *Phys Rev A* 33:1141–1151
- Hentschel HGE, Procaccia I (1983) The infinite number of generalized dimensions of fractals and strange attractors. *Physica D* 8:435–444
- Hooge C, Lovejoy S, Pecknold S, Malouin JF, Schertzer D (1994) Universal multifractals in seismicity. *Fractals* 2:445–449
- Hurst HE (1951) Long-term storage capacity of reservoirs. *Trans Am Soc Civ Eng* 116:770–808
- Kolmogorov AN (1940) Wiener's spirals and some other interesting curves in Hilbert's space. *Doklady Akademii Nauk SSSR* 26:115–118
- Lamperti J (1962) Semi-stable stochastic processes. *Trans Am Math Soc* 104:62–78
- Landais F, Schmidt F, Lovejoy S (2019) Topography of (exo)planets. *MNRAS* 484:787–793. <https://doi.org/10.1093/mnras/sty3253>
- Lovejoy S (2018) The spectra, intermittency and extremes of weather, macroweather and climate. *Nat Sci Rep* 8:1–13. <https://doi.org/10.1038/s41598-018-30829-4>
- Lovejoy S (2019) *Weather, macroweather and climate: our random yet predictable atmosphere*. Oxford University Press, p 334
- Lovejoy S, Schertzer D (2007) Scaling and multifractal fields in the solid earth and topography. *Nonlin Processes Geophys* 14:1–38
- Lovejoy S, Schertzer D (2012) Haar wavelets, fluctuations and structure functions: convenient choices for geophysics. *Nonlinear Proc Geophys* 19:1–14. <https://doi.org/10.5194/npg-19-1-2012>
- Lovejoy S, Schertzer D (2013) *The weather and climate: emergent Laws and Multifractal cascades*, 496 pp. Cambridge University Press
- Lovejoy S, Schertzer D, Ladoy P (1986) Fractal characterisation of inhomogeneous measuring networks. *Nature* 319:43–44
- Lovejoy S, Pecknold S, Schertzer D (2001) Stratified multifractal magnetization and surface geomagnetic fields, part 1: spectral analysis and modelling. *Geophys J Inter* 145:112–126
- Lovejoy S, Piel J, Schertzer D (2012) The global space-time Cascade structure of precipitation: satellites, gridded gauges and reanalyses. *Adv Water Resour* 45:37–50. <https://doi.org/10.1016/j.advwatres.2012.03.024>
- Lovejoy S, Tuck AF, Schertzer D, Hovde SJ (2009) Reinterpreting aircraft measurements in anisotropic scaling turbulence. *Atmos Chem Phys* 9: 1–19.
- Mandelbrot BB (1977) *Fractals, form, chance and dimension*. Freeman
- Mandelbrot BB (1982) *The fractal geometry of nature*. Freeman
- Mandelbrot BB, Van Ness JW (1968) Fractional Brownian motions, fractional noises and applications. *SIAM Rev* 10:422–450
- Marsan D, Schertzer D, Lovejoy S (1996) Causal space-time multifractal processes: predictability and forecasting of rain fields. *J Geophys Res* 31D:26333–326346
- Orlanski I (1975) A rational subdivision of scales for atmospheric processes. *Bull Amer Met Soc* 56:527–530
- Parisi G, Frisch U (1985) A multifractal model of intermittency. In: Ghil M, Benzi R, Parisi G (eds) *Turbulence and predictability in geophysical fluid dynamics and climate dynamics*. North Holland, pp 84–88
- Pecknold S, Lovejoy S, Schertzer D (2001) Stratified multifractal magnetization and surface geomagnetic fields, part 2: multifractal analysis and simulation. *Geophys Inter J* 145:127–144
- Schertzer D, Lovejoy S (1983) On the dimension of atmospheric motions, paper presented at IUTAM Symp. In: *On turbulence and chaotic phenomena in fluids*, Kyoto, Japan
- Schertzer D, Lovejoy S (1985a) The dimension and intermittency of atmospheric dynamics. In: Bradbury LJS, Durst F, Launder BE, Schmidt FW, Whitelaw JH (eds) *Turbulent shear flow*. Springer-Verlag, pp 7–33
- Schertzer D, Lovejoy S (1985b) Generalised scale invariance in turbulent phenomena. *Physico-Chemical Hydrodynamics Journal* 6: 623–635
- Schertzer D, Lovejoy S (1987) Physical modeling and analysis of rain and clouds by anisotropic scaling of multiplicative processes. *J Geophys Res* 92:9693–9714
- Taleb NN (2010) *The Black Swan: the impact of the highly improbable*. Random House, p 437pp

Scattergram

Richard J. Howarth
Department of Earth Sciences, University College London
(UCL), London, UK

Synonyms

Crossplot; Dotplot; Pointplot; Scattergraph; Scatterplot; xy-plot

Definition

Scattergram is a contraction of *scatter diagram*, a term which also occurs in the literature. It is a bivariate graph in which pairs of values of two variates (x , y) are plotted as points, using coordinates based on their numerical values on two orthogonal axes in which the x -axis is conventionally horizontal and the y -axis vertical. If one variate is known to be dependent in some way on the other, the y -axis is assigned to the dependent variate. The axes are scaled with regard to a continuous range of the possible, or observed, values of a variate, and each may use either arithmetic or logarithmic scaling. Figure 1 is an early example from the geological literature.

The synonyms *crossplot*, *dotplot*, *pointplot*, *scattergraph*, and *scatterplot* are sometimes written as two separate words which may, or may not, be hyphenated, but such forms are much less frequent; *xy-plot* is also occasionally used. Since the 1990s, *scatterplot* has become established as the most frequently used term (e.g., in Reimann et al. (2008), which, despite its general title, deals largely with regional geochemical data). The terms *scattergram*, *scatter plot*, and *scattergraph* were all in use by the early 1920s, *crossplot* was in use by 1940, and *dotplot* was introduced in the 1960s. Apart from “*xy-plot*,” hyphenation is uncommon today.

Data on the annual frequency of usage of these terms were retrieved, ignoring case, from the Google Books 2012 English corpus, which comprises all the words extracted by optical character recognition from the scanned text of some eight million “books” (which include bound journal volumes) using the *ngramr* package of Carmody (2015). Words occurring less than 40 times in the corpus are not recorded in the publicly available database (Michel et al. 2011), so in some

cases it may be impossible to find when the actual “first” use of a given word took place.

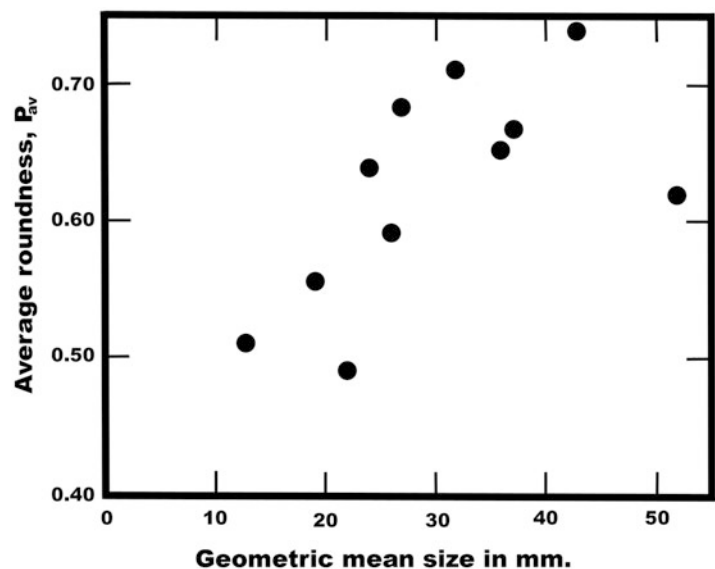
The raw frequency counts for each term were normalized relative to the average annual frequency of occurrence of a set of “neutral” English “words with little or no specific meaning,” *the, of, and, in, a, is, was, not, and other*, and are shown in Fig. 2 as the number of occurrences per million words, as advocated by Younes and Reips (2019). The time trends were obtained by smoothing, using robust locally weighted regression with a 7-year moving window. On account of the large variation in the frequency of usage of different terms, the time trends are shown as a semilogarithmic chart, in which an exponential increase in usage with time becomes linear (the term *scatterplot* exhibits this type of behavior).

Early examples of usage in the Earth sciences include *scatter diagram* by Krumbein and Pettijohn (1938), *scattergram* by Burma (1948), and *scatterplot* by Mason and Folk (1958). Plots in which the x -axis corresponds to time or to an ordered list of abbreviations for chemical element names, etc., are usually referred to by other titles (such as a time-series plot, spider plot, and so on). Scattergrams have long been used in igneous and sedimentary petrology to show compositional trends or to assign samples to a class using one of many existing classification schemes (e.g., the $(\text{Na}_2\text{O} + \text{K}_2\text{O})$ vs. SiO_2 “total alkali silica” or “TAS diagram,” for volcanic and plutonic rocks of Cox et al. (1979); see also Janoušek et al. (2016)), for plotting isotopic ratios (e.g., $^{87}\text{Sr}/^{86}\text{Sr}$ vs. $^{206}\text{Pb}/^{204}\text{Pb}$), and so on.

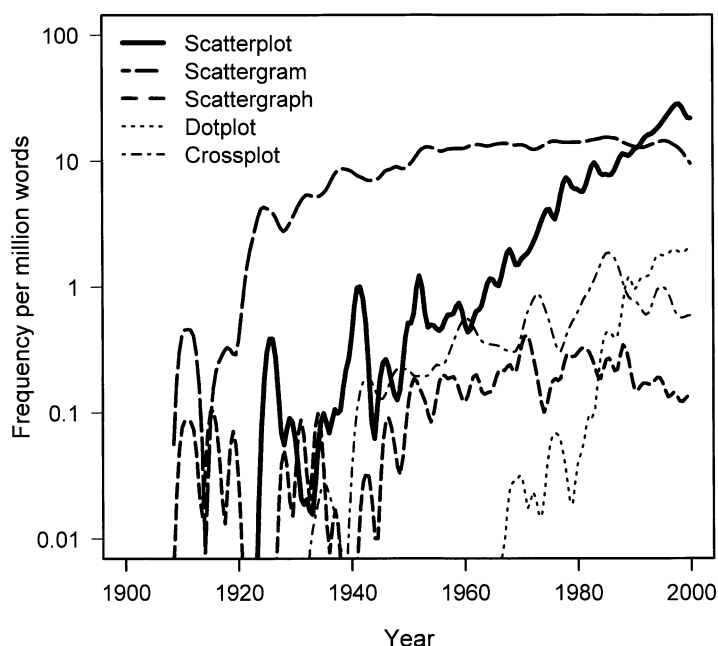
In some cases, the values of an associated categorical attribute, such as geographical location or class affinity, or the value of a third continuous variate, may be shown by means of discrete symbols of different shapes, proportional symbol sizes, or colored symbols of the same or different shapes. If there is too large a number of datapoints of the same type to distinguish them individually, then the value of a third

Scattergram,

Fig. 1 Scattergram, redrawn after Krumbein and Pettijohn (1938, Fig. 89), showing the relationship between average roundness and geometric mean size of a sample of beach pebbles



Scattergram, Fig. 2 Frequency of usage of the term *scattergram* and its most common synonyms (1900–2000) in the Google Books 2012 English corpus



variate may be shown using isolines. Spatial point density within a plot may be shown in a similar way.

Correlation between the x - and y -variates in a scattergram is visually assessed by the closeness of fit of the datapoints to a line, or curve. If a formal fit by a function is required, then it should be accomplished using an appropriate linear or non-linear regression algorithm. In doing so, it should be noted whether measurement errors (or similar) are associated with x , y , or both, and an appropriate regression method applied. It should also be borne in mind that correlation is inherent in the so-called “closed” percentaged and other constant-sum data. Specialized techniques for dealing with such “compositional data” exist (e.g., Buccianti et al. 2006; Tsagris et al. 2020).

Authors, referees, and editors should all bear in mind that because of the expense of color-printed figures in journals (often charged to the author), some illustrations may appear in the printed, as opposed to online, version of an article using monochrome tones of gray. Care should therefore be taken to ensure that the essential information to be conveyed by an illustration, which may appear evident enough when viewed in color on a computer screen or in a color print, can also be distinguished in a monochrome version. Some “temperature” color scales used to depict the value of a third variate often use pastel colors which can be difficult to distinguish individually. A check should also be made for use of colors which those who are color-blind cannot distinguish (e.g., using <https://www.color-blindness.com/coblis-color-blindness-simulator/>). Ensuring clear distinction may require submission of two differently drawn versions of the same illustration for publication.

Scatterplots may also be used as visual goodness-of-fit tests when the observed frequency distribution for a set of data is fitted by a theoretical model, e.g., in quantile-quantile plots, percentile-percentile plots, cumulative probability

plots, and also to find possible outliers using chi-square plots (see Reimann et al. 2008 for discussion).

Summary

A scattergram has long been one of the most commonly used graphical displays in the geological literature, providing a useful visual synopsis of the relationship between two (or three) variates which the eye can easily assimilate. However, some thought is required to ensure that its design is such that its message can be most effectively conveyed in all circumstances.

Cross-References

- [Compositional Data](#)
- [Correlation Coefficient](#)
- [Data Visualization](#)
- [Exploratory Data Analysis](#)
- [Krumbein, William Christian](#)
- [Locally Weighted Scatterplot Smoother](#)
- [Ordinary Least Squares](#)
- [Reduced Major Axis Regression](#)
- [Regression](#)
- [Total Alkali-Silica Diagram](#)

Bibliography

- Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V (eds) (2006) Compositional data analysis in the geosciences: from theory to practice.

- Geological society special publication, vol 264. The Geological Society, London
- Burma BH (1948) Studies in quantitative paleontology: I. some aspects of the theory and practice of quantitative invertebrate paleontology. *J Paleontol* 22:725–761
- Carmody S (2015) Package ‘ngramr.’ <https://github.com/seancarmody/ngramr>. Accessed 9 June 2020
- Cox KG, Bell JD, Pankhurst RJ (1979) The interpretation of igneous rocks. Allen & Unwin, London
- Janoušek V, Moya J-F, Martin H, Erban V, Farrow CM (2016) Geochemical modelling of igneous processes – principles and recipes in R language. Bringing the power of R to a geochemical community. Springer, Heidelberg
- Krumbein WC, Pettijohn FJ (1938) Manual of sedimentary petrography. Appleton-Century, New York
- Mason CC, Folk RL (1958) Differentiation of beach, dune and aeolian flat environments by size analysis, Mustang Island, Texas. *J Sediment Petrol* 28:211–226
- Michel J-B, Shen YK, Aiden AP, Veres A, Gray MK, The Google Books Team, Pickett JP, Hoiberg D, Clancy D, Norvig P, Orwant J, Pinker S, Nowak MA, Aiden EL (2011) Quantitative analysis of culture using millions of digitized books. *Science* 331:176–182
- Reimann C, Filzmoser P, Garrett RG, Dutter R (2008) Statistical data analysis explained. Applied environmental statistics with R. Wiley, Chichester
- Tsagris M, Athineou G, Alenazi A (2020) Package ‘Compositional.’ <https://cran.r-project.org/web/packages/Compositional/Compositional.pdf>. Accessed 14 June 2020
- Younes N, Reips U-D (2019) Guideline for improving the reliability of Google Ngram studies: evidence from religious terms. *PLoS One* 14(3):e0213554. <https://doi.org/10.1371/journal.pone.0213554>. Accessed 4 Apr 2019

Schuenemeyer, John H.

Donald Gautier

DonGautier L.L.C., Palo Alto, CA, USA



Fig. 1 Jack Schuenemeyer (Courtesy of J. Schuenemeyer)

Biography

John H. (Jack) Schuenemeyer is an American mathematical statistician known for his work in mass balance stochastic analysis, discovery process modeling, probabilistic resource assessment, uncertainty evaluation, and statistical education. He is emeritus professor of mathematical sciences at the University of Delaware, an elected member of the International Statistics Institute, Fellow of the American Statistical Association, Distinguished Lecturer for the International Association of Mathematical Geosciences, and recipient of the John Cedric Griffiths Award for excellence in teaching mathematical geology.

Schuenemeyer was born in St. Louis, Missouri, on October 19, 1938, to descendants of German immigrants. He grew up in the St. Louis area with a love of baseball and mathematics. Two years at Washington University showed that an engineering career was not for him, so he joined the US Air Force. While stationed at Lowry Air Force Base, near Denver, Colorado, he met his future wife, Judy, with whom he would ultimately have three children.

Upon return to civilian life, he was hired by RCA in New Jersey as a computer programmer. He then returned to Denver where he worked for the US Bureau of Mines and enrolled in the University of Colorado, earning B.S. and M.S. degrees in Applied Mathematics. In 1972, he was accepted to the University of Georgia, where Prof. Ralph Bargmann supervised his doctoral work in multivariate statistics; he received his Ph. D. in 1975.

The following year, Dr. Schuenemeyer joined the University of Delaware faculty, initially as an assistant professor and then as full professor of mathematics with joint appointments in Geology and Geography. At Delaware, he taught graduate and undergraduate courses in statistics and supervised numerous graduate students, while simultaneously conducting research in applied statistics. An important part of his research was a fruitful collaboration with the US Geological Survey, for whom he developed quantitative techniques for resource evaluation and provided methodological leadership in spatial analysis, nonparametric regression analysis, discovery process modeling, and dependency evaluation. In 1998, he retired from the university to work full time at the USGS as a research statistician, before moving to Cortez, Colorado, to found Southwest Statistical Consulting L.L.C.

Schuenemeyer is a devoted advocate of high-quality education for all members of society, including the financially disadvantaged. For many years, he has volunteered his expertise as a member and president of his local school board in Cortez where he lives with his wife, Judy, a retired attorney.

Schwarzacher, Walther

Jennifer McKinley

Geography, School of Natural and Built Environment,
Queen's University, Belfast, UK



Fig.1 Professor Walther Schwarzacher. (Taken by Jennifer McKinley, 23 April, 2015)

Biography

Professor Walther Schwarzacher was born in 1925 in Graz, Austria. He completed both his undergraduate studies and his doctoral dissertation in a total of 4 years at Innsbruck University, Austria, under the supervision of the distinguished geologist Bruno Sander. Walther was awarded a British Council Scholarship to Cambridge University, where he became a member of the University Natural Science Club and participated in an expedition to Spitsbergen. In recognition of Walther's respect for his academic supervisor, Bruno Sander, he was proud to name a glacier in Spitsbergen after him. Walther made many lifelong friends during his professional life. His advisor at Cambridge, Percival Allen, suggested that he should consider an open position in the Geology Department of Queen's University Belfast, UK. Walther joined Queen's University Belfast as an assistant lecturer in 1949. He was promoted to lecturer, reader, and eventually professor when he was appointed to a personal chair in mathematical geosciences in 1977. In 1967/1968, Walther was distinguished visiting lecturer in the newly formed mathematical geology section at the Geological Survey of Kansas. During this year, he applied statistical time series analysis to local Pennsylvanian limestone-shale sequences and performed experiments simulating the Kansas cyclothems. Walther also spent sabbaticals at the Christian Albrechts University Kiel.

His former head of the school at Queen's University Belfast, Professor Julian Orford, recounts "Walther had a strong belief that geology had a story that could be explained by maths – in particular cycles and rhythms in sediment deposition and planetary motions that such cycles might be associated with." After 65 years including 25 years as emeritus professor, Walther retired but continued to inspire many through his pioneering research on Markov processes. Walther was one of the first researchers to find evidence for Milankovitch cycles, periodic variations in the Earth's orbit, and in the thickness of sedimentary facies. He wrote two influential books in the field of sedimentology and many international publications and chapters. The first book *Sedimentation Models and Quantitative Stratigraphy* (dedicated to A.B. Vistelius and W.C. Krumbein) was published by Elsevier in 1975. In it he proposed stochastic models for sedimentary processes introducing applications of Markov chains and the use of semi-Markov processes. The second book *Cyclostratigraphy and the Milankovitch Theory* which was published in 1993 contains his original evidence for Milankovitch cycles in the thicknesses of coupled limestone and marl beds.

Professor Walther Schwarzacher was a founding member of the International Association for Mathematical Geosciences (IAMG). From 1980 to 1986, Walther was a lecturer in the "Flying Circus" offering short courses in nine countries on four continents under the auspices of the Quantitative Stratigraphy Project of the UNESCO-sponsored International Geological Correlation Program. This included field trips in Brazil, India, and China, accompanied by his wife June. He was extremely proud to have published in the first ever edition of *Mathematical Geology* (now *Mathematical Geosciences*) the first and flagship journal of the Association. The IAMG recognized his scientific contributions through the award in 1977 of the William Christian Krumbein Medal, the second Krumbein Medal to be awarded. This is the highest award given by the Association in recognition of outstanding career achievement and distinction in the application of mathematics or informatics in the Earth sciences, service to the IAMG, and support to professions involved in the Earth sciences. This fully reflects Walther's dedication to advancing mathematical geoscience in quantitative sedimentology and stratigraphy.

I knew Walther for many years, as one of my lecturers in the former Geology Department at Queen's University Belfast and later as a colleague in Queen's and the IAMG. I was pleased that his role in establishing mathematical geoscience and as a founding member of the IAMG was acknowledged through recognition as an honorary IAMG member. Walther's research continues to inspire researchers to use a quantitative cyclo-stratigraphic approach.

Acknowledgments I would like to thank Walther's wife, June, family, and friends for the information and inspirational stories for this biography.

Bibliography

- Schwarzacher W (1942) Zur Morphologie des Wallersees. Arch Hydrobiol 42:372–376
- Schwarzacher W (1966) Sedimentation in subsiding basins. Nature 210(5043):1349
- Schwarzacher W (1969) The use of Markov chains in the study of sedimentary cycles. J Int Assoc Math Geol 1(1):17–39
- Schwarzacher W (1975) Sedimentation models and quantitative stratigraphy. Developments in Sedimentology, vol 19. 382pp, Elsevier Scientific Publishing, Amsterdam, Oxford, New York. ISBN 0 444 41302 2
- Schwarzacher W (1993) Cyclostratigraphy and the Milankovitch theory, vol 52. ELSEVIER Amsterdam, London, New York, Tokyo
- Schwarzacher W (2007) Sedimentary cycles and stratigraphy. Stratigraphy 4(1):77–80

Sedimentation and Diagenesis

Nana Kamiya¹ and Weiren Lin²

¹Laboratory of Earth System Science, School of Science and Engineering, Doshisha University, Kyotanabe, Japan

²Earth and Resource System Laboratory, Graduate School of Engineering, Kyoto University, Kyoto, Japan

Synonyms

Sedimentation: Deposition, Accumulation.

Diagenesis: Lithification.

Definition

Sedimentation and diagenesis are the act and process of forming sedimentary rocks distributed on the surface of Earth's crust. During sedimentation, fluid sediment composed of particles flows, deposits, and accumulates with varying degrees of sorting. The accumulating sediments forms layers and becomes buried. After sedimentation, the process gradually progresses to diagenesis. Conditions such as temperature, pressure, and the chemical environment change with burial, in which sediment changes into solid rock.

Sedimentation is a phenomenon that occurs on the surface of Earth's crust, in which terrigenous clastic sediments, which are particle grains from rocks formed by weathering and/or erosion sink, accumulate and pile onto the surface and/or bottoms of water bodies by fluid motion. The origin of sedimentary particles can be divided into two types: clastic and biogenic. Gravel, sand, and mud are particles derived from sedimentary, igneous, and metamorphic rocks. Biogenic sediment includes skeletal remains of mollusks, echinoderms, foraminifera, diatoms, coral, etc.

Diagenesis is a gradual process that constitutes the first stage of burial after sedimentation. Variations in temperature and pressure with burial change the physical and chemical properties of the sediment. Dehydration, by increasing the overburden pressure, and the dissolution of minerals caused by an increase in the contact surface, leads to decreased porosity and increased density, which is the process of lithification.

Introduction

On the surface of the solid Earth, particles of the crust are mobile. Crustal surface features, such as mountains, are decomposed physically and chemically by weathering. The eroded particles are transported to the ocean and/or lake by fluids, wind, glaciers, and tides. Transported particles deposit at the bottom of the water and/or on the surface of the ground. This deposition process is called sedimentation. The sediments change their properties during burial. As the sediments are buried deeper, the sedimentary rocks notably change. On the surface of Earth's crust, weathering, erosion, transportation, deposition, burying, and metamorphism occur in that order. In this process, deposition and lithification during burial are “sedimentation” and “diagenesis,” respectively. These two processes are complicated and interact with each other. Because sedimentation and diagenesis are related to the generation of natural disasters and natural resources, these phenomena are also relevant to human society. Therefore, understanding the processes of sedimentation and diagenesis is important.

Sedimentation Processes and Sedimentary Structures

When a particle becomes gravitationally unstable due to the flow of air and/or water, the particle begins to move. When the flow speed decreases or the particle reaches a gravitationally stable position, the particles stop and deposit. Terrigenous clastic sediment generated in mountainous areas are transported along rivers and are deposited in lakes and the ocean. Lakes and the ocean are stable (low potential) environments for transported sediment particles. Large-grained gravels are deposited in estuaries and coasts with strong currents. On the other hand, fine silt and clay particles are carried offshore and deposited where the flow is calm.

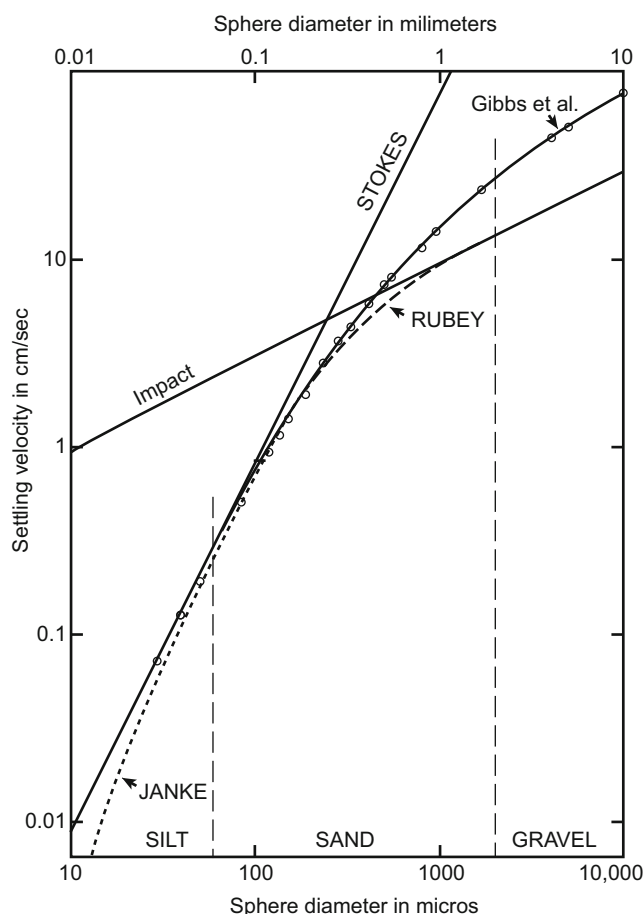
The settling velocity of sedimentary particles depends on the density of the fluid and the particle size. In the case of quartz with a perfect spherical shape, when the particle Reynolds number is less than 1 ($Re < 1$), the settling velocity depends on Stokes law:

$$v = \frac{(\sigma - \rho)g}{18\eta} d^2$$

where d , σ , ρ , η , and g are the radius of the particle, density of the particle, density of the fluid, viscosity coefficient, and gravitational acceleration, respectively. On the other hand, when the Reynolds number for the particle is more than 1 ($R_e > 1$), the settling velocity depends on the impact law (Newton's law):

$$v^2 = \frac{4(\sigma - \rho)g}{3C_D\rho} d$$

where C_D is drag coefficient. Figure 1 shows the relationship between the radius of the sphere and the settling velocity of quartz with a perfect spherical shape. The settling velocity of quartz particles whose radius is less than 100 μm mostly follow Stokes law, which is less than 1.0 cm/s. Large particles of quartz with radii greater than 2000 μm have settling velocities closer to the impact law, although there can be some differences.



Sedimentation and Diagenesis, Fig. 1 Settling velocity versus sphere diameter of quartz particles. (After Gibbs et al. 1971)

Sedimentary conditions such as the sedimentation rate, particle size, fluid density, etc. are recorded in sedimentary rocks as sedimentary structures (Fig. 2). Figure 2a shows the gradual distribution of particles. Large particles are distributed at the bottom and small particles are distributed at the top, which is called graded bedding. Graded bedding shows that the sorting of the particles depends on density differences. Figure 2b depicts a flame structure. When heavier sediments are layered on top of unconsolidated sediments immediately after deposition, the upper sediments sink due to their weight, creating a sedimentary structure in which the lower sediments flow upwards. In Fig. 2b, the white layer is composed of pumice, and the black layer includes scoria. The densities of pumice and scoria are generally $\sim 1.0 \text{ g/cm}^3$ and $\sim 2.0 \text{ g/cm}^3$, respectively. After the sedimentation of pumice, scoria was deposited on the pumice layer. At that time, a part of the pumice layer rose up to the surface because there was a density difference at the boundary of the two layers.

When particles are deposited via fluid currents, winds and waves, the surface of sediment is called a bedform and is related to the condition of the sediments. When the current speed increases at the same depth, a bedform composed of medium sand changes from flat, to ripple, dune, transition, plane bed, standing wave, and antidune. Evidence of the current is recorded in the sediments and/or rocks as a sediment structure called a lamination. Figure 2c shows a stripe pattern indicative of lamination. This pattern would typically have formed in a slow fluid flow and indicates that the sedimentary surface had previously been rippled.

Human Society and Sedimentation

Turbidity currents are one type of sedimentary gravity flows. Particles of sand and mud from land enter seawater, form high-density current flows, and are deposited. These current flows erode the seafloor surface. Terrestrial sediment includes organic matter such as plants and organisms, which become natural resources; therefore, turbidites play an important role in transporting organic sediments to the seafloor. On the other hand, turbidites can be generated by earthquakes (seismoturbidites) and tsunami deposits and are thus important in the study of disasters. Because seismoturbidites are evidence of past earthquake occurrences, recurrence intervals of earthquakes can be estimated by analyzing the distribution of seismoturbidites (cf. Adams 1990). Tsunami deposits are important records of large earthquakes in the past. Recently, new technology has been used in the field of sedimentology to understand the complex condition and structure of sedimentary flows. Mitra et al. (2021) reconstructed the flow conditions of tsunami deposits using a deep neural network inverse model and obtained a simulation result that was consistent with the observed values of the flow conditions, such as the



Sedimentation and Diagenesis, Fig. 2 Sedimentary structures in rocks. (a) Graded bedding. Large particles are distributed at the bottom and small particles are distributed at the top. (b) Flame structure. The white layer is composed of pumice, and the black layer includes scoria.

maximum inundation distance, flow velocity, and maximum flow depth. Their results indicate that the deep neural network inverse model has the potential to estimate the physical characteristics of modern tsunamis.

Diagenesis

As sediment is buried deep underground, the temperature and pressure rise; therefore, its physical and chemical properties change continuously over time. In the process of diagenesis, the following phenomena occur as the process generates rock from sediments.

- **Consolidation:** The porosity of the sediments decreases, and the density increases. The consolidation typically progresses with dehydration.
- **Cementation:** The constituent particles become bound to each other, and the pore space is filled by minerals. Representative filled minerals are carbonate and silicate minerals.
- **Replacement:** Primary minerals are replaced by secondary minerals. Calcareous sediment replaces dolomite during diagenesis (dolomitization).
- **Differential solution:** Differential solutions are caused by differences in constituent minerals. The phenomenon in which one of them dissolves discriminatively due to the pressure by the contact part of minerals and/or gravel is called the pressure solution.
- **Recrystallization:** The existing minerals grow crystals.
- **Authigenesis:** This process is a self-sustaining action of new minerals. The generation of authigenic minerals relates to the temperature and chemical state in diagenesis.

A part of the pumice layer rose up to the surface because there was a density difference at the boundary of the two layers. (c) Lamination. A stripe pattern would typically have formed in a slow fluid flow and indicates that the sedimentary surface had previously been rippled

Representative authigenic minerals are clay minerals such as kaolinite, smectite, montmorillonite, and illite.

- **Organic maturation:** Organic matter included in the sediments changes properties with increasing temperature and pressure due to burial.

Consolidation, cementation, recrystallization, replacement, differential solution, and authigenesis make sediments harder and transform sediments into rock (lithification). On the other hand, organic maturation is the reaction of organic matter included in the sediment, which is involved in the generation process of natural resources such as natural gas and oil. The generation of natural resources from organic matter is controlled by the environment conditions, especially temperature, and pressure.

Interaction of Sedimentation and Diagenesis

Generally, diagenesis is a process that occurs after sedimentation. However, there is a reverse situation in which diagenesis generates sediments. For example, consolidation, which is one of the diagenetic processes, decreases the pore volume due to a decrease in porosity with dehydration. The progress of consolidation is governed by the difference between the overburden pressure and the pore pressure. Immediately after deposition, the permeability is high because it is in an unconsolidated state. Therefore, dehydration easily occurs and consolidation progresses. However, in the case of low permeability, for instance, due to a high sedimentation rate, pore water cannot escape to the surface, and the pore pressure increases, in which consolidation is decelerated. The generated high pore pressure reduces the stability of the ground. As a result, the seafloor collapses, which forms mass transport

depositions (MTDs). In fact, MTDs distributed in the high pore pressure zone are confirmed. Kamiya et al. (2018) calculated the level of consolidation of formations including MTDs in the Boso Peninsula, Central Japan. As a result, the consolidation level (consolidation yield stress) of the layers including MTDs is less than that of other formations, which indicates that the high pore pressure prevented consolidation and generated mass transport. Thus, sedimentation and diagenesis can interact. The surface layer of the solid Earth is undergoing complex and diverse dynamic actions.

Summary

During sedimentation, different sedimentary structures are created depending on what, where, and how sediments are deposited. After deposition, the alteration process of minerals and organic matter differs depending on the environment in which they are buried. Through such a process of deposition and diagenesis, the complex surface material of the Earth is produced. In addition to research that explores past environments through observational analyses, numerical analyses based on probability theory are also being conducted. From the perspective of disaster prevention and the production of natural resources such as minerals and petroleum, research on Earth's surface will continue to be important.

Cross-References

- [Earth Surface Processes](#)
- [Earth System Science](#)
- [Exploration Geochemistry](#)
- [Flow in Porous Media](#)
- [Geohydrology](#)
- [Geomechanics](#)
- [Geotechnics](#)
- [Geothermal Energy](#)
- [Grain Size Analysis](#)
- [Mathematical Minerals](#)
- [Mine Planning](#)
- [Mineral Prospectivity Analysis](#)
- [Porosity](#)
- [Porous Medium](#)
- [Predictive Geologic Mapping and Mineral Exploration](#)
- [Turbulence](#)

Bibliography

- Adams J (1990) Paleoseismicity of the Cascadia Subduction Zone: Evidence from turbidites of the Oregon-Washington Margin. Vol. 9, Issue. 4, pp. 569–583. <https://doi.org/10.1029/TC009i004p00569>
- Gibbs JR, Matthews DM, Link AD (1971) The relationship between sphere size and settling velocity. *J Sediment Petrol* 41(1):7–18. <https://doi.org/10.1306/74D721D0-2B21-11D7-8648000102C1865D>
- Kamiya N, Utsunomiya M, Yamamoto Y, Fukuoka J, Zhang F, Lin W (2018) Formation of excess fluid pressure, sediment fluidization and mass-transport deposit in the Plio-Pleistocene Boso forearc basin, Central Japan. Geological society. London Special publications 477: 255–264. <https://doi.org/10.1144/SP477.20>
- Mitra R, Narusee H, Fujino S (2021) Reconstruction of flow conditions from 2004 Indian Ocean tsunami deposits at the Phra thong island using a deep neural network inverse model. *Natural Hazards and Earth System Sciences* 21:1677–1683. <https://doi.org/10.5194/nhess-21-1667-2021>

Self-Organizing Maps

Sid-Ali Ouadfeul¹, Leila Aliouane² and Mohamed Zinelabidine Doghmane³

¹Algerian Petroleum Institute, Sonatrach, Boumerdes, Algeria

²LABOPHT, Faculty of Hydrocarbons and Chemistry, University of Boumerdes, Boumerdes, Algeria

³Department of Geophysics, FSTGAT, University of Science and Technology Houari Boumediene, Algiers, Algeria

Definition of the Biological and Formal Neurons

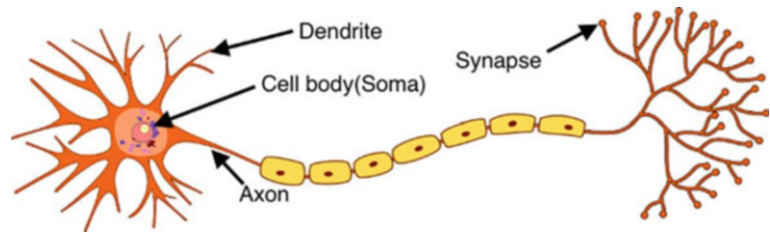
In the brain, neurons are linked together through axons and dendrites; Fig. 1 is a synthetic presentation of the human neuron. As a first approach, we can consider that these links are conductors and can thus convey messages from one neuron to another. The dendrites are the inputs of a neuron and its axon are the outputs (Mejia 1992).

A neuron emits an electrical signal based on coming signals from other neurons. We observe, in fact, at the level of a neuron, a summation of signals received over time and the neuron in turn emits an electrical signal when the sum exceeds certain threshold.

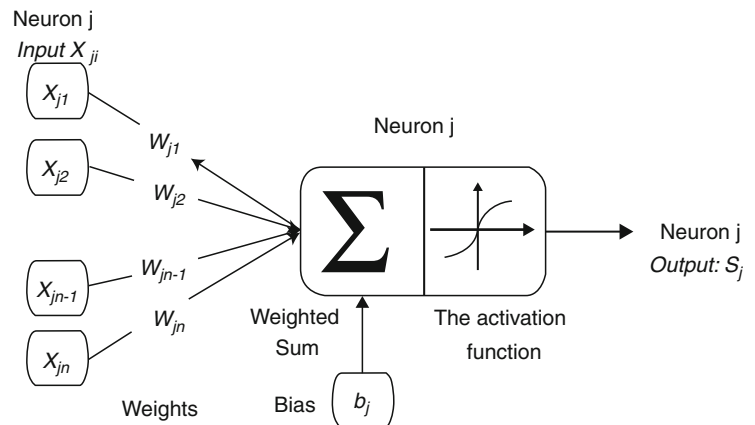
A formal neuron is a mathematical and computational presentation of a biological neuron (McCulloch and Pitts 1943); it is a simple unit able to carrying out some elementary calculations. A large number of these units are then connected together, and an attempt is made to determine the computing power of the network thus obtained.

The formal neuron is therefore a mathematical modeling that mimics the functioning of the biological neuron, in particular the summation of inputs, at a biological level, synapses do not all have the same “value” (the connections between the neurons being more or less strong). McCulloch and Pitts (1943) and Rosenblatt (1958) created an algorithm that weights the sum of its inputs by synaptic weights (weighting coefficients). In addition, the 1 s and −1 s at the input are used

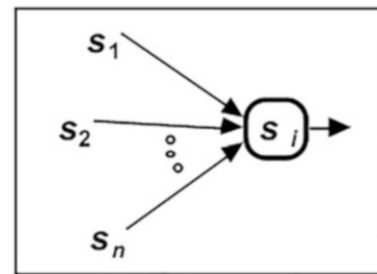
Self-Organizing Maps,
Fig. 1 Biological neuron
 (McCulloch and Pitts 1943)



Self-Organizing Maps,
Fig. 2 Formal neuron
 (McCulloch and Pitts 1943)



there to represent an excitatory or inhibitory synapse, Fig. 2 is a graphical presentation of the formal neuron. The self-organizing map represents a specific type of artificial neural network where its main utility is to map to study and map the distribution of data in a large-dimension, thus, its application can provide many computational advantages in Geosciences even its learning process is unsupervised.



Overview of Artificial Neuron Characteristics

A connectionist network is composed of neurons, which interact to give to the network its global behavior (Tian et al. 2021). In connectionist models, these neurons are elementary processors whose definition is made in analogy with nerve cells (Le Cun 1985). These basic units receive signals from outside or from other neurons in the network, they calculate a function, usually simple, of these signals and in turn send signals to one or more other neurons or to the outside. A neuron is characterized by three parameters: its state, its connections with other neurons, and its transition function (Anderson and Rosenfeld 1988).

The State

An artificial neuron is an element that has an internal state, in fact, it receives signals that eventually allow it to change its state (MacLeod 2021). We will denote S_i the set of possible states of a neuron. It could be for example $\{0, 1\}$, where 0 will be interpreted as the inactive state and 1 the active state (Tian et al. 2021).

Self-Organizing Maps, Fig. 3 Presentation of a neuron (Mejia 1992)

The state S_i of the neuron is a function of the inputs $S_1, \dots, S_n$. The neuron produces an output, which will be transmitted to the connected neurons. To calculate the state of a neuron we must therefore consider the connections between this neuron and other neurons (Fig. 3).

Connections Between Neurons or Architecture

A connection is an established link between two neurons; connections are also called synapses, in analogy with the name of the connectors of real neurons. A connection between two neurons has an associated numerical value called the connection weight. The connection weight W_{ij} between two neurons j and i can take discrete values in $\mathbb{Z}$ or continuous in $\mathbb{R}$. The information passing through the connection will be affected by the value of the corresponding weight. A connection with a weight $W_{ij} = 0$ is equivalent to no connection.

The Transition Function

We are interested here in neurons that calculate their state from the information they receive. We will use the following notation below:

S : The set of possible states of neurons

X_i : The state of a neuron i , where $X_i \in S$.

A_i : The activity of neuron i

W_{ij} : The weight of the connection between neurons j and i

The activity of a neuron is calculated according to the states of the neurons in its neighborhood and the weights of their connections, according to the following formula:

$$A_i = \sum_j W_{ij} X_j \quad (1)$$

The state of a neuron i is a function of the states of neurons j , of its neighborhood, and of the weights of the connections W_{ij} , this function is called the transition function.

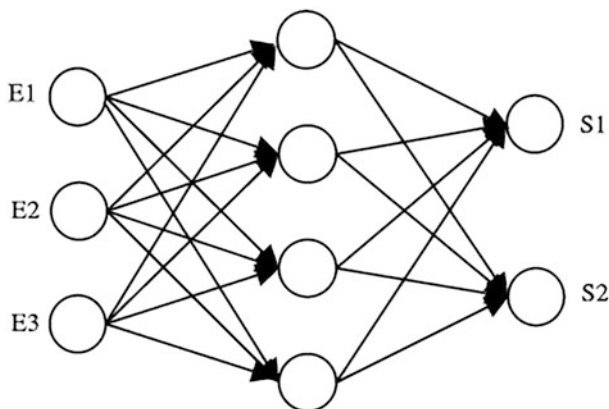
These elements are assembled according to certain architecture of a network. This architecture defines a composition of elementary functions that can be used in several ways during the operation of the network; this is called dynamic of a network (Mejia 1992).

Overview of Topologies of Neural Networks

There are several topologies of neural networks:

Multilayer Networks

They are organized in layers. Each neuron generally takes as input all the neurons of the lower layer (Fig. 4). They do not have cycles or intra-class connections. We then define an input layer, an output layer, and n hidden layers (Mejia 1992).



Self-Organizing Maps, Fig. 4 Multilayer perceptron with one hidden layer

Networks with Local Connections

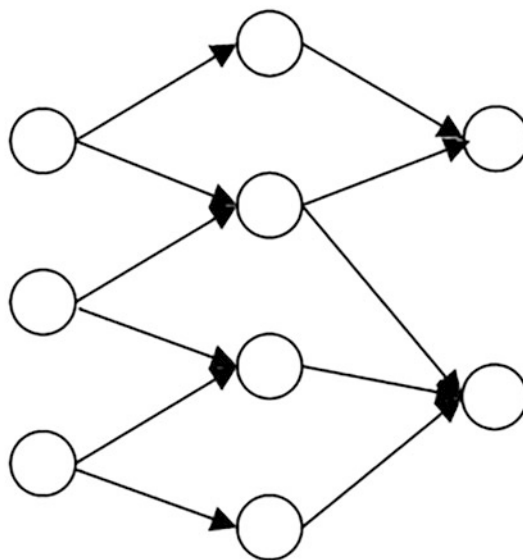
A neuron is not necessarily connected to all the neurons of the previous layer (Fig. 5)

Recurrent Connection Networks

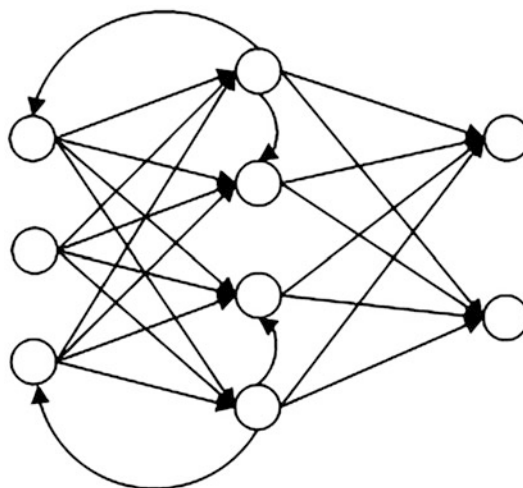
We always have a layered structure, but with returns or possible connections between neurons of the same layer (Fig. 6)

Networks with Complete Connections

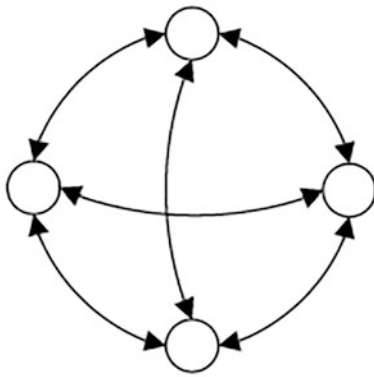
All neurons are interconnected (Fig. 7) (Karacan 2021).



Self-Organizing Maps, Fig. 5 Neural network with connections



Self-Organizing Maps, Fig. 6 Neural network with recurrent connections



Self-Organizing Maps, Fig. 7 Neural network with complete connections

Definition of Learning and Overview of Modes

One of the most interesting characteristics of neural networks is their ability to learn. Learning will allow the network to modify its internal structure (synaptic weights) to adapt to its environment (Rumelhart and Mc Clelland 1986)

In the context of neural networks, a network is defined by its connection graph and the activation function of each neuron. Each choice of synaptic coefficients (weights of connection) corresponds to a system, the learning operation is the seeking of the best weights that solve a given problem. To be able to evaluate a particular system, we make a series of experiments to observe the behavior of the network. An experiment is the presentation of the input to the system; the response is collected at the output. The network is evaluated by the value of an error function; a learning problem is then to find a network that minimizes this function.

Learning Modes

We distinguish three learning modes: supervised, unsupervised and reinforced learning.

Supervised Learning

In this mode, a teacher who is perfectly familiar with the desired or correct output guides the network by teaching it the right result at each step. The supervised learning consists to compare the obtained result by the artificial network with the desired output, then to adjust the weights of the connections to minimize the difference between the calculated and the desired output (MacLeod 2021).

Unsupervised Learning

In the unsupervised learning, the network modifies its parameters taking into account only local information. This learning doesn't need any predetermined desired outputs. Networks using this technique are called self-dynamic networks and are considered to be regularity detectors because the network

learns by detecting the regularities in the structure of the input patterns and produces the most satisfactory output.

Reinforced Learning

It is used when the feedback on the quality of performance is provided, the desired output is not completely specified by a teacher. So learning is less directed than supervised learning. Unlike unsupervised learning where no feedback is given, the reinforced Learning Network can use the reinforcement signal to find the most needed desirable weights.

Concept and Construction of Self-Organizing Maps

Self-adaptive or self-organizing map is a class of artificial neural network based on unsupervised learning mode (Tian et al. 2021). It is often referred to by the English term Self-Organizing Map (SOM), or even Teuvo Kohonen's map named after the statistician who developed the concept in 1992.

They used a map in the space to study the distribution of data in a large-dimension. In practice, this mapping can be used to perform discretization, vector quantization, or classification tasks (Silversides 2021). These intelligent data representation structures are inspired, like many other creations of artificial intelligence, from human biology (MacLeod 2021). They involve reproducing the principle of the vertebrate brain: stimuli of the same nature excite a very specific region of the brain. Neurons are organized in the cortex to interpret every kind of imaginable stimuli. In the same way, the self-adaptive map is deployed in order to represent a set of data (Wen et al. 2021), and each neuron represents a very particular group of data according to the common characteristic that bring them together. It allows a multi-dimensional visualization of crossed data. The map performs vector quantization of the data space, it means discrediting the space and dividing it into zones. A significant characteristic called the referent vector is assigned to each zone. The space V is divided into several zones and W_r presents a referent vector associated with a small area of the space V_r and $M, r(2,3)$ presents its associated neuron in the grid A (Fig. 8). Each area can be easily addressed by the indexes of the neurons in the grid (Kohonen 1992 1998).

From an architectural point of view, a Kohonen's self-organizing maps is made up of a grid (most often one-dimensional or two-dimensional). In each node of the grid there is a neuron. Each neuron is linked to a referent vector, responsible for an area in the data space (called the input space) (see Fig. 8). In a self-organizing map, the referent vectors provide a discrete representation of the input space. They are positioned in such a way that they retain the topological shape of the entry space. By keeping the neighborhood

relations in the grid, they allow an easy indexing (via the coordinates in the grid). This is useful in various fields, such as classification of textures, data interpolation, visualization of multidimensional data (Silversides 2021). Let A the rectangular neural grid of a self-organizing map. A neuron map assigns to each input vector V a neuron designated by its position vector, such that the reference vector W_r is closest to V .

Mathematically, we express this association by a function (Kohonen 1992):

$$\varphi_m : V \rightarrow A \quad (2)$$

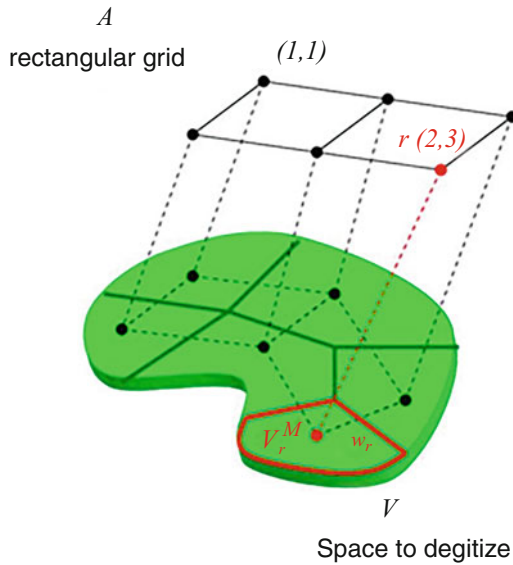
$$r = \phi_m(v) = \arg \min_{r \in A} \|v - w_r\| \quad (3)$$

This function allows us to define the applications of the card. For the quantifier vector, we approximate each point in the input space by the closest referent vector by:

$$W_r = \varphi_m^{-1}(\varphi_w(v)) \quad (4)$$

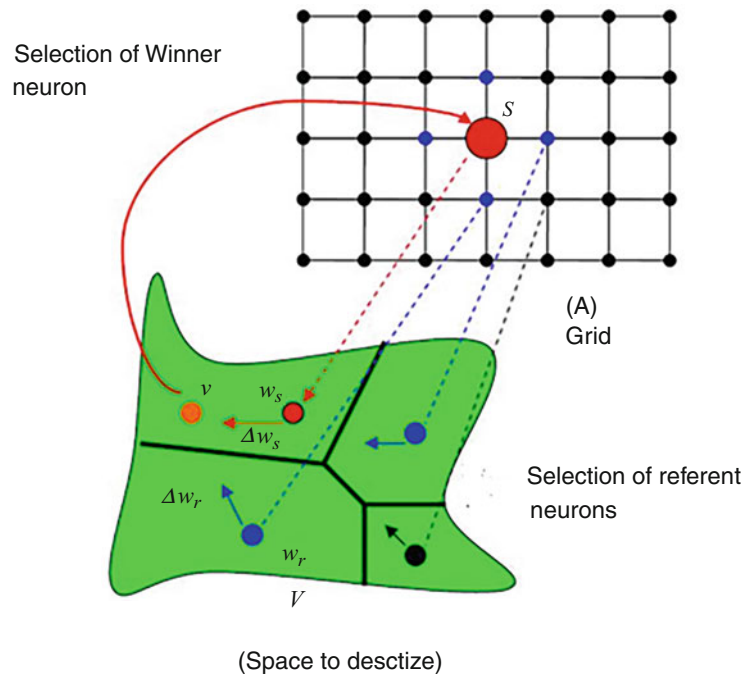
For the classifier we use the function $r = \phi_w(v)$. Each neuron in the grid is assigned a label corresponding to a class. All the points of the entry space which are projected on the same neuron belong to the same class. The same class can be associated with several neurons.

The learning algorithm is based on the concept of the winner neuron, after a random initialization of the values of each neuron, the data are submitted one by one to the self-adaptive map (Agterberg and Cheng 2022). Depending on the values of the neurons, there is one that will respond best to the stimulus (the one whose value will be closest to the data presented). Then this neuron will be rewarded with a change in value so that it responds even better to another stimulus of the same nature as the previous one. In the same way, the neurons neighboring the winner are also somewhat rewarded with a gain multiplying factor of less than one. Thus, it is the entire region of the map around the winning neuron that specializes. At the end of the algorithm, when the neurons no longer move, or very little, at each iteration, the self-organizing map covers the entire topology of the data (see Fig. 9)



Self-Organizing Maps, Fig. 8 Self-Organizing maps architecture

Self-Organizing Maps, Fig. 9 Self-organization algorithm for the Kohonen model



The mapping of the input space is carried out by adapting the referent vectors W_r . The adaptation is made by a learning algorithm whose power lies in the competition between neurons and in the importance given to the notion of neighborhood.

A random sequence of input vectors is presented during training. With each vector, a new adaptation cycle is started. For each vector V in the sequence, we determine the winning neuron, i.e., the neuron whose referent vector approaches V as best as possible:

$$S = \varphi_m(v) = \arg \min_{r \in A} \|v - W_r^t\| \quad (5)$$

The winning neuron S and its neighbors (defined by a neighborhood membership function) move their referent vectors to the input vector

$$W_r^{t+1} = W_r^t + \Delta W_r^t \quad (6)$$

With

$$\Delta W_r^t = \varepsilon \cdot h(v - W_r^t) \quad (7)$$

where $\varepsilon = \varepsilon(t)$ represents the learning coefficient and $h = h(r, s, t)$ the function which defines membership in the neighborhood (Sreevalsan-Nair 2021).

The learning coefficient defines the amplitude of the overall displacement of the map. The notion of neighborhood is inspired from the cortex, where neurons are connected to each other. This is the topology of the map. The shape of the map defines the neighborhoods of the neurons and therefore the links between neurons. The neighborhood function describes how the neurons in the vicinity of the winner are drawn into the corrective movement (Sreevalsan-Nair 2021). In general, we use:

$$h(r, s, t) = \exp\left(-\frac{\|r - s\|}{2\sigma^2(t)}\right) \quad (8)$$

Where σ is called the neighborhood coefficient. Its role is to determine a neighborhood radius around the winning neuron. The neighborhood function h forces neurons in the neighborhood of S to move their referent vectors closer to the input vector V (Sreevalsan-Nair 2021). The closer a neuron is to the winner in the grid, the less its displacement is important. The correction of referent vectors is weighted by the distances in the grid. This reveals, in the input space, the order relations in the grid.

During learning, the map described by the referent vectors of the network evolves from a random state to a state of stability in which it describes the topology of the input

space while respecting the order relations in the grid (Kohonen 1998).

Application to Petroleum Geosciences

In petroleum exploration and production, the prediction of lithofacies is an important issue, where lithofacies is the first task to identify in petroleum reservoir characterization in order to determine layer limits of a crossed formation of a well and geological rock type using petrophysical recordings crossed a well named well-logs data with core analysis. The recuperation of core is an expensive process and not always continuous (Chen and Zhang 2021).

Here, we show the efficiency of self-organizing maps to predict lithofacies from raw well-logs data of two boreholes named Well01 and Well02, located in the Algerian Sahara.

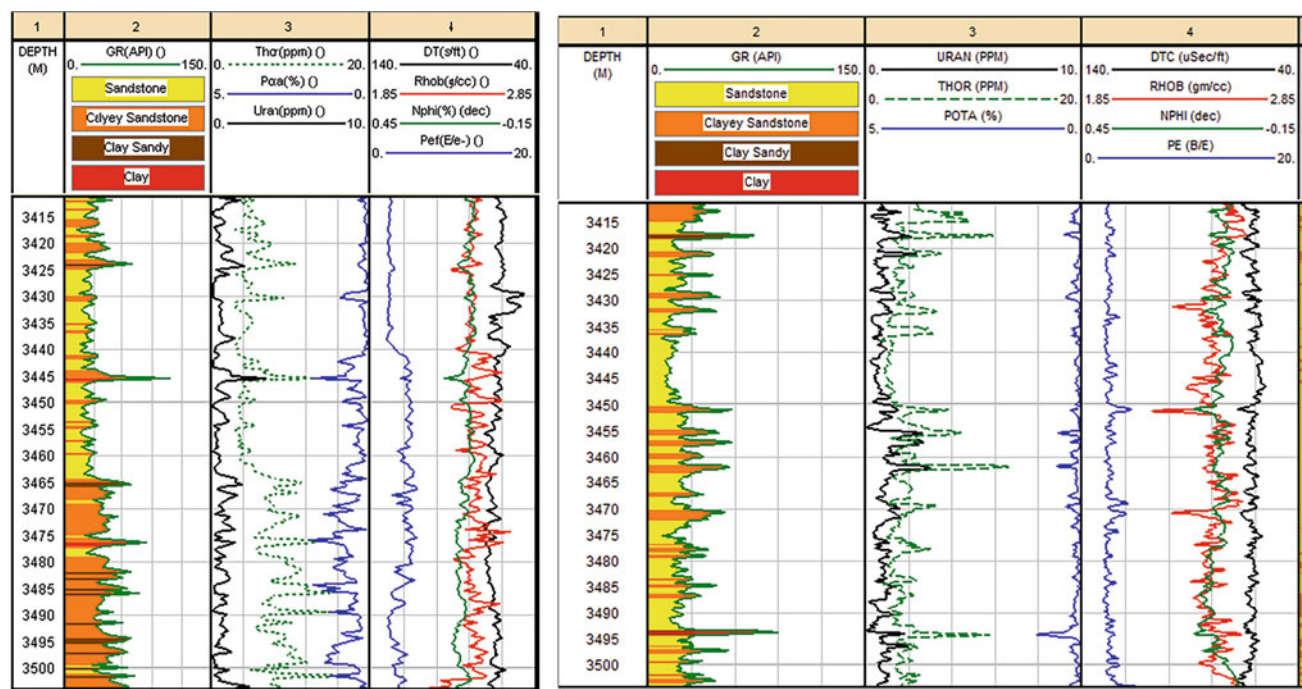
Raw well-logs data are: Natural Gamma ray, Natural Gamma ray spectroscopy (Thorium, Potassium and Uranium concentration), Bulk density, Neutron porosity, Photoelectric absorption coefficient and Slowness of the P wave. Figure 10 presents these raw logs data versus the depth, only the depth interval [3410 m, 3505 m] is investigated with a sampling interval of half a foot and the obtained lithofacies using the natural Gamma ray log (see track 02). In this case we take the following criteria: $0 < \text{Gamma ray} < 30$ is a sandstone; $30 < \text{Gamma ray} < 60$ is a Clayey Sand $60 < \text{Gamma ray} < 90$ is a Sandy; Gamma ray > 90 is a Clay.

Raw well-logs are injected in a self-organizing map composed of 08 neurons in the input layer and 04 neurons in the output layer, the goal is to train this map, at the end of the learning weights of connection between neurons are optimized. The obtained lithofacies using the Gamma ray log are used to index the map and one of the following four lithologies: Clay, Sandstone, Clayey Sandstone, Sandy Clay, is attributed to each class number. After the learning stage, Raw well-logs data of the two wells are propagated through the machine using the optimized weights of connection.

Figure 11 shows the obtained lithofacies using the Gamma ray log (track A) and the implemented Kohonen Self-Organizing map.

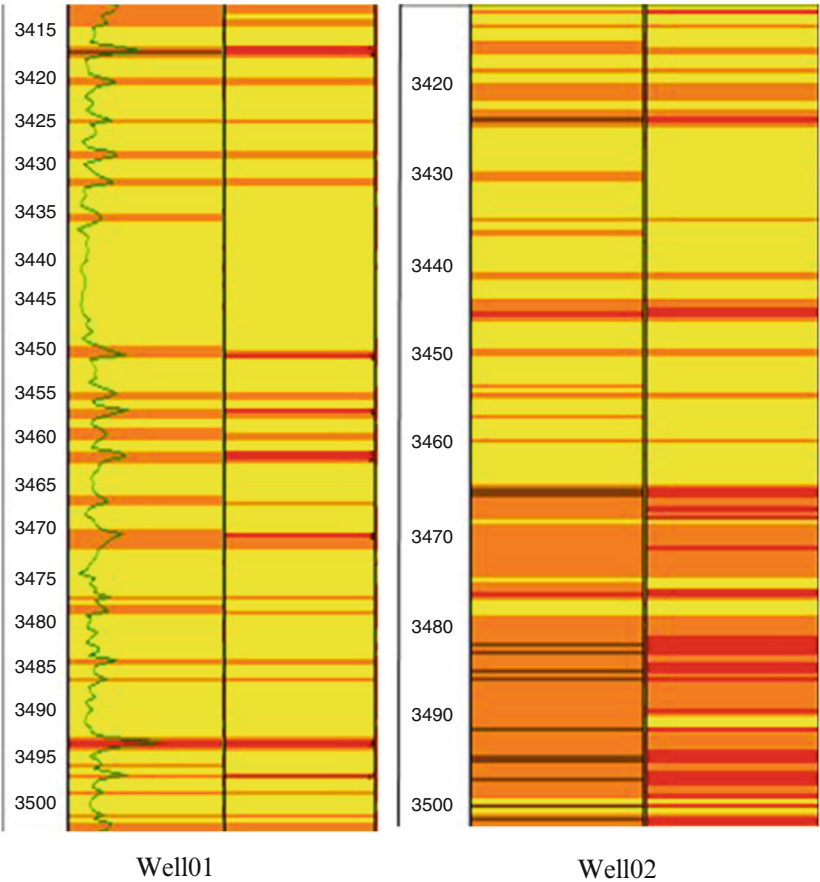
Conclusions

Self-Organizing Maps (SOMs) is a class of neural network based on the unsupervised learning, they are used to map a real space, to study the distribution of data in a large-dimension and involve reproducing the neuronal principle of the vertebrate brain. SOMs are based on the principle of neighborhood and winner neuron. Application to real well-logs data proves the power of the SOMs to classify



Self-Organizing Maps, Fig. 10 Raw well-logs data and lithofacies classification using the Gamma ray log

Self-Organizing Maps,
Fig. 11 Lithofacies Prediction
using the Gamma ray log (track A)
and the implemented Self-
Organizing map (track B) for the
pilot well (Well01) and the
validation well (Well02)



lithofacies in case of absence of core cock measurements, they have high economic benefit.

Cross-References

- Artificial Intelligence in the Earth Sciences
- Artificial Neural Network
- Chaos in Geosciences
- Data Mining
- K-nearest Neighbors
- Multilayer Perceptrons
- Pattern Classification
- Wavelets in Geosciences

Bibliography

- Anderson JA, Rosenfeld E (1988) Neuro computing foundations of research. MIT Press, Cambridge
- Kohonen T (1992) Self organizing and associative memory. Springer series in information sciences, 8, 2nd edn. Springer, Berlin, 2012, 316 Pages.
- Kohonen T (1998) Self organization and associative memory. Springer series in information sciences, 8, 2nd edn. Springer, Berlin
- Le Cun Y (1985) Une procédure d'apprentissage pour réseau à seuil 331 asymétrique. COGNITIVA 85, Paris
- McCulloch W, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 5:115–133
- Mejia C (1992) Architectures Neuronales pour l'Approximation des Fonctions de Transfert: application à la télédétection, thèse de doctorat. Université de Paris Sud
- Rosenblatt F (1958) The perceptron: probabilistic model for information storage and organization in the brain. Psychol Rev 65:386–408
- Rumelhart DE, McClelland JL (1986) Parallel distributed processing: exploration in the MicroStructure of cognition. MIT Press, Cambridge

Sensitivity Analysis

Valentina Ciriello¹ and Daniel M. Tartakovsky²

¹Dipartimento di Ingegneria Civile, Chimica, Ambientale e dei Materiali (DICAM), Università di Bologna, Bologna, Italy

²Department of Energy Resources Engineering, Stanford University, Stanford, CA, USA

Definition

Sensitivity analysis (SA) is a procedure used to rank the various input components by the degree to which changes in their values impact the model output. The input may include model parameters, boundary and initial conditions, and model assumptions. SA has become an integral part of mathematical

modeling (e.g., Razavi et al. 2021). It is used to estimate the robustness of model calibration, to reduce the input space's dimensionality for subsequent uncertainty quantification, and to guide data collection campaigns or system design (e.g., Ciriello et al. 2015; Um et al. 2018) or decision-making under uncertainty (e.g., Tartakovsky 2013).

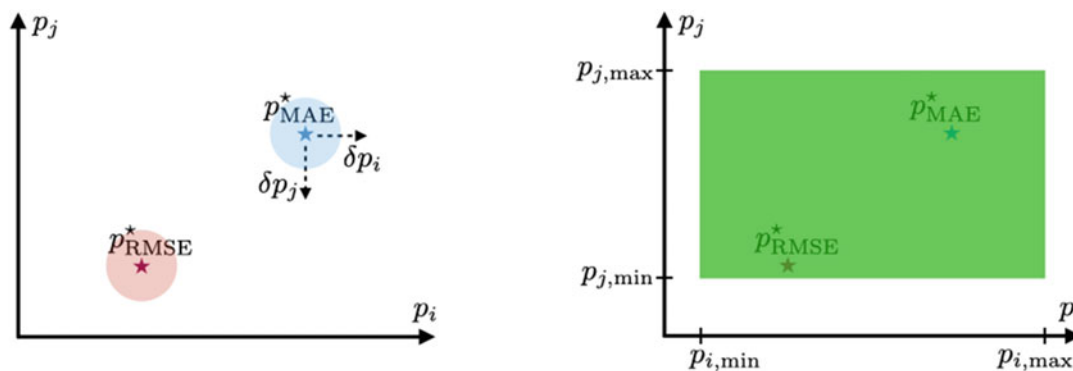
Local Sensitivity Analysis (LSA)

Suppose that a model input $\mathbf{p}$ consists of N_{par} components $\mathbf{p} = \{p_1, \dots, p_{N_{\text{par}}}\}$. Model calibration consists of identifying an input $\mathbf{p}^* = \{p_1^*, \dots, p_{N_{\text{par}}}^*\}$ —a point in the input space Ω_p of all possible combinations of the input $\mathbf{p}$ —that minimizes a discrepancy between the model predictions and observations. LSA is a suite of techniques whose aim is to gauge the impact of a slight deviation $\delta\mathbf{p}$ from the point or “locale” $\mathbf{p}^*$ on the model output (Fig. 1); the components p_i ($i = 1, \dots, N_{\text{par}}$) are ranked by the magnitude of change in the model output due to the change from p_i^* to $p_i^* + \delta p_i$, while keeping the remaining components p_k ($k \neq i$) fixed at their respective values p_k^* . LSA remains popular in geosciences because of its ease of implementation and scalability with N_{par} . Yet, it has a number of shortcomings and limitations.

First, the ubiquitous data scarcity and spatiotemporal variability of the input $\mathbf{p}$ often undermine the reliability of model calibration efforts by rendering the “optimal” model parameterization, i.e., the identification of the point $\mathbf{p}^*$, nonunique. Second, the existence of alternative metrics of the discrepancy between model predictions and observations, e.g., root mean square error or mean absolute error, serves as an additional source of nonuniqueness of model calibration (Fig. 1). The absence of a uniquely defined point $\mathbf{p}^*$ undermines *raison d'être* for LSA. Third, in its standard form, LSA assumes both model predictions and observations to be error-free (deterministic). That assumption seldom, if ever, holds in geosciences due to pervasive uncertainty.

Global Sensitivity Analysis (GSA)

GSA overcomes these limitations by adopting the probabilistic framework and equipping each input component, p_i , with its range of variability, $[p_i^{\min}, p_i^{\max}]$, and with a probabilistic model (i.e., cumulative distribution function F_{p_i}) defined on that range. The choice of the ranges and distributions is subjective and reflects the degree of uncertainty in input values. Rather than focusing on small deviations from a single point $\mathbf{p}^*$ one input element at a time, GSA allows the input $\mathbf{p}$ to span the whole input space $\Omega_p = [p_1^{\min}, p_1^{\max}] \times \dots \times [p_{N_{\text{par}}}^{\min}, p_{N_{\text{par}}}^{\max}]$ simultaneously (Fig. 1). In doing so, GSA makes no assumption about either uniqueness or optimality of a



Sensitivity Analysis, Fig. 1 Left: Local sensitivity analysis is conducted by perturbing the calibrated input set $\mathbf{p}^*$ by a small amount $\delta \mathbf{p}$. The point $\mathbf{p}^*$ is nonunique, depending, among other things, on the choice of the discrepancy metric (e.g., root mean square error or mean

absolute error) between model predictions and observations. Right: Global sensitivity analysis allows the input $\mathbf{p}$ to span the whole range of its possible values

particular set of parameters obtained via calibration. GSA estimates the degree to which each random (uncertain) input component, p_i , and interactions between different components, p_i and p_k with $i \neq k$, contribute to uncertainty in the model predictions. The past decades witnessed the widespread adoption of GSA by the computational community, including in geosciences (Ferretti et al. 2016).

Variance-based GSA quantifies the predictive uncertainty of a model in terms of the variance of the model output. Thus, the Sobol' indices, the most widely used GSA metrics, rely on the analysis of variance (ANOVA) to apportion the output variance among the variances of the input components and their interactions (e.g., Ciriello et al. 2015; Ferretti et al. 2016). Variance-based GSA is strictly applicable to uncorrelated random inputs. Its application to systems, whose input components are correlated random fields and/or are mutually correlated, is valid as an uncontrollable approximation. This limitation can be overcome by deploying the Rosenblatt transform to map correlated input components onto independent, identically distributed random variables (e.g., Um et al. 2019, and references therein).

Variance-based GSA is of limited value for the systems with highly non-Gaussian (e.g., multimodal) input and output. Such problems call for the use of distribution-based GSA techniques. Examples of the latter include analysis of the difference between the unconditional distribution of the output and its conditional counterparts when one or more inputs are fixed (e.g., Ciriello et al. 2019), and information-theoretic approaches and Bayesian networks (e.g., Um et al. 2019).

By identifying the most influential factors governing a given system, these and other variants of GSA can be used to guide data acquisition campaigns and monitoring activities; additional data collected on the basis of GSA findings narrow the degree of uncertainty in model output (Ciriello et al. 2015).

Numerical Implementation of GSA

GSA is typically performed via Monte Carlo simulations, which can be prohibitively expensive for complex models common in geosciences. A viable solution to this problem is to use surrogates (emulators). A surrogate is a computationally efficient model, such as a polynomial function or a deep neural network (e.g., Zhou and Tartakovsky 2021, and references therein), which is calibrated on a dataset generated by running the original model multiple times.

Polynomial chaos expansions (PCEs) are used to construct surrogate models and to accelerate variance-based GSA (e.g., Ciriello et al. 2017). For example, Sobol' metrics can be expressed as an analytical post-processing of PCE coefficients, drastically reducing the computational cost of variance-based GSA (Sudret 2008). PCEs also allow one to accelerate the numerical implementation of distribution-based GSA (Ciriello et al. 2019).

Summary

Decision-making under uncertainty in geosciences benefits from the use of SA as part of mathematical modeling. SA can be applied both in forward and inverse uncertainty quantification problems. The increasing adoption of GSA in geosciences is associated with the adoption of surrogate modeling techniques and/or efficient and robust algorithms, which drastically reduce the computational cost associated with the definition of global sensitivity metrics.

Bibliography

Ciriello V, Edery Y, Guadagnini A, Berkowitz B (2015) Multimodel framework for characterization of transport in porous media. *Water Resour Res* 51:3384–3402

- Ciriello V, Lauriola I, Bonvicini S, Cozzani V, Di Federico V, Tartakovsky DM (2017) Impact of hydrogeological uncertainty on estimation of environmental risks posed by hydrocarbon transportation networks. *Water Resour Res* 53(11):8686–8697
- Ciriello V, Lauriola I, Tartakovsky DM (2019) Distribution-based global sensitivity analysis in hydrology. *Water Resour Res* 55:8708–8720
- Ferretti F, Saltelli A, Tarantola S (2016) Trends in sensitivity analysis practice in the last decade. *Sci Total Environ* 568:666–670
- Razavi S, Jakeman A, Saltelli A, Prieur C, Iooss B, Borgonovo E, Plischke E, Lo Piano S, Iwanaga T, Becker W, Tarantola S, Guillaume JHA, Jakeman J, Gupta H, Melillo N, Rabitti G, Chabridon V, Duan Q, Sun X, Sheikholeslami R, Smith S, Hosseini N, Asadzadeh M, Puy A, Kucherenko S, Maier HR (2021) The future of sensitivity analysis: An essential discipline for systems modeling and policy support. *Environ Model Softw* 137:104954
- Sudret B (2008) Global sensitivity analysis using polynomial chaos expansions. *Reliab Eng Syst Saf* 93:964–979
- Tartakovsky DM (2013) Assessment and management of risk in subsurface hydrology: A review and perspective. *Adv Water Resour* 51:247–260
- Um K, Zhang X, Katsoulakis M, Plechac P, Tartakovsky DM (2018) Global sensitivity analysis of multiscale properties of porous materials. *J Appl Phys* 123(7):075103
- Um K, Hall EJ, Katsoulakis MA, Tartakovsky DM (2019) Causality and Bayesian network PDEs for multiscale representations of porous media. *J Comput Phys* 394:658–678
- Zhou Z, Tartakovsky DM (2021) Markov chain Monte Carlo with neural network surrogates: Application to contaminant source identification. *Stoch Env Res Risk A* 35(3):639–651

Sequence Classification

Mehala Balamurali

Australian Center for Field Robotics, University of Sydney, Sydney, NSW, Australia

Definition

In general, a sequence can be defined as a series of events that are happened in order. These events can be arranged as a simple symbolic values, a vector of complex sequence, and a time series sequences where a single or multiple numerical vectors associate with timestamps or any other complex data types (Xing et al. 2010). In machine learning, sequence analysis is used for inferring the next value, the class label of sequence, or the next sequence based on the prior pattern of the data in the sequence. Sequence classification is a method to infer the class label of unseen sequence by training the classification model with labeled sequence data.

Sequence Classification in Geosciences

In geology, layers of sediment deposits are known as stratigraphic sequences. Stratigraphic data and geophysical logs

provide information about the deposits and the seismic sequences. Hidden Markov models are used to classify log facies and identify seismic events by considering the transition probability from one facies to another, where the unspecified facies profiles are considered as a sequence of unknown states. The model utilizes expectation–maximization algorithm to estimate the unknown parameters (Lindberg and Grana 2015; Cárdenas-Peña et al. 2013).

Random forests (RF) and support vector machines (SVM) are successfully applied in other classification tasks such as geochemical and material classifications. Of these studies, RF and SVM methods have been adopted in identifying and classifying pattern in sequential data. Cracknell and Reading (2013) compared the performance of RF and SVM to categorize high-grade meta-sedimentary and meta-volcanic rocks of the Broken Hill region, Australia, using airborne geophysical and Landsat multispectral imagery. SVM was found superior to other supervised and unsupervised methods in classifying lithofacies using the combined well log data from Devonian Bakken and Mahantango-Marcellus (Bhattacharya et al. 2016). SVM incorporating dynamic time warping DTW feature extractor has been used for classifying time series data (Gudmundsson et al. 2008).

Many studies have been conducted, on the development of automatically classified geophysical log data based on machine learning applications. However, choosing the optimal features that can classify signatures from the signals is difficult in traditional machine learning techniques. Unlike machine learning techniques, deep learning models can automatically learn the optimal features through multiple stages of learning. Among them RNN (Tang et al. 2016; Chung et al. 2014) is an effective neural network for sequence analysis, such as natural language processing, speech recognition, and music generation. Long short-term memory or LSTM (Hochreiter and Schmidhuber 1997) networks are a special kind of RNNs that deal with the long-term dependency problem effectively. These models have been successfully used to extract the sedimentary facies information from well logs (Song et al. 2020). The aptitude of bidirectional long short-term memory (Bi-LSTM) was tested for retrieving information and text mining from the geoscience reports (Qiu et al. 2018).

One-dimensional CNN models were successfully applied in many sequence analysis problems such as solving inversion of electromagnetic problem (Puzyrev and Swidinsky 2021) and classifying hyperspectral data that are used in various applications of geosciences and remote sensing (Wei et al. 2019). A combination of deep CNN model with LSTM and RF has shown promising results on simulating the dynamics of land use change (Xing et al. 2020).

The following section briefly describes the sequence classification methods using deep learning. Please refer the entries in this book for other machine learning algorithms

that are used in sequence classification: random forest, neural network, Markov chains.

Sequence Classification Using Deep Learning Algorithms

Deep neural network (DNN) models are compelling machine learning algorithms that showed promising results on difficult tasks such as image classification, speech recognition, and object detection. DNNs are flexible, powerful, and they work well when there is enough ground truth data, but their applicability is limited for analyzing sequences because their inputs and targets need to be sensibly programmed as fixed dimensionality vectors. Many real-world problems are happened as sequential events where the length of event is unknown so as their dimensions. Below are some deep learning algorithms that are successfully applied in sequence learning.

Recurrent Neural Networks (RNNs)

Unlike traditional neural network, recurrent neural networks have loop in them. Hence the model allows the cells to continue the information to adjacent cells (Fig. 1). As RNNs can remember the features of preceding inputs and outputs, the model is much suitable for inferring the information of following sequence. However, the model cannot learn the longtime dependency of the sequence.

LSTM Networks

Long short-term memory networks are a special kind of recurrent network that can handle the long-term dependencies effectively by removing the limitation of RNN, such as gradient vanishing and exploding, complex training, and difficulty to process very long sequences. Similar to standard RNN, LSTMs are made of chain of repeating modules.

However, the module of LSTM has multiple recurrently connected memory cells and three multiplication units: a forget gate, input gate, and output gate (Graves and Schmidhuber 2005).

As seen in the Fig. 2, the classification network consists of a sequence input layer, LSTM layer, fully connected layer, a SoftMax layer, and an output layer. The model can be further deepening by introducing additional LSTM layers. Furthermore, LSTM combined with CNN layers can be used for analyzing video or time-stamped images where CNN is used to learn the spatial information.

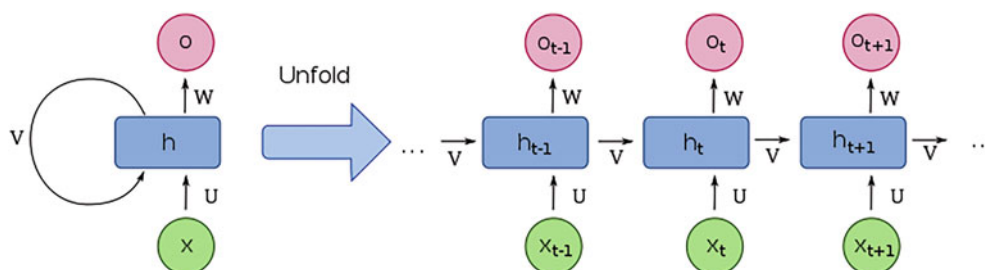
LSTM networks can be used for sequence-to-label classification and sequence-to-sequence classification. In **sequence-to-label classification**, the network is trained with different sequences and the corresponding label for each sequence. The trained model is then used to classify the unseen sequences. In **sequence-to-sequence classification**, the network is trained with a single sequence X_{Train} that can have multiple dimensions and the corresponding sequence of categorical labels Y_{Train} . The trained model can make prediction on every time step with categorical label.

1D CNNs

Figure 3 shows an architecture of 1D CNN that is used to classify sequences. The convolution layer in this model uses 1D filters/kernels to process the input data. Similar to 2D and 3D CNN models, 1D coevolution blocks also include activation function and pooling layers. Please refer ▶ “Deep CNN-Based AI for Geosciences” for further details.

Summary

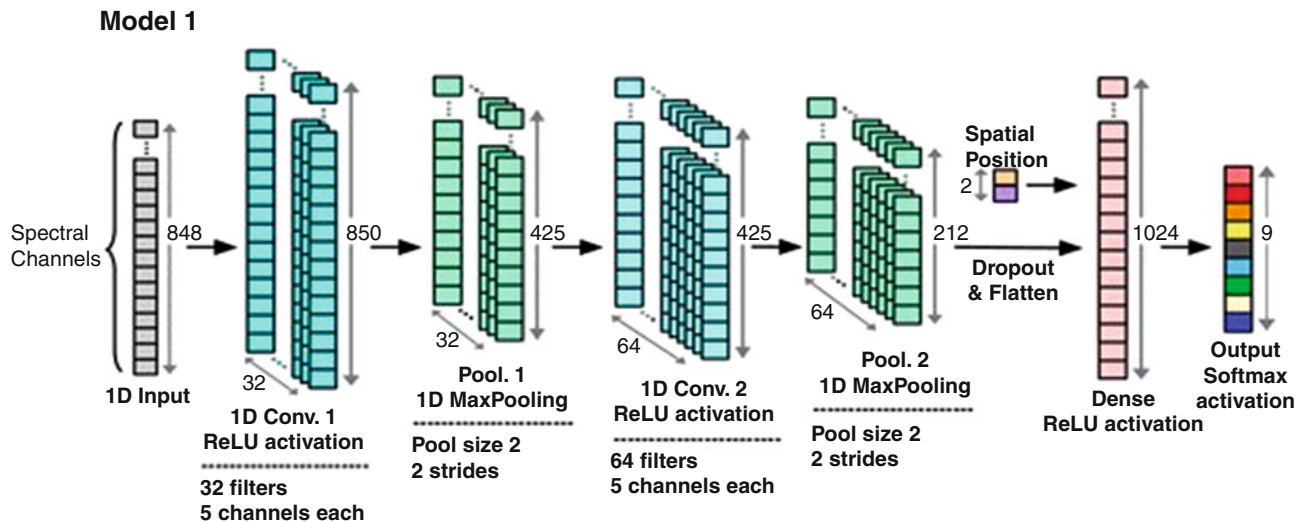
Machine learning algorithms have emerged as useful tools in geosciences that provide researchers with openings for



Sequence Classification, Fig. 1 Recurrent neural network (RNN) and its unfold version



Sequence Classification, Fig. 2 Layers in the classical network



Sequence Classification, Fig. 3 Architecture of CNN model for classifying hyperspectral data (Qamar and Dobler 2020)

improved prediction, interpretation, anomaly detection, event classification, and hence contribute to better decision-making. Growing technology facilitates these methodologies by further offering fast computation and better storage for large data. The models demonstrated in this chapter are some of the machine learning applications that are extensively valid in geological investigations where the data can be represented as a series of events that are happened in order.

Cross-References

- Cluster Analysis and Classification
- Data Mining
- Deep Learning in Geoscience
- Machine Learning
- Mathematical Geosciences
- Pattern Classification
- Pattern Recognition

Bibliography

- Bhattacharya S, Carr TR, Pal M (2016) Comparison of supervised and unsupervised approaches for mudstone lithofacies classification: case studies from the Bakken and Mahantango-Marcellus Shale, USA. *J Nat Gas Sci Eng* 33:1119–1133
- Cárdenas-Peña D, Orozco-Alzate M, Castellanos-Dominguez G (2013) Selection of time-variant features for earthquake classification at the Nevado-del-Ruiz volcano. *Comput Geosci* 51:293–304. ISSN 0098-3004. <https://doi.org/10.1016/j.cageo.2012.08.012>
- Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning*, December 2014
- Cracknell MJ, Reading AM (2013) The upside of uncertainty: identification of lithology contact zones from airborne geophysics and satellite data using random forests and support vector machines. *Geophysics* 78:WB113–WB126
- Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18(5–6):602–610. ISSN 0893-6080. <https://doi.org/10.1016/j.neunet.2005.06.042>
- Gudmundsson S, Runarsson TP, Sigurdsson S (2008) Support vector machines and dynamic time warping for time series. In: *IEEE International Joint Conference on Neural Networks, IJCNN 2008. IEEE World Congress on Computational Intelligence. IEEE*, pp 2772–2776
- Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9:1735–1780
- Lindberg DV, Grana D (2015) Petro-Elastic Log-Facies Classification Using the Expectation–Maximization Algorithm and Hidden Markov Models. *Math Geosci* 47:719–752. <https://doi.org/10.1007/s11004-015-9604-z>
- Puzyrev V, Swidinsky A (2021) Inversion of 1D frequency- and time-domain electromagnetic data with convolutional neural networks. *Comput Geosci* 149:104681. ISSN 0098-3004. <https://doi.org/10.1016/j.cageo.2020.104681>
- Qamar F, Dobler G (2020) Pixel-wise classification of high-resolution ground-based urban hyperspectral images with convolutional neural networks. *Remote Sens* 12:2540. <https://doi.org/10.3390/rs12162540>
- Qiu Q, Xie Z, Wu L, Li W (2018) DGeoSegmenter: a dictionary-based Chinese word segmenter for the geoscience domain. *Comput Geosci* 121:1–11. ISSN 0098-3004. <https://doi.org/10.1016/j.cageo.2018.08.006>
- Song S, Hou J, Dou L, Song Z, Sun S (2020) Geologist-level wireline log shape identification with recurrent neural networks. *Comput Geosci* 134:104313
- Tang D, Qin B, Feng X, Liu T (2016) Effective LSTMs for Target-Dependent Sentiment Classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307, Osaka, Japan. The COLING 2016 Organizing Committee.
- Wei X, Yu X, Liu B, Zhi L (2019) Convolutional neural networks and local binary patterns for hyperspectral image classification. *Eur J Remote Sens* 52(1):448–462
- Xing Z et al (2010) A brief survey on sequence classification. *SIGKDD Explor* 12:40–48
- Xing W, Qian Y, Guan X, Yang T, Wu H (2020) A novel cellular automata model integrated with deep learning for dynamic spatio-temporal land use change simulation. *Comput Geosci* 137:104430. ISSN 0098-3004. <https://doi.org/10.1016/j.cageo.2020.104430>

Sequential Gaussian Simulation

J. Jaime Gómez-Hernández

Research Institute of Water and Environmental Engineering
Universitat Politècnica de València, Valencia, Spain

Definition

Sequential Gaussian simulation is a computer-based technique for the generation of realizations $z(\mathbf{x})$ from a multi-Gaussian random function $Z(\mathbf{x})$ defined on a finite point set D , generally discretizing into N voxels a one-, two-, or three-dimensional area of interest. There are many other techniques capable of generating realizations from a multi-Gaussian random function, but what makes unique the sequential Gaussian simulation algorithm is the possibility of generating realizations with a reasonable computing effort over arbitrarily large domains with N easily reaching values larger than millions. This is achieved thanks to the sequential simulation principle derived and implemented by André Journel's group at Stanford University in the late 1980s (Gómez-Hernández and Srivastava 2021).

Random Function

A random function $Z(\mathbf{x})$ can be defined as a rule that assigns a realization to the outcome of an experiment

$$Z(\mathbf{x}) \sim \{z(\mathbf{x}, \theta)\}, \forall \theta \in \Omega, \mathbf{x} \in \mathcal{D}. \quad (1)$$

In the case of numerical simulations in discretized domains, D is defined as the set of N points falling at the

centers of the voxels discretizing the area of study; $\mathbf{x}$ is any such point, θ is the outcome of an experiment, and Ω is the sample space containing all possible outcomes. In the remainder, the dependency of any realization on θ will be dropped and lower capital letters will be used to refer to a realization or the values it could take, and upper capital letters will be used for random variables or random functions.

A random function can also be defined by all the n -variate distributions

$$F(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n; z_1, z_2, \dots, z_n) = \text{Prob}(Z(\mathbf{x}_1) \leq z_1, Z(\mathbf{x}_2) \leq z_2, \dots, Z(\mathbf{x}_n) \leq z_n), \quad (2)$$

for any subset of n points out of the N points discretizing the study area. In the previous expression, $Z(\mathbf{x}_i)$ is a random variable defined at $\mathbf{x}_i$ through the ensemble of realizations.

When the random function is multi-Gaussian, each random variable can take any value in $\Re$ and the distributions in Eq. (2) have the following expression (Anderson 1984)

$$F(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n; z_1, z_2, \dots, z_n) = (2\pi)^{-\frac{n}{2}} \left| \Sigma \right|^{\frac{1}{2}} \int_{-\infty}^{z_1} \int_{-\infty}^{z_2} \dots \int_{-\infty}^{z_n} e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{z}-\boldsymbol{\mu})} d\mathbf{z}, \quad (3)$$

where $\boldsymbol{\mu}$ is a column vector with the expected values of the random variables

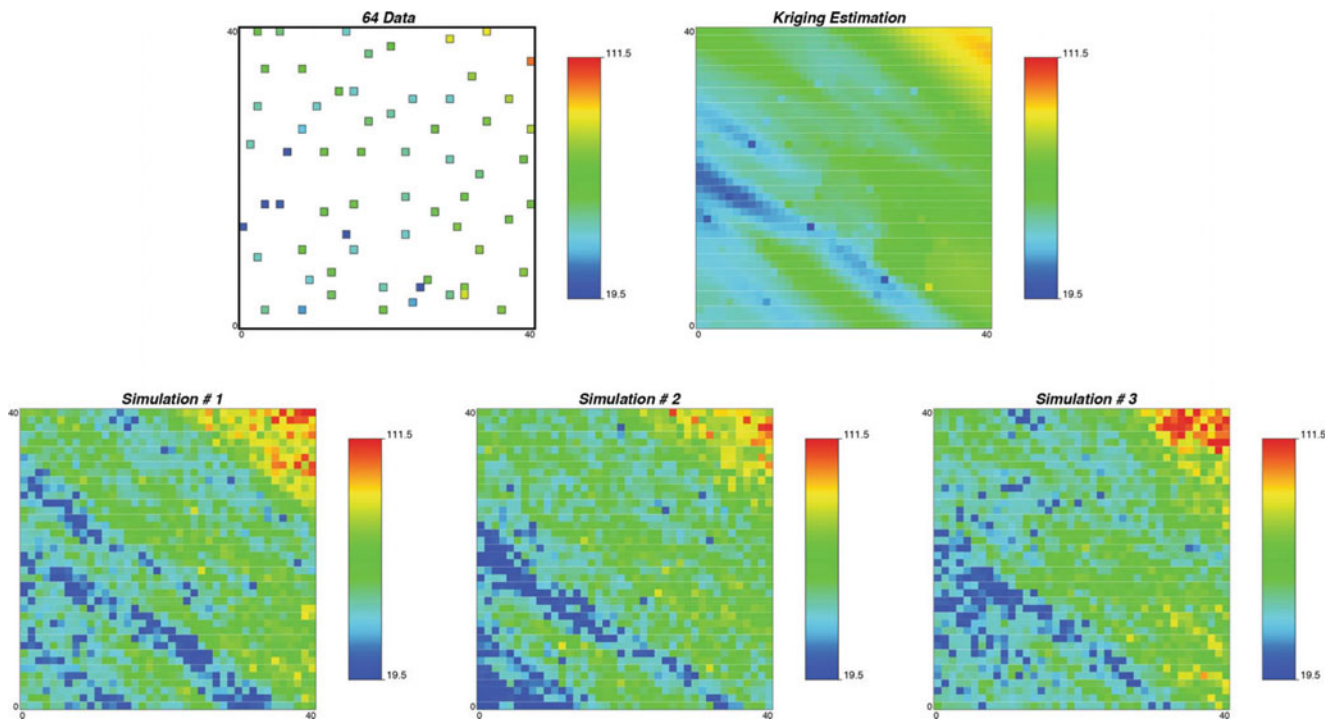
$$\boldsymbol{\mu} = \begin{Bmatrix} E\{Z(\mathbf{x}_1)\} \\ E\{Z(\mathbf{x}_2)\} \\ \dots \\ E\{Z(\mathbf{x}_n)\} \end{Bmatrix}, \quad (4)$$

$\mathbf{z}$ is an n -component column vector with the integration variables, and Σ is a matrix of covariances

$$\Sigma = \begin{pmatrix} C(Z(\mathbf{x}_1), Z(\mathbf{x}_1)) & C(Z(\mathbf{x}_1), Z(\mathbf{x}_2)) & \dots & C(Z(\mathbf{x}_1), Z(\mathbf{x}_n)) \\ C(Z(\mathbf{x}_2), Z(\mathbf{x}_1)) & C(Z(\mathbf{x}_2), Z(\mathbf{x}_2)) & \dots & C(Z(\mathbf{x}_2), Z(\mathbf{x}_n)) \\ \vdots & \vdots & \ddots & \vdots \\ C(Z(\mathbf{x}_n), Z(\mathbf{x}_1)) & C(Z(\mathbf{x}_n), Z(\mathbf{x}_2)) & \dots & C(Z(\mathbf{x}_n), Z(\mathbf{x}_n)) \end{pmatrix}. \quad (5)$$

The attractiveness of the multi-Gaussian distribution is that it is fully characterized by its expected value and its covariance between any two points in its domain. At the same time, this is its major drawback since such a characterization prevents any control on the higher-order moments, which are fully determined as functions of the mean and covariance. Once a multi-Gaussian random function is chosen, the user

cannot expect higher-order moments different from the ones to be derived from the distributions in Eq. (3). For instance, multi-Gaussianity will always prevent high continuity of extreme values, a property much desirable in the geosciences when modeling fracture zones or impermeable shale barriers.



Sequential Gaussian Simulation, Fig. 1 Estimation versus simulation. Top row, sample data and estimated map obtained by kriging. Bottom row, three realizations. All maps conditioned to the sample data and corresponding to the same random function

Estimation Versus Simulation

The behavior of a random function can be analyzed by looking at the average value of all the realizations that compose it or by looking at an ensemble of individual realizations.

When talking about estimation, the stack of realizations that make up the random function is collapsed into a single spatial function obtained by averaging, point by point, all the realizations. Equivalently, this would be the same as computing the expected value of each random variable, and this is what kriging does. Such a representation is locally optimal in some sense (Journal 1989), meaning that, focusing on a single location and computing the deviation of the expected value from all possible values in the ensemble of realizations, the mean deviation would be zero and its variance would be minimal. This does not mean that the spatial representation is realistic; in fact, the mean value will not coincide with any specific realization, but rather it will display lower variance and a more continuous, smoother spatial variability than any specific realization. Figure 1 shows, on the top row, a set of 64 data and the resulting estimated map obtained by kriging.

While estimation seeks a locally optimal estimate of the attribute, simulation seeks to generate alternative realizations from the random function. These realizations have specific patterns of variability and continuity. Figure 1 shows, on the bottom row, three realizations from the ensemble of realizations that make up the random function, the average of which

is the kriging map on the top row. It is evident that there is more variability in any of the realizations than in the estimated map both in terms of variance and in terms of spatial correlation. In fact, any of the realizations could be the reality that stands behind the 64 sample data.

When should estimation be used instead of simulation or vice versa? The answer is simple, if the resulting map is going to be used to compute some kind of attribute that is a linear combination of the map values, a kriging estimate is good enough. This is the case when computing the total reserves in a mine, and the random function is used to model ore grades. However, if the map is going to be used to compute a response variable that depends non-linearly on the attribute, as it would be the case of computing the travel time of a contaminant traversing an aquifer in which the property mapped is hydraulic conductivity: the time that the contaminant would take to cross the aquifer on the estimated map of Fig. 1 will be very different from the time computed as the average of the times that the contaminant takes to cross each one of the realizations in the ensemble.

Sequential Simulation

The sequential simulation algorithm was proposed by André Journé in the late 1980s to address the problem of how to generate a realization from a non-Gaussian random function

defined on the basis of indicator functions (functions that take a value of 0 or 1 depending if the argument is above or below a given threshold). The first publicly available code was published by Gómez-Hernández and Srivastava (1990), and it describes the principles of the algorithm and many implementation issues. Soon after, it was realized that sequential simulation could also be applied for the generation of realizations from other random functions, particularly, from multi-Gaussian random functions, and sequential Gaussian simulation was born (Deutsch and Journel 1992; Gómez-Hernández and Journel 1993).

The basic idea behind sequential simulation is quite simple (Gómez-Hernández and Cassiraga 1994). Given a random function defined over a finite point set of size N , the probability distribution in Eq. (2) can be decomposed, by recursively applying the definition of conditional probability, as follows:

$$\begin{aligned} F(Z(\mathbf{x}_1), Z(\mathbf{x}_2), \dots, Z(\mathbf{x}_N)) &= F(Z(\mathbf{x}_1)) \cdot \\ &F(Z(\mathbf{x}_2)|Z(\mathbf{x}_1)) \cdot \\ &F(Z(\mathbf{x}_3)|Z(\mathbf{x}_1), Z(\mathbf{x}_2)) \cdot \dots \\ &F(Z(\mathbf{x}_N)|Z(\mathbf{x}_1), \dots, Z(\mathbf{x}_{N-1})). \end{aligned} \quad (6)$$

Using this decomposition, drawing a realization from a full N -variate distribution can be replaced by drawing, sequentially, N values from N conditional univariate distributions. The only requirement would be the ability to compute the conditional distribution of any random variable given any number of random variables, which, for the case of a multi-Gaussian random function, is trivial.

The conditional distribution of a random variable given any number of random variables when these random variables are members of a multi-Gaussian random function is a Gaussian distribution with mean and covariance given by the solution of a set of normal equations (Anderson 1984), also known as the simple kriging equations. Its expression is

$$\begin{aligned} \text{Prob}(Z(\mathbf{x}_i) \leq z_i | Z(\mathbf{x}_1) = z_1, Z(\mathbf{x}_2) = z_2, \dots, Z(\mathbf{x}_n) = z_n) = \\ \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{z_i} e^{-\frac{(z-\mu)^2}{2\sigma^2}} dz \end{aligned} \quad (7)$$

with μ being the estimate of $z(\mathbf{x}_i)$ using as data $\{z_1, z_2, \dots, z_n\}$ at locations $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, respectively, and σ^2 the corresponding simple kriging variance.

The conditional distribution in Eq. (7) is generally approximated, when the number n is large, by restricting the data used to solve the kriging equations to a predefined maximum number within a given search neighborhood around $\mathbf{x}_i$. Only the subset of $z_1, z_2, \dots, z_n$ that falls within the neighborhood and only the closest values up to a certain maximum number

are used to build the simple kriging system. This approximation is necessary since, for a realization with N larger than a few hundreds, the decomposition in Eq. (6) results in conditional distributions depending in too many conditioning data; determining the parameters of those conditional distributions would amount to solving kriging systems with hundreds of equations; in order to keep the computation of the conditional realizations – that has to be repeated for each point – at a reasonable computing cost, there is a need to reduce the size of the kriging system, which can be achieved using the concept of a search neighborhood and limiting the maximum number of points.

Sequential Gaussian Simulation Algorithm

Drawing a realization from a multi-Gaussian random function can be achieved as follows:

1. Define a random permutation of the natural numbers 1 to N that will serve to define how expression (6) is built and also the order in which the points will be visited.
2. Following the order in the random permutation, visit each point and
 - (a) Search for the values that have already been simulated falling within a search neighborhood centered at the point to simulated.
 - (b) Solve the simple kriging equations to determine the mean and the variance of the Gaussian conditional distribution.
 - (c) Draw a random number from the conditional distribution and assign it to the point.
3. Return to the beginning for the generation of a new realization.

Remarks

Equation (6) is valid for any permutation of numbers 1 to N ; a random permutation is suggested because it was found that the approximation of the conditional distribution retaining only the closest simulated points introduces some artifacts in the realizations when a structured sequence is used (for instance, simulating the points following the rows or columns of a two-dimensional realization).

Conditional realizations, that is, realizations that honor the observed data at data locations can be generated simply by including their locations in the domain D and by including the observed values as known (already simulated) values at their locations prior to the start of any simulation.

There are other implementation issues, many of which can be consulted in the review paper by Gómez-Hernández and Srivastava (2021). Some of them try to speed up the simulation time, for instance, simulating not one point at a time, but a whole bunch of nearby points at once, or relocating all

observed data and simulation locations over a regular grid that would permit the precomputation of the covariance values that will be repeatedly used to build the kriging systems. Others try to solve conceptual problems, like when the conditioning data do not have a Gaussian histogram, in which case the data should be transformed into normal-scores – where they will have a standardized Gaussian histogram–, sequential Gaussian simulation will be carried out in normal-score space, and the results will be back-transformed to the original space.

Summary and Conclusions

Sequential Gaussian simulation is a stochastic simulation technique to draw realizations from a multi-Gaussian random function. It is a specific implementation of the more general sequential simulation algorithm that could generate realizations from any random function for which the conditional distribution of any random variable given any number of random variables can be determined or approximated. The attractiveness of the multi-Gaussian random function is that it is characterized by a mean and a covariance, parameters that can be decided upon an exploratory data analysis.

Cross-References

- [Journel, André](#)
- [Kriging](#)
- [Monte Carlo Method](#)
- [Multivariate Analysis](#)
- [Random Function](#)
- [Random Variable](#)
- [Realizations](#)
- [Simulation](#)
- [Spatial Statistics](#)
- [Stochastic Geometry in the Geosciences](#)

References

- Anderson TW (1984) Multivariate statistical analysis. Wiley, New York
 Deutsch CV, Journel AG (1992) GSLIB, Geostatistical Software Library and User's Guide. Oxford University Press, New York
 Gómez-Hernández JJ, Cassiraga EF (1994) Theory and practice of sequential simulation. In: Armstrong M, Dowd P (eds) Geostatistical simulations. Kluwer Academic Publishers, pp 111–124
 Gómez-Hernández JJ, Journel AG (1993) Joint sequential simulation of Multi-Gaussian fields. In: Soares A (ed) Geostatistics Tróia '92, Kluwer Academic Publishers, Dordrecht, vol 1, pp 85–94
 Gómez-Hernández JJ, Srivastava RM (1990) ISIM3D: an ANSI-C three dimensional multiple indicator conditional simulation program. Comput Geosci 16(4):395–440

- Gómez-Hernández JJ, Srivastava RM (2021) One step at a time: the origins of sequential simulation and beyond. Math Geosci 53 (2):193–209. <https://doi.org/10.1007/s11004-021-09926-0>
 Journel AG (1989) Fundamentals of geostatistics in five lessons, Short courses in Geology, vol 8. Agu, Washington D.C

Serra, Jean

Philippe Salembier
 Department of Signal Theory and Communication,
 Universitat Politècnica de Catalunya, BarcelonaTech,
 Barcelona, Spain



Fig. 1 Jean Serra, 1970 Leningrad. Courtesy of Jean Serra



Fig. 2 Jean Serra, 1994 Barcelona. Courtesy of Jean Serra

Biography

Jean Serra was born in Algeria in 1940, where he lived until 1962. After high school, he studied in France at Nancy School of Mines and received an engineering degree in 1962. He also got a Bachelor degree in philosophy/psychology in 1965 from the University of Nancy.

From 1962 to 1966, while a research engineer at the Iron and steel research institute in France, Serra undertook a Ph.D. under the supervision of Georges Matheron, about geostatistics of iron ore deposit of Lorraine. During that period, Serra came up with the idea of using structuring elements for transforming images of thin cross sections of the ore in order to extract information. This work led to a device called “Texture Analyser,” and the hit-or-miss transformation, which was extended by Matheron to the concepts of erosion, dilation, opening and closing, and granulometry. In 1966, Matheron and Serra decided to give a name to the approach they were developing: “Mathematical morphology.” Jean Serra obtained a Ph.D. in Mathematical Geology from the University of Nancy in 1967, and a Doctorat d’état in Mathematics, from the University Paris-VI in 1986.

In 1968, the Centre of Mathematical Morphology of the École des Mines de Paris was created, combining geostatistics and mathematical morphology (after 5 years, the two groups became independent). The center itself was localized at Fontainebleau, France. Matheron was appointed director, and Serra was hired as master of research (“Maître de Recherches”) and assistant director. This was an important anchoring point. Very soon, the chaining of successive hit or miss transforms were shown to create new and interesting operations. In parallel, the German company Leitz commercialized the texture analyzer. The 1970s correspond to the period where many classical Mathematical Morphology tools were created, including in particular: binary thinning, skeleton by influence zones, ultimate erosion, particle analysis, conditional bisector, morphological gradient, top hat transform and watershed. In 1971, Serra took a sabbatical year at the Lomonosov University, Moscow.

During the 1970s and the 1980s, Jean Serra actively participated to the dissemination of Mathematical Morphology in Europe, the USA, and Australia. At the scientific level, he worked on a new formulation of Mathematical Morphology relying on the framework of complete lattices. In 1986 he officially became the director of a new Centre of Mathematical Morphology, separated from that of Geostatistics.

At the beginning of the 1990s, Jean Serra spent a sabbatical year in Barcelona, Spain, and started to work on multimedia applications. In 1993 he founded the International Symposium for Mathematical Morphology (ISMM). Since this time,

12 other symposia were held. Important contributions related to this period are connections, connected filters, and segmentation. After his retirement, he has been very active in terms of research as a professor Emeritus of ESIEE, Paris.

Beside his scientific activity, Jean Serra is also a musician. He participated to the Russian Liturgical Choir of the Holy Trinity Church, Paris, and is organist for Saint Peter Church of Avon, since 1988. He speaks French, Russian, English, and Spanish.

Shape

Norman MacLeod

Department of Earth Sciences and Engineering, Nanjing University, Nanjing, Jiangsu, China

Definition

Shape – the geometric form of an object’s boundary, outline, or surface independent of its size and distinct from other properties such as material, color, luster, and/or texture.

Introduction

When earth scientists think or speak of shape, they usually reference one of two very different sets of concepts, perceptions and traditions which also differ in important ways from the mathematician’s or geometer’s views on shape. In common speech shape is most often taken to mean the degree to which the outline of an object conforms to the expected pattern of one or more reference shapes (e.g., a circle, triangle, or square in the case of two-dimensional [2D] shapes, a sphere, prism, or cube in the case of three-dimensional [3D] shapes). This analogistic method of describing shape works well so long as the shape being described is regular and matches the expected geometry of the reference shape closely and so long as an appropriate reference shape exists. Unfortunately, many shapes of interest to earth scientists fail to conform to these simple criteria and so remain difficult to characterize or describe using appeals to analogy.

Perhaps the first thing to note is that the concept of shape is distinguished routinely from the concept of size. A reference shape can adopt a wide variety of sizes, but this variation in no way compromises its utility as a basis for the description of simple shapes. In formal shape analysis, the description or measurement of size is almost always distinguished explicitly from the description or measurement of shape. Indeed, one

popular formal definition of shape is those aspects of a form that remain after position, size, and orientation have been removed from consideration (Kendall, 1984). While pragmatic and practical from a computational point-of-view, this widely cited definition suffers from the obvious logical problems that arise from adopting a definition predicated exclusively on what the concept being defined is not. Kendall's definition also assumes agreement can be reached on how to describe, measure, and/or characterize an object's position, size, and orientation.

Those concerned with the analysis of inorganic particle shapes, such as those of sediment grains and crystals, have tended to follow the recommendations of Wentworth (1922) and Sneed and Folk (1958) who considered shape to be a "tri-dimensional characteristic" via reference to its three, orthogonal, linear dimensions: length, breadth (or width), and thickness. Those concerned with analyses of the shapes of organic bodies, such as animals, plants, and/or the fossilized remains thereof, began by adopting the methods of particle analysis to describe shapes. In recent years, though, these practitioners have tended to follow the recommendations of Blackith and Reyment (1971), who regarded form as arising from the combined effects of size and shape and represented both via sets of linear distances between comparable, or landmark, point locations that could be found on all specimens in a sample. These linear distances did not necessarily need to be confined to the specimen's gross dimensions, and their orientation did not need to respect the constraint of orthogonality. Later Bookstein (1991) refined this "morphometric" approach to form analysis by pointing out that the geometric character of the landmarks used previously to define the endpoints of the scalar linear distances could be preserved in a multivariate morphometric-style analysis by using the 2D or 3D point coordinate locations (= paired or triplet scalar linear distances from the origin of a common coordinate system) as direct input into data analysis procedures, usually after a Procrustes-style standardization had been performed to remove patterns of variance that could be attributed to position, size, and orientation from the point coordinate values. This extension of the multivariate morphometric approach is commonly referred to as "geometric morphometrics" in order to distinguish it from its linear distance-based "multivariate morphometrics" precursor.

Naturally, the main distinguishing feature of the particle analysis and morphometric approaches to the characterization of shape revolved around the perceived absence or presence of topographic point locations that could form a logical basis for the comparison of different shapes. Sedimentary and crystal grains have no such privileged and specific point locations that constitute intrinsic aspects of their forms. As a consequence, maximal or Feret distances (Feret, 1930) defined by the length of hypothetical enclosing

2D or 3D boxes (as measured with rulers or calipers) were used to quantify their shapes. But biological forms can and do exhibit such privileged and specific point locations, often in abundance in the case of morphologically similar species. To many biologically trained shape analysts it seemed self-defeating to ignore the implications of the existence of such point locations and forego the opportunity to employ them as a basis for shape comparison irrespective of how this topological comparability arose (e.g., phylogenetic descent, functional convergence). The existence of these privileged landmark point locations in forms of biological interest also accounts for the suspicion – shared by many biological morphometricians – that use of ratio-based shape characterization indices, calculated from linear distance data, amounts to the unwarranted discarding of aspects of information relevant to geometric shape description and analysis in addition to the representational ambiguities such an operation can entail, especially if the numerator and denominator variables are themselves correlated. Adherents to the tradition of sedimentary particle shape analysis have far fewer scruples when it comes to the use of ratio data. Nonetheless, both traditions, or approaches, to shape description/analysis accept data derived from the 2D and 3D outlines of the forms under consideration though, in the case of geometric morphometrics, not without considerable early resistance.

Mathematical Aspects of Shape

To mathematicians, shape is the external boundary, outline, or surface of any closed-form object as opposed to its color, texture, or composition. Mathematical points have positions, but no shape. Mathematical lines and vectors have position, orientation, and magnitude, but no shape. The simplest 2D Euclidean object that has a shape is a triangle (= three-sided polygon). The simplest 3D Euclidean object that has a shape is a pyramid with a triangular base. Both regular and irregular polygons have the property of shape though, in the case of the latter, this can be difficult to describe. Complex, self-intersecting polygons and pyramids exist as mathematical concepts, but cannot exist as real objects in normal two- or three-dimensional Euclidean spaces. According to Kendall's (1984) definition though, these forms could be thought of as having the property of shape. Circles, ellipses, their 3D counterparts, and any form that includes a curve, are not polygons but do have the property of shape.

One shape may be transformed into another shape by sequences of rotations, translations, and/or reflections. These transformations may be applied isotropically (uniformly in all directions) or anisotropically (non-uniformly in direction). Depending on the character of their transformation sequence,

shapes may be said to be congruent, similar, isotropic, or anisotropic. The mirror transformation of a non-isosceles shape is not usually regarded as being congruent with the original shape as there is no way to devise a sequence of translations or less than 360° rotations that will produce a perfect superposition.

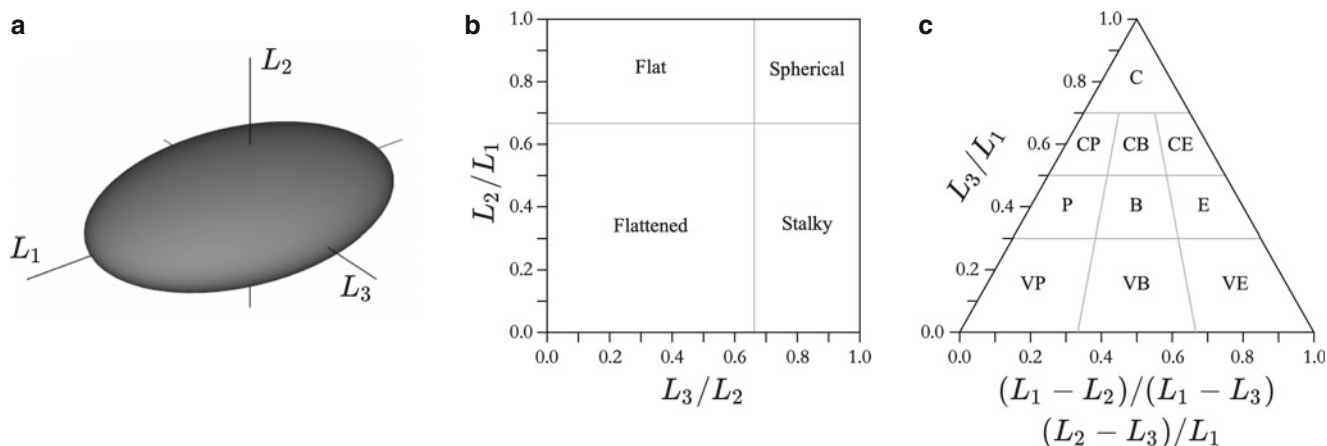
Form Ratios, Form Factors, and Particle Shape Spaces

By the 1920s, it was realized that simple ratios of length variables could be used to define broad categories of particle shapes more objectively and repeatably than text-based descriptions. Wentworth (1922) and Zingg (1935) based their calculations on three objective distance measurements: the particle's longest dimension (L_1), longest dimension perpendicular to L_1 (L_2), and longest dimension perpendicular to both L_1 and L_2 (L_3) (Fig. 1a). These definitions suffice for the majority of particles, especially if the (artificial) requirement of all three distances intersecting at a single point is relaxed. Still, their implementation can be ambiguous or counterintuitive for highly symmetrical (e.g., perfect squares or cubes) and very irregular shapes.

Wentworth (1922) used the relation $(L_1 + L_2)/2 L_3$ to express flatness or equancy and r_1/R (where r_1 = the radius of curvature of sharpest edge and R = the mean radius) to express the roundness of beach pebbles collected from two localities on the US east coast. These indices were used to document an inverse relation between pebble roundness and pebble flatness in what was likely the first quantitative particle shape analysis published in the earth sciences. Wentworth's (1922) data were also used to demonstrate that pebble abrasion was linked directly to the intensity of wave action.

Wentworth's (1922) approach to shape description was later expanded by Zingg (1935) in the form of two ratio-based shape indices $i_1 = L_1/L_2$ and $i_2 = L_3/L_2$. These ratios capture aspects of an object's flatness and elongation, respectively. Zingg (1935) used these ratios as the axes of a Euclidean coordinate system within which disk-shaped ($i_1 < 0.667$, $i_2 \geq 0.667$), spherical ($i_1 \geq 0.667$, $i_2 \geq 0.667$), bladed ($i_1 < 0.667$, $i_2 < 0.667$), and rod-like ($i_1 \geq 0.667$, $i_2 < 0.667$) morphotypic shape categories could be defined and portrayed graphically (Fig. 1b). Sneed and Folk (1958) objected to the fact that Zingg's (1935) system had so few shape categories and was composed of unequal ratio ranges. As an alternative, they proposed a shape modeling system based on a ternary chart in which the horizontal and vertical axes represented the ratios L_2/L_1 and $(L_2-L_3)/(L_2-L_1)$ (Fig. 1c). This triangular space was then subdivided into ten morphotypic shape categories using equally spaced boundaries. Later Hockey (1970) showed that Sneed and Folk's (1958) shape space was isomorphic with a ternary shape space in which the vertical and horizontal axes were calculated as L_2/L_1 and $(L_2-L_3)/L_1$, respectively (Fig. 1c).

Both the Zingg (1935) and Sneed and Folk (1958) particle shape spaces and shape categories are still used and each has its adherents who criticize use of the opposite technique because it introduces "distortions," in the form of density heterogeneities, into the resulting shape identifications. This debate was resolved by Blott and Pye (2008) who used sets of computer-generated random blocks to show that both spaces are prone to density-based distortions at the extreme elongate ends of their respective shape spectra. Accordingly, particle shape analysts need to keep these aspects of the Zingg (1935) and Sneed and Folk (1958) shape spaces in mind when using



Shape, Fig. 1 Particle analysis axes, shape factors, and shape spaces. (a). Tridimensional lengths used originally by Wentworth (1922) to quantify particle shape. These lengths are equivalent to the Feret diameters that are used to characterize more irregular shapes. (b). Particle-axes length ratios, the mathematical space, and the particle shape

categories proposed by Zingg (1935). (c). The triangular particle shape space and shape categories proposed by Sneed and Folk (1958) to characterize particle shapes: C (compact), CP (compact-platy), CB (compact-bladed), CE (compact-elongate), P (platy), B (bladed), E (elongate), VP (very platy), VB (very bladed), VE (very elongate)

them to test statistical hypotheses concerning particle shape distributions and/or shape-group distinctions.

Wentworth's (1922) roundness shape index has also been subject to debate and revision over time, especially since the sharpest edge of a particle or grain might not be representative of the degrees of roundness exhibited by all edges along the particle or grain's periphery. In 1932, Wadell proposed the definition of Wentworth's r_1 index be modified to represent the average radius of all the corners of the particle or grain's projected 2D outline using one of two alternative formulae.

$$P_{d1} = \frac{\left(\frac{\sum_{i=1}^n D_i}{n} \right)}{D_r} \quad (1)$$

$$P_{d2} = \frac{n}{\sum_{i=1}^n \frac{D_i}{D_r}} \quad (2)$$

where:

n = number of corners

D_i = diameter of curvature of corner i

D_r = diameter of largest inscribed circle.

In order to facilitate the collection of corner curvature data across a particle's entire outline, Wadell developed a transparent scale of concentric circles that could be placed over a magnified representation of the outline. Subsequently many authors proposed class definitions for a variety of shape categories based on Wadell's (1932) index.

As noted by Blott and Pye (2008), angularity is an aspect of particle shape commonly thought to be the inverse of roundness. Geometrically though, angularity constitutes a distinct geometric concept based on the angles formed by planes bounding corners along the grain's outline or surface rather than the outline or surface's localized curvature. Particle angularity indices have been proposed, but these have proven difficult to implement owing to scale-related factors. Mandelbrot's concept of the fractal dimension represents an attractive alternative to the difficulties that arise in attempting to implement particle angularity measures, but suffers itself from a lack of predictable geometric specificity as well as the absence of a readily interpretable classification scheme for particle shapes that exhibit defined fractal dimension ranges.

Circularity (2D) and sphericity (3D) are aspects of shape related to, but distinct from, roundness and have usually been assessed by comparing the outline of a grain to that of either an inscribed or mean circle or sphere, quantified (where necessary or appropriate) as a ratio of areas, perimeters, or diameters that express the extent of an outline's or surface's deviation from perfect circularity or sphericity. A large number of indices are available to quantify this deviation.

Finally, just as the concept of roundness focuses attention on a specific category of deviations from circularity, the concept of shape irregularity focuses attention on a different, though related, aspect of geometric deviation. In this case, Blott and Pye (2008) suggested that the magnitude of the deviation from 2D circularity for each local concavity and convexity along the projected outline be quantified and summed across the entire outline according to the following equation:

$$I_{2D} = \sum_{i=1}^n \frac{d_{convex} - d_{concave}}{d_{convex}} \quad (3)$$

where:

n = number of irregularity observations

d_{convex} = distance from the center to the largest inscribed circle to the outline's convex hull in the direction of the i^{th} concavity

$d_{concave}$ = distance from the center to the largest inscribed circle to the nearest point in the direction of the i^{th} concavity

In those cases, where d_{convex} is difficult to assess, an estimate can be obtained by taking an average of the distances between the projections that define the localized aspect of the convex hull according to

$$y = \frac{a \cdot \cos A + b \cdot \cos B}{2} \quad (4)$$

where:

n = number of irregularity observations

a & b = distance from the center of the largest inscribed circle to the tip of the projection on either side of the concavity

A & B = angles between a and x and b and x respectively

The Comparison of Shapes: Shape Deformation Grids

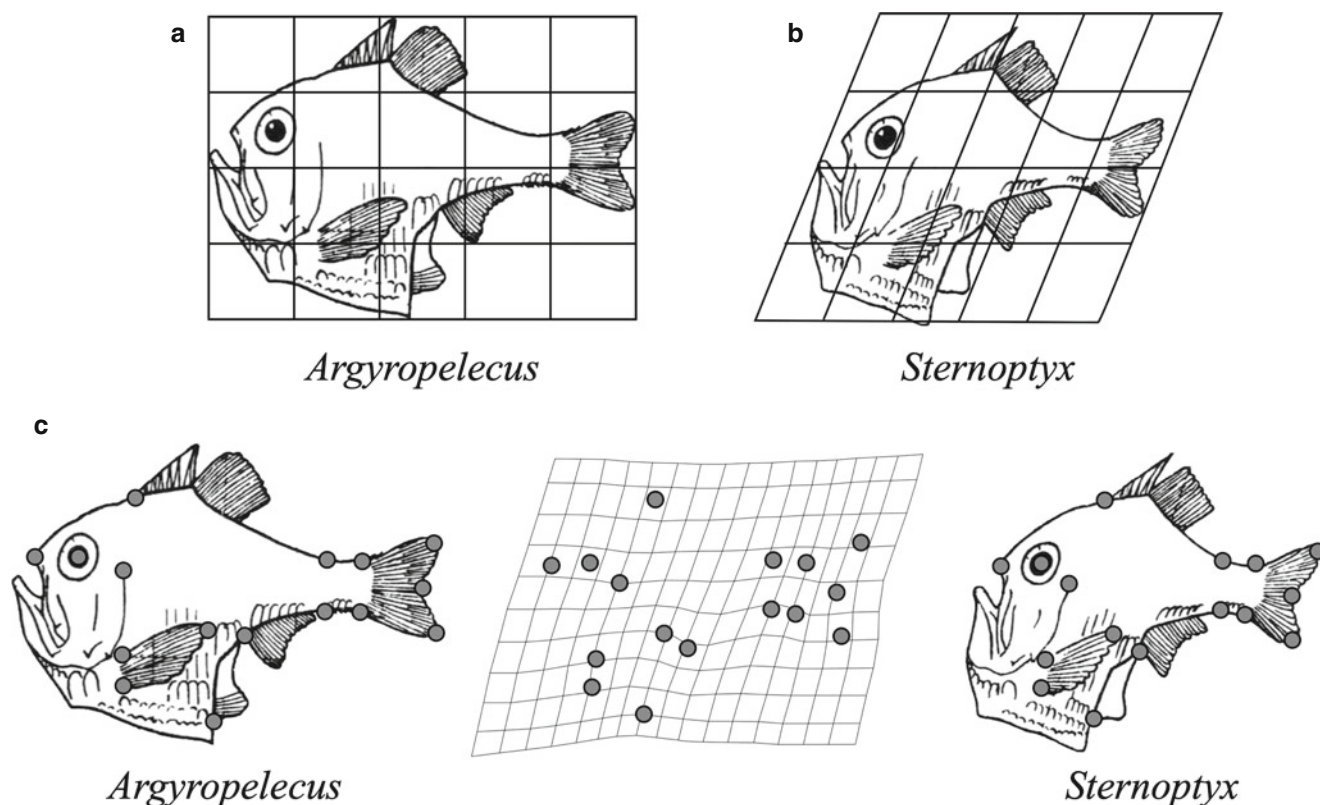
Originally the assessment, description, and analysis of biological shapes were accomplished using many of the same conventions, concepts, and tools that had been developed for the sedimentary particle analysis. A dinosaur cranium, for instance, has a length, width, and depth in the same way a limestone cobble does and form ratios that remove size differences among sets of crania can be calculated just as they can for crystal and grain shapes. But owing to the greater degree of morphological complexity displayed by the different bones of a dinosaur cranium, not to mention their spatial arrangement relative to one another, it is exceedingly difficult

to describe or quantify organic shapes adequately if one restricts oneself to the consideration of only the lengths of three mutually orthogonal axes. Moreover, if the mutual orthogonality of the lengths being used to describe and/or quantify shape cannot be maintained, it becomes exceedingly difficult to recall how the distances being used to describe or quantify the organic shapes in question relate to one another, either in the tables of descriptive characteristics published in the primary taxonomic literature or in the matrices of values analyzed using numerical methods.

In an early attempt to overcome this problem, the Victorian biologist and classics scholar D'Arcy Wentworth Thompson (1917) developed a method – possibly based on Albrecht Dürer's graphical studies of proportional variation in human faces – of representing shape transformations between different vertebrate species by recording the positions of sets of topologically corresponding point locations and summarizing 2D spatial displacements of corresponding point, or landmark, configurations graphically as a smooth deformation of an originally rectilinear grid (Fig. 2a, b). This method of

summarizing shape differences was entirely qualitative and, in a number of cases, not very accurate. But, despite these shortcomings, Thompson's deformation grid approach to the comparison of shapes was highly intriguing in its demonstration that simple sets of smooth deformations could be used to summarize the differences between an astonishing number of radically different shapes. To generations of morphologists, this implied that gaining an understanding of simple variations in the (presumably) small number of geometric variables that controlled the spatial deformation systems responsible for changes in organismal morphology was all that would be needed to understand the perplexing morphological diversity of life.

Thompson was unable to develop mathematical methods that could be used to quantify his deformation-grid approach to the description and comparison of organic shapes. The form ratio and form factor methods developed by the particle and grain analysts were also unsuited to this task because of the limitations imposed by mutual orthogonality and the small number of variables employed in particle shape



Shape, Fig. 2 The characterization of shape change by means of a deformation grid. (a). D'Arcy Thompson's image of the fish genus *Argyropelecus* with a superimposed rectilinear grid. (b). d'Arcy Thompson's image of the fish genus *Sternoptyx* with a sheared and antero-posteriorly compressed version of the *Argyropelecus* grid that appears to bring the major features of the *Sternoptyx* morphology into comparative alignment with those of *Argyropelecus*. Thompson used

this graphic device to argue that even radical morphological differences could be the result of simple geometric transformations. (c). A contemporary mathematical realization of the deformation grid shape transformation between *Argyropelecus* and *Sternoptyx* based on 16 common landmark positions and calculated as a thin-plate spline. Fish images redrawn from Thompson (1917)

characterization. The American quantitative biologist F. L. Bookstein made an early attempt at quantifying Thompson's ideas with his biorthogonal grid technique (Bookstein 1978). Nonetheless, by 1991 a reasonably full, mathematically elegant, and highly useful reconceptualization of what polygonal shapes were from a mathematical point of view, and how they could be represented in a Thompson-esque manner, had been developed, largely as the result of noteworthy contributions by the British statistician David Kendall, Bookstein, and the American statistician Colin Goodall (see Bookstein, 1993).

The modern (re)formation of the Thompsonian shape deformation grid involves calculating a polyharmonic thin plate spline (TPS), which is related computationally to a form of kriging. This mathematical spline attempts to mimic the behavior of a defect-free, uniform, and infinitely thin metal plate that is bent in the z direction to conform to the geometry of the x,y shape displacement vectors at corresponding locations across the reference and target shapes. The metal plate metaphor is important because, unlike elastic surfaces, metal plates bend in ways that minimize the energy required to achieve the bend in all directions over the entire plate. In applying this physical

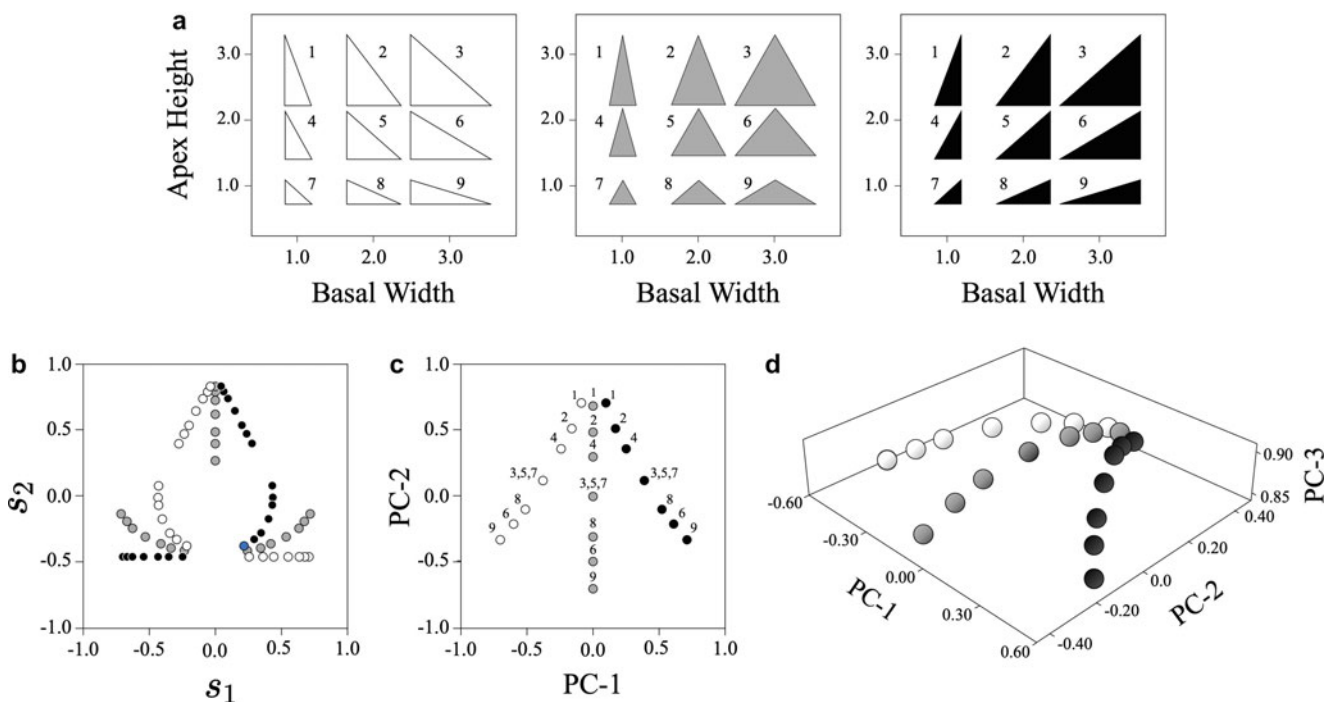
metaphor to shape analysis, the TPS represents the surface of minimal bending energy implied by the transformation of one shape into another.

In the TPS calculation, bending energy is quantified by the following expression:

$$U(r_{ij}) = r_{ij}^2 \ln r_{ij}^2 \quad (5)$$

where r_{ij}^2 is the square of the distance between landmarks i and j in the reference shape's landmark configuration and $\ln$ is the natural logarithm. This calculation quantifies the relative amount of "energy" required to achieve a bend between all pairs of landmarks. The spacing of the landmarks represents an important constraint on the spline because it requires more "energy" to achieve a bend between closely spaced landmarks than between landmarks located at a distance from one another.

All possible modes of bending across a set of landmarks are specified by a partitioned matrix (L) whose structure is summarized as follows:



Shape, Fig. 3 The empirical shape space of triangles. (a). Feret diameter shape spaces for collections of left-verging right triangles (white), isosceles triangles (gray), and right-verging right triangles (black). Note that, although these triangles have different shapes, they plot in exactly the same positions in the particle analysis space. This suggests that each point in the spaces used by sedimentary and particle analyses (e.g., Fig. 1b, c) actually corresponds to families of differently shaped objects. (b). Procrustes shape coordinate plots of the triangle sets shown in a. Note that each of the non-congruent triangles has vertices that plot in

unique positions within this space. (c). The 2D covariance-based PCA space of the shape coordinate data shown in b. Note that families of isosceles, and of right- and left-verging right triangles, plot along consistent trajectories within the PC space. (d). The 3D covariance-based PCA space of the shape coordinate data shown in b. Note that the 2D curved trajectories shown in c exhibit a 3D curvature, as if the triangular shapes they represent are located at unique positions along a spherical surface

$$L = \begin{pmatrix} P & Q \\ Q' & 0 \end{pmatrix} \quad (6)$$

The P matrix partition summarizes the distances between landmarks using the U function (Eq. 5).

$$P = \begin{pmatrix} 0 & U_{12} & U_{13} & \dots & U_{1p} \\ U_{21} & 0 & U_{23} & \dots & U_{2p} \\ U_{31} & U_{32} & 0 & \dots & U_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ U_{p1} & U_{p2} & U_{p3} & \dots & 0 \end{pmatrix} \quad (7)$$

The Q matrix summarizes the coordinates of the reference landmark configuration.

$$Q = \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \\ \vdots & \vdots & \vdots \\ 1 & x_p & y_p \end{pmatrix} \quad (8)$$

Finally, the 0 matrix is a 3×3 matrix of zeros. The overall TPS surface is calculated from the L^{-1} matrix by padding the matrix of x and y shape coordinate values for the target configuration (X_t) out with a 3×2 matrix of zeros (X_{t+}) to give it the same number of rows as the L^{-1} matrix and then multiplying these two matrices together as follows:

$$W = L^{-1}X_{t+} \quad (9)$$

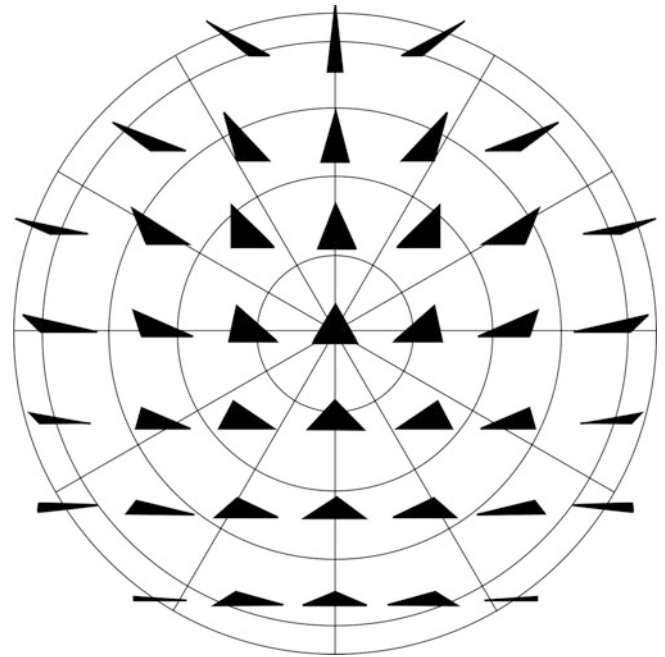
This creates a weight matrix (W) which can be used to calculate the z coordinate of the TPS. An example of the result of this calculation for Thompson's example of the shape transformation from *Argyropelecus* to *Sternoptyx* is shown in Fig. 2c.

Geometric Shape Spaces

With regard to the current mathematical concept of shape and its relation to the concept used in particle shape analysis, this is also best illustrated by way of an example. Figure 3a shows three sets of triangles with identical Feret diameters. All three sets display a variety of different shapes, yet all plot at the same positions in the shape space analogous to those used by particle shape analysts. This result shows clearly how and why the conventions used by particle shape analysts are useful for specifying broad classes of similar shapes but cannot be employed in the detailed characterization of shape when information regarding the topological identity of parts or structures is present.

Of course the shapes of these same triangles can also be specified by recording the x, y coordinates of their vertices. Figure 3b shows the superposition of these vertices in the space of the two Procrustes shape coordinates, s_1 and s_2 . Note the clear distinction between the geometries of the isosceles (gray) and left- (white) and right- (black) verging right triangles. Note also that each set of vertex points contains only seven apparent positions. The reason for this apparent reduction is that triangles along the upward trending diagonals of the triangle sets shown in Fig. 3a all have the same aspect ratio and so have the same shapes but different sizes. Therefore, use of corresponding landmark locations is able to represent polygonal shapes correctly, whereas the use of Feret diameters cannot.

Figures 3c, d show results of a covariance-based principal component analysis (PCA) of these Procrustes-aligned shape coordinate data both in terms of the first two (Fig. 3c) and a complete set of three (Fig. 3d) eigenvectors. Each point in this space represents a unique configuration of the six coordinate values that, together, comprise any triangle. But unlike the width/height space (or a PCA-determined representation



Shape, Fig. 4 The mathematical shape space of triangles. All conceivable polygonal shapes of constant size can be located at a unique position on the surfaces of a $2K-3$ dimensional manifold where K = the number of polygon-defining landmarks. In the case of triangles, the dimensionality of this manifold is 3, so it has the form of a sphere. Only the upper hemisphere (= Procrustes pre-shape space) of the shape manifold is shown here with an equilateral triangle chosen arbitrarily to occupy the pole position. Circles centered on the pole represent lines of latitude spaced 18° apart. The lower hemisphere is occupied by identical triangular shapes whose apices fall below the baseline. The shape spaces' equator is occupied by the set of colinear triangles. (Redrawn from an original by F. J. Rohlf)

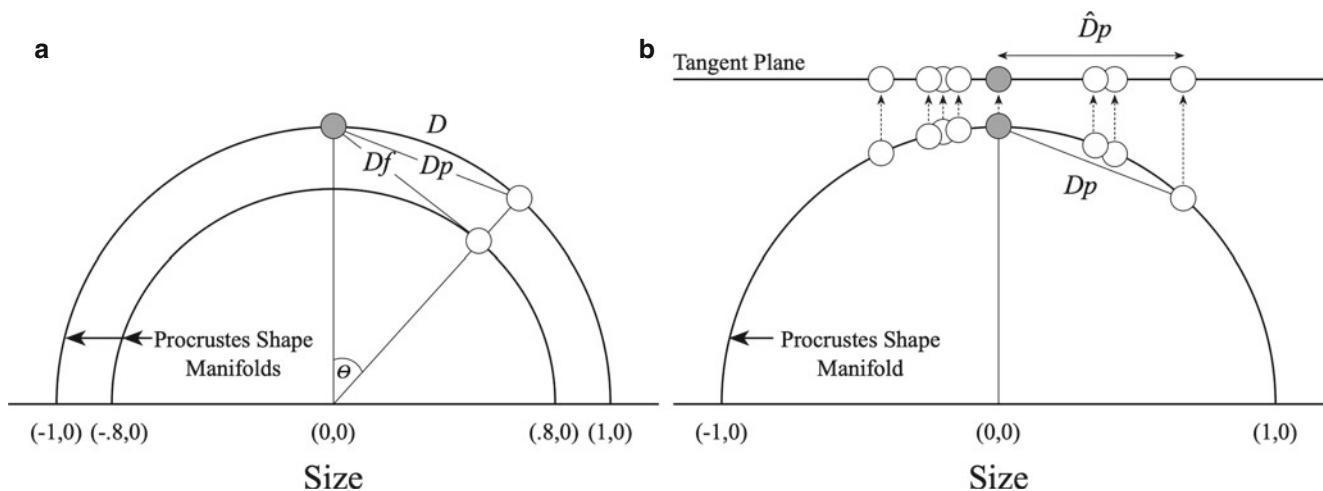
thereof), all but the three congruent triangles in each shape group exhibit unique coordinate locations in the PCA-decomposed Procrustes shape-coordinate space. These unique locations reflect and summarize their unique shapes.

Note also that the equilateral triangles in this series (triangles 3, 5, and 7 in the gray group) plot at the origin of the PC-1 vs. PC-2 coordinate system (Fig. 3c, left) and at the unit coordinate along PC-3. This is the pole position of a unit hemisphere, the surface of which contains all the other triangles – and all conceivable triangles – arrayed in a systematic manner that reflects their intrinsic and geometrically organized shape differences accurately. All triangular configurations of landmark points are located on the surface of this hemisphere because differences in their sizes (along with any differences in position and orientation) have been discarded during the Procrustes alignment procedure.

This simple mathematical experiment illustrates Kendall's (1984) conclusion that all polygonal shapes exist on the surfaces of hyperdimensional mathematical manifolds of dimension $2k-3$ (where k is the number of polygon vertices) and in which the radius of the manifold reflects the size of the vertex-defined polygon. In the case of triangles (which are the simplest geometric objects that have a shape), this manifold is a sphere.

As can be seen in Fig. 4, all conceivable triangles (including colinear triangles) whose sizes have been adjusted to a unit value can be located on the surface of the three-dimensional triangle shape sphere. The two hemispheres of this sphere separated by the “equator” of colinear triangles contain sets of triangles that are mirror images of each other with the apex vertex reflected across the basal chord. During Procrustes alignment, all three corresponding vertices are rotated to corresponding positions, which reduces the full triangle shape sphere to a Procrustes-aligned shape hemisphere.

Analogous shape manifolds can be constructed for any collection of objects whose forms can be represented, or approximated, by sets of topologically corresponding landmark locations. These manifolds represent precise mathematical realizations of the more ambiguous particle shape spaces created by Wentworth (1922), Zingg (1935), and Sneed and Folk (1958). Such shape spaces can be used either for describing and categorizing the shapes of objects or testing hypotheses concerning the shape similarities or differences between objects and groups (Fig. 5). If care is taken to match topologically corresponding points located either on, or internal to, the outlines of objects, these geometric methods can be extended to the consideration of any polygon-defined shape in either two or three dimensions.



Shape, Fig. 5 Cross sections through the Procrustes shape hemisphere (= Kendall's pre-shape space) illustrating various spatial relations between the reference (gray) and target (white) landmark-defined shapes. (a). Between any two configurations, the Procrustes distance (D) is the length of the shortest arc between the configuration's points on the shape manifold's surface. The partial Procrustes distance (D_p) is the linear distance between these points. For point comparisons for which the arc angle (θ) is small, D_p will provide a good estimate of D . Interestingly, a closer match between the target and reference configuration's points can often be found by relaxing the constraint of size similarity. The full Procrustes distance (D_f) is the minimal distance between the reference and the target configurations when the configuration size of the target is allowed to vary, but not its shape. (b). The distribution of a shape or series of shapes with respect to a reference shape may be portrayed in the

linear spaces that have become traditional in pre-geometric morphometrics by orthogonally projecting their positions on the Procrustes shape manifold to a linear plane. This operation can be accomplished via covariance-based PCA of the complete set of partial warp scores (= relative warps analysis) or direct covariance-based PCA of the Procrustes shape coordinates expressed as deviations from the reference shape. Under this convention the projection plane will be tangent to the shape manifold at the point of the reference shape and exhibit a reduced dimensionality from that of the shape manifold's surface. If the total range of shape variation present within the sample is small, the apparent Procrustes distances ($\hat{D}_p$) between point locations in this tangent space will be good estimates of both the partial Procrustes distance (D_p) and (by extension) the Procrustes distance (D).

Conclusion

The existence of two distinct traditions of shape characterization and analysis in the earth sciences presents challenges to communication across this highly interdisciplinary field. Many of the mathematical formalisms and data analysis methods that are currently applied to biological/paleontological shapes can also be applied successfully to non-biological shapes, often to the advantage of hypothesis tests (e.g., avoidance of the problems associated with the calculation of ratios, reduction in geometric ambiguity). In many practical cases, though, a high degree of geometric fidelity and mathematical sophistication is not necessary in order to characterize a sample of shapes or test a shape-related hypothesis. Moreover, newly available machine-learning approaches to generalized object analysis seem certain to add a new set of options for the analysis of shapes in an earth-science context. While no generalized advice can be offered regarding an approach to pursue in any particular case, familiarity with both the concepts and methods of particle analysis and geometric morphometrics (in addition to machine learning) will ensure access to the greatest possible range of shape-related concepts and shape analysis tools.

References

- Blackith RE, Reyment RA (1971) Multivariate morphometrics. Academic Press, London
- Blott SJ, Pye K (2008) Particle shape: a review and new methods of characterization and classification. *Sedimentology* 55:31–63
- Bookstein FL (1991) Morphometric tools for landmark data: geometry and biology. Cambridge University Press, Cambridge
- Bookstein FL (1993) A brief history of the morphometric synthesis. In Marcus LF, Bello E, García-Valdecasas A, (eds) *Contributions to morphometrics*. Museo Nacional de Ciencias Naturales 8, Madrid
- Bookstein FL (1978) The measurement of biological shape and shape change. Springer, Berlin
- Feret LR (1930) La grosseur des grains des matières pulvérulentes. *Premières Communications de la Nouvelle Association Internationale pour l'Essai des Matériaux* D:428–436
- Hockey B (1970) An improved coordinate system for particle shape representation. *J Sediment Petrol* 40:1054–1056
- Kendall DG (1984) Shape manifolds, procrustean metrics and complex projective spaces. *Bull Lond Math Soc* 16:81–121
- Sneed EDR, Folk L (1958) Pebbles in the Lower Colorado River, Texas a study in particle morphogenesis. *J Geol* 66:114–150
- Thompson DW (1917) *On growth and form*. Cambridge University Press, Cambridge
- Wadell H (1932) Volume, shape and roundness of rock particles. *J Geol* 40:443–451
- Wentworth CK (1922) The shapes of beach pebbles. *USGS Prof Pap* 131–C:75–83
- Zingg T (1935) Beitrag zur schotteranalyse. *Schweiz Mineral Petrogr Mitt* 15:39–140

Signal Analysis

Mohamed Zinelabidine Doghmane¹, Sid-Ali Ouadfeul² and Leila Aliouane³

¹Department of Geophysics, FSTGAT, University of Science and Technology, Houari Boumediene, Algiers, Algeria

²Algerian Petroleum Institute, Sonatrach, Boumerdes, Algeria

³LABOPHT, Faculty of Hydrocarbons and Chemistry, University M'hamed Bougara of Boumerdes, Boumerdes, Algeria

Definitions

Signal Analysis

Signal analysis is an essential and important step for all earth sciences including geophysics; its main aim is to process and interpret all types of registered signals in order to carry out all possible required information (Dimri 2005). Therefore, its usefulness can be judged by its success in extracting maximum useful information for a given signal, which is generally disturbed by different noises (Li et al. 2018); the reliability of the obtained results is strongly dependent on the electronic measurement devices and computational process. This chapter is dedicated to overview the signal analysis steps applied in earth sciences, with some focus on geophysical applications (Herrera et al. 2013).

The Signal

A signal is a term that is applied for any physical representation of information in different forms including mathematical functions, which are related to the inputs of a given system to its outputs. The mathematical description of a signal s provides the possibility to represent, analyze, and treat the information by using mathematical technique sets in the so-called processing systems (Robinson 1981). For example, in geophysics, the signal can represent the information obtained from the response of the geological layers when submitting an artificial energy signal through the earth (Dimri 2005). The interpretation of these seismic signals allows us to construct an image on the different structures of the subsurface (Liu et al. 2016).

The Noise

Any kind of signals that is disturbing the registration, transmission, or interpretation of the main signal is considered to be a noise N . Since the noise is also a type of signal, the concepts of signal and noise are relative to what interest the interpreters in the first place (Li et al. 2018). For example, in astronomy the signal coming from satellites is considered to be noise, and the main signal is the one received from astrophysical resources (i.e., the sun), while for telecommunications, the satellite

signals are the targets. In geophysics, the signals coming from the subsurface are considered to be the main target, while other signals from surface are seen as noise (Li et al. 2018).

The Signal to Noise Ratio (s/N)

This ratio is conventionally used to measure the importance of the noise in comparison to the main signal; it gives a direct indication on how the signal is noised (Wang et al. 2012). If this ratio is close to 1, it means the signal is very noisy and an important part of the signal is overlapped in time with the noise (Li et al. 2018). It is practically calculated by dividing the power of the signal (P_s) by the power of the noise (P_N), and it is often expressed in decibels (dB) as demonstrated by Eq. (1)

$$\left(\frac{S}{N}\right)_{dB} = 10 \log\left(\frac{P_s}{P_N}\right) \quad (1)$$

Signal Classification

The signals are classified according to their properties into different classes, where the most important characteristic of the signal is its evolution as a function of time. Based on that, there are two types of signals.

Deterministic Signals

The evolution of such type of signals over time can be modeled by well-defined mathematical functions; in this type of signal, many classes can be found such as periodic signals, transient signals, and pseudo-random signals.

Random Signals

These signals are characterized by the property that their temporal behavior is unpredictable; therefore, it is necessary to appeal to their statistical properties to describe them. If their statistical properties are invariant in time, they are said to be stationary.

Other Classification

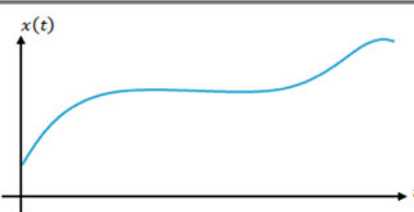
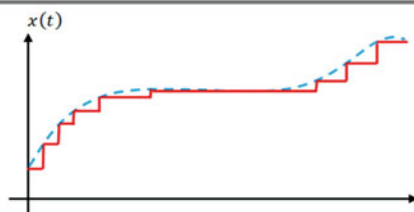
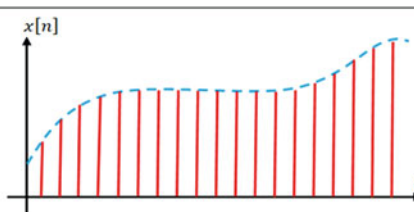
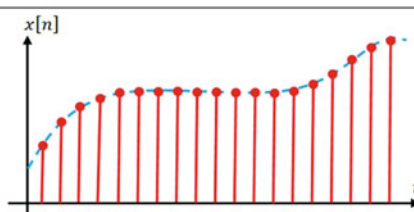
Other classifications of the signals are based on the nature of their variables; continuous signals are distinguished from discrete signals by the nature of the time variable as well as those whose amplitude is discrete or continuous as detailed in Table 1.

Based on Table 1, the signals can be therefore classified into four main types:

- Type 1 (analog signals): their amplitude and time are both continuous.
- Type 2 (quantized signals): its amplitude is discrete while the time is continuous.
- Type 3 (sampled signals): its amplitude is continuous while its time is discrete.
- Type 4 (digital signals): both its amplitude and time are discrete.

Since the seismic signals are analyzed using black box software, they are all digitalized in both amplitude and time; thus, they are all type 4 (Liu et al. 2016). Nevertheless, other types can be also obtained with some predefined requirements.

Signal Analysis, Table 1 Classification of signals based on the nature of its variable and amplitude

		Amplitude	
		Continuous	Discrete
Time	Continuous		
	Discrete		

Important Functions for Signal Analysis

In order to proceed to signal analysis, it is necessary to define some reference signals that will be used later to simplify the mathematical calculations of other techniques for domain transformation. The main signals to ensure are the following.

Sign Function

The sign function is mathematically defined by Eq. (2), and its graphical presentation is given by Fig. 1a. It is conventionally admitted that the value at the origin is $\text{sgn}(t = 0) = 0$.

$$\text{sgn}(t) = \begin{cases} -1 & \text{for } t < 0 \\ +1 & \text{for } t > 0 \end{cases} \quad (2)$$

Step Function

The step function is mathematically defined by Eq. (3) and graphically represented by Fig. 2b. It is conventionally admitted that $u(t = 0) = 1/2$. In some literature, the value given at the origin is 1.

$$u(t) = \begin{cases} 0 & \text{for } t < 0 \\ 1 & \text{for } t > 0 \end{cases} \quad (3)$$

Ramp Function

The ramp function is mathematically defined by Eq. (4) and graphically represented by Fig. 2c.

$$r(t) = t \cdot u(t) = \int_{-\infty}^t u(\tau) d\tau \quad (4)$$

Rectangular Function

The rectangular function is mathematically defined by Eq. (5) and graphically represented by Fig. 2d. It is also called the window function because it serves as a basic windowing function.

$$\text{rect}\left(\frac{t}{T}\right) = \begin{cases} 1 & \text{for } \left|\frac{t}{T}\right| < \frac{1}{2} \\ 0 & \text{for } \left|\frac{t}{T}\right| > \frac{1}{2} \end{cases} \quad (5)$$

Dirac Pulse Function

The Dirac function can be seen as a rectangular function with a width T that tends towards 0 and whose area is equal to 1. It is mathematically defined by Eq. (6) and graphically schematized by Fig. 2e.

$$\delta(t) = \begin{cases} \infty & \text{for } t = 0 \\ 0 & \text{for } t \neq 0 \end{cases} \quad (6)$$

Figure 2e shows the schema of the function but does not represent it because it contains an infinite term; the arrow in this figure indicates the function whose surface should be equal to 1. It should also be noticed that the Dirac function is the derivative of the step function $\delta(t) = \frac{du(t)}{dt}$. The importance of this function is that it is used to sample other complex signals; the passage from one state to another in a given signal is seen as a pulse (Dirac function) if the rise time is very negligible in comparison to the original signal global time (Wang et al. 2012).

Dirac Comb Function

The Dirac comb function is a periodic succession of Dirac impulses, its mathematical presentation is given by Eq. (7), and its graphically appearance is highlighted in Fig. 1f.

$$\delta_T(t) = \sum_{k=-\infty}^{+\infty} \delta(t - kT) \quad (7)$$

T is the period of the comb. This sequence is sometimes referred to as a pulse train or a sampling function because this type of signals is mainly used in sampling.

Cardinal Sinus Function

This function plays an important role in the signal analysis due to its mathematical properties; it is defined mathematically by Eq. (8) and highlighted graphically by Fig. 1g

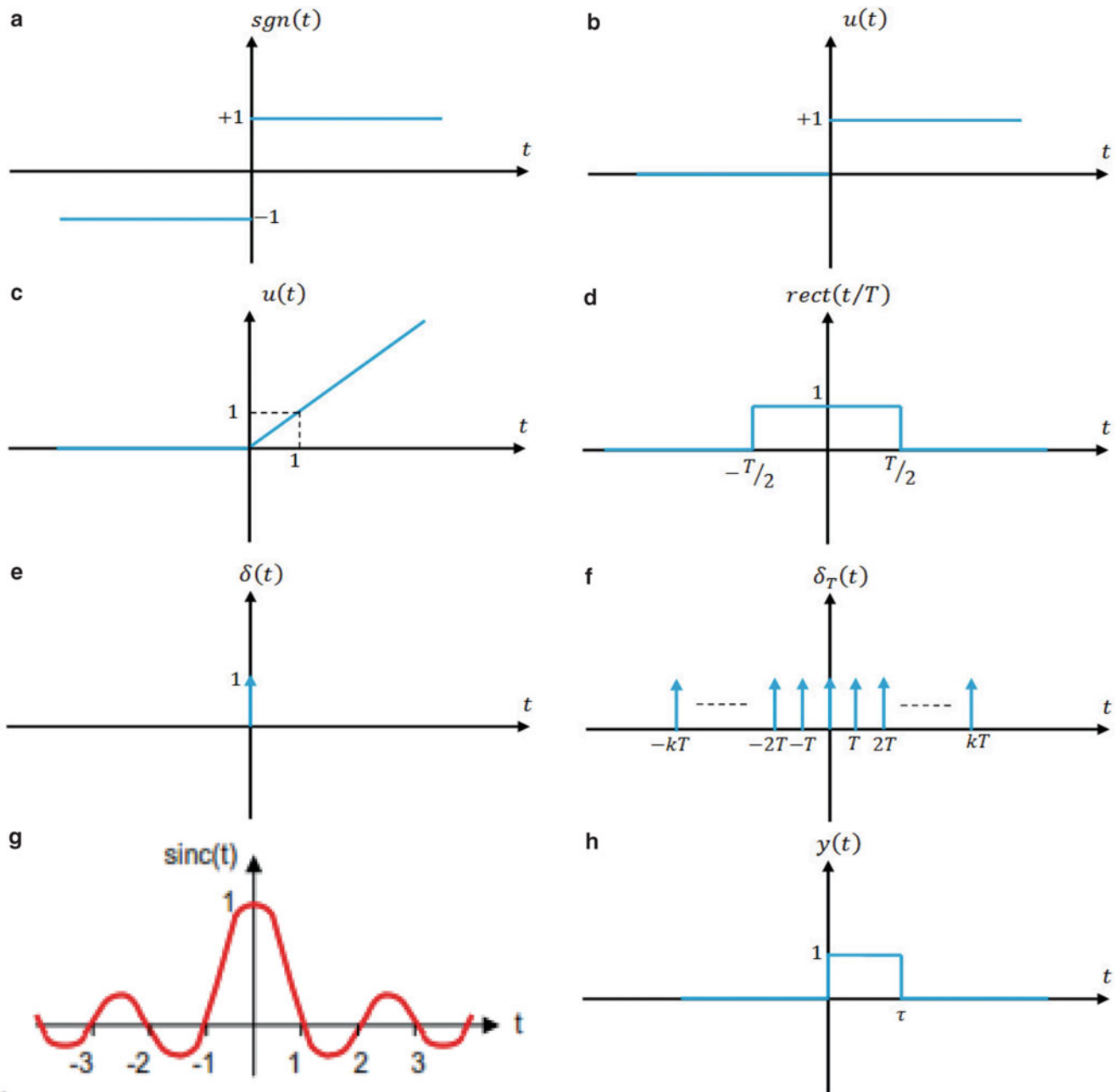
$$\text{sinc}(t) = \frac{\sin(\pi t)}{\pi t} \quad (8)$$

One of the main important properties of the cardinal sinus function is expressed by

$$\int_{-\infty}^{+\infty} \text{sinc}(t) dt = 1, \quad \int_{-\infty}^{+\infty} \text{sinc}^2(t) dt = 1$$

Frequency Representation

The signals are generally presented as a function of time because our perception of physical world is based on taking the time as the most important variable for all phenomena. However, in many cases this presentation cannot provide deep analysis especially for semi-periodic and quasiperiodic complex phenomenon in time (Benedetto 2010). Therefore, the necessity to transfer these signals into other domains rather than time became primordial. The knowledge of the spectral properties of a signal from its energy is essential (Huang et al. 2015). Thus, the frequency representation of the signal in the frequency domain can be very efficient for energy-phenomenon characterization. This transformation

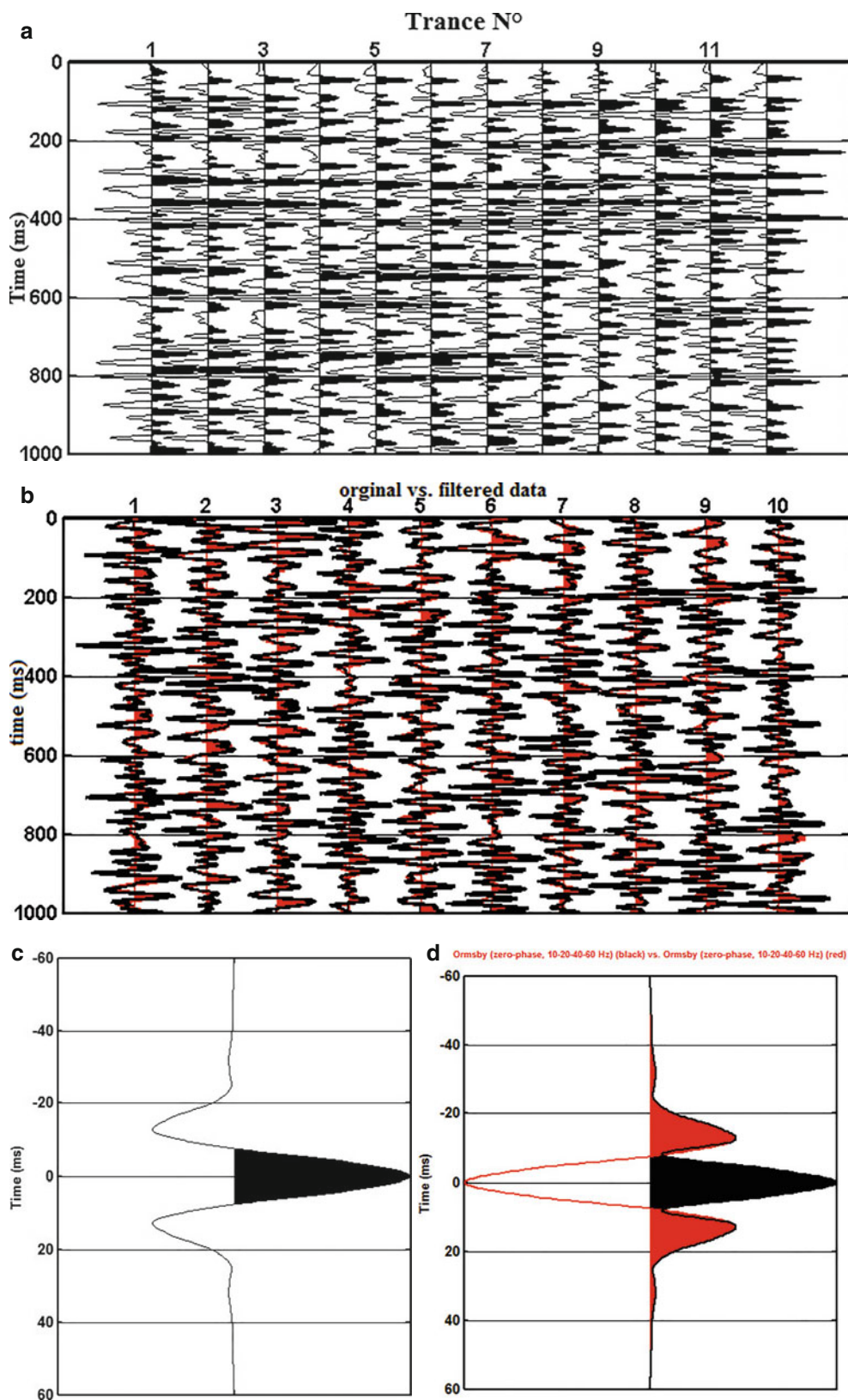


Signal Analysis, Fig. 1 Most important functions for signal analysis: (a) sign function, (b) step function, (c) ramp function, (d) rectangular function, (e) Dirac pulse function, (f) Dirac comb function, (g) cardinal sinus function, (h) window function for band-pass filter

of a signal from time domain to the frequency domain has many advantages, i.e., it can help in filtering the noise from original signal since the noise is generally characterized by high-frequency dynamics in seismic data (Liu et al. 2016). The Fourier series have been firstly proposed in order to decompose the perfectly periodic signals into sinusoids so that passage can be easily (Chakraborty and Okaya 1995). Later on, the Fourier transform has been introduced as the generalization of Fourier series for all nonperiodic functions.

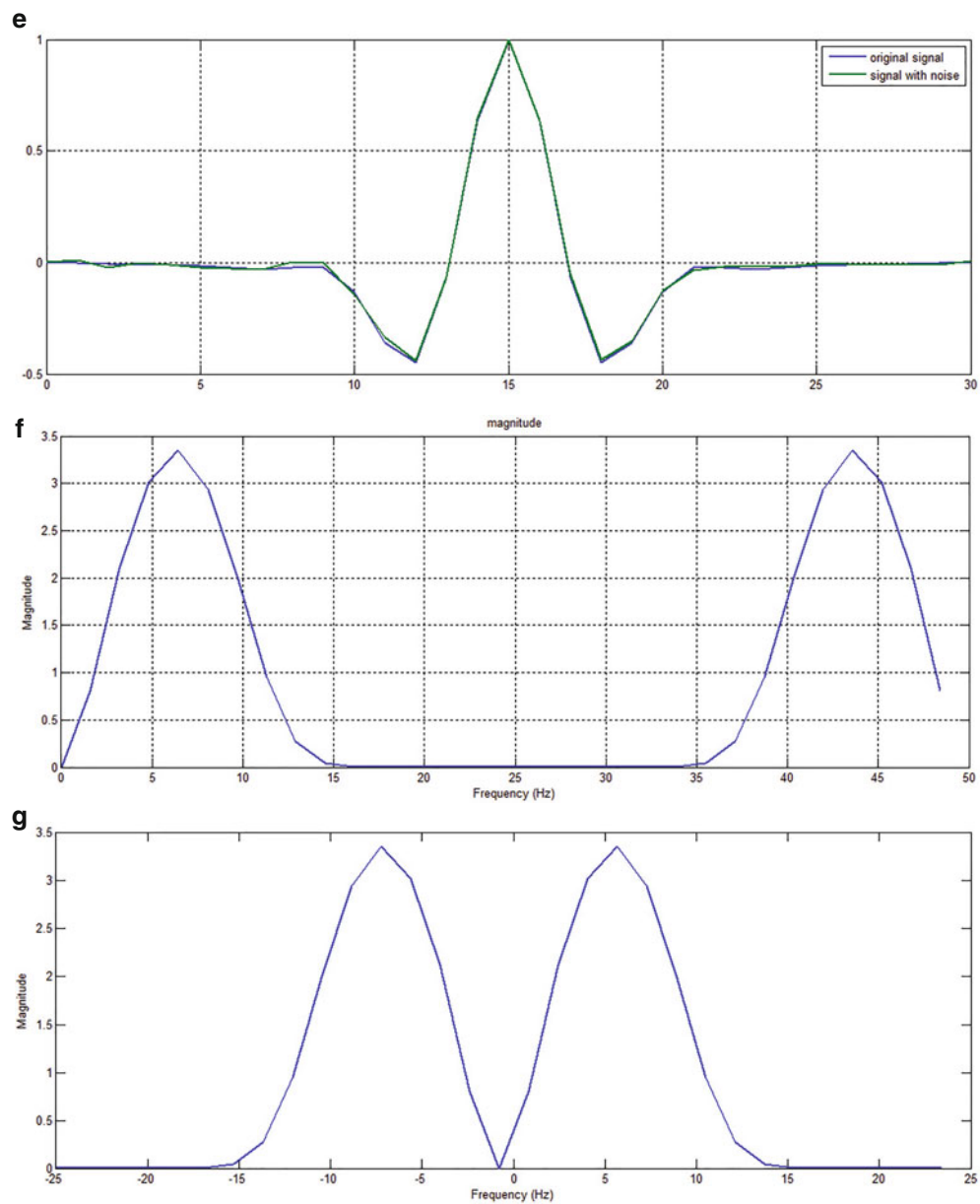
Fourier Series

The Fourier series principle is based on decomposing the periodic signal into a sum of sinusoids. Thus, it permits to transform easily from time domain to the frequency domain but with the necessary condition that the decomposed signal should have bounded variations. Let's consider the periodic bounded signal $s(t)$; its Fourier series can be written as:

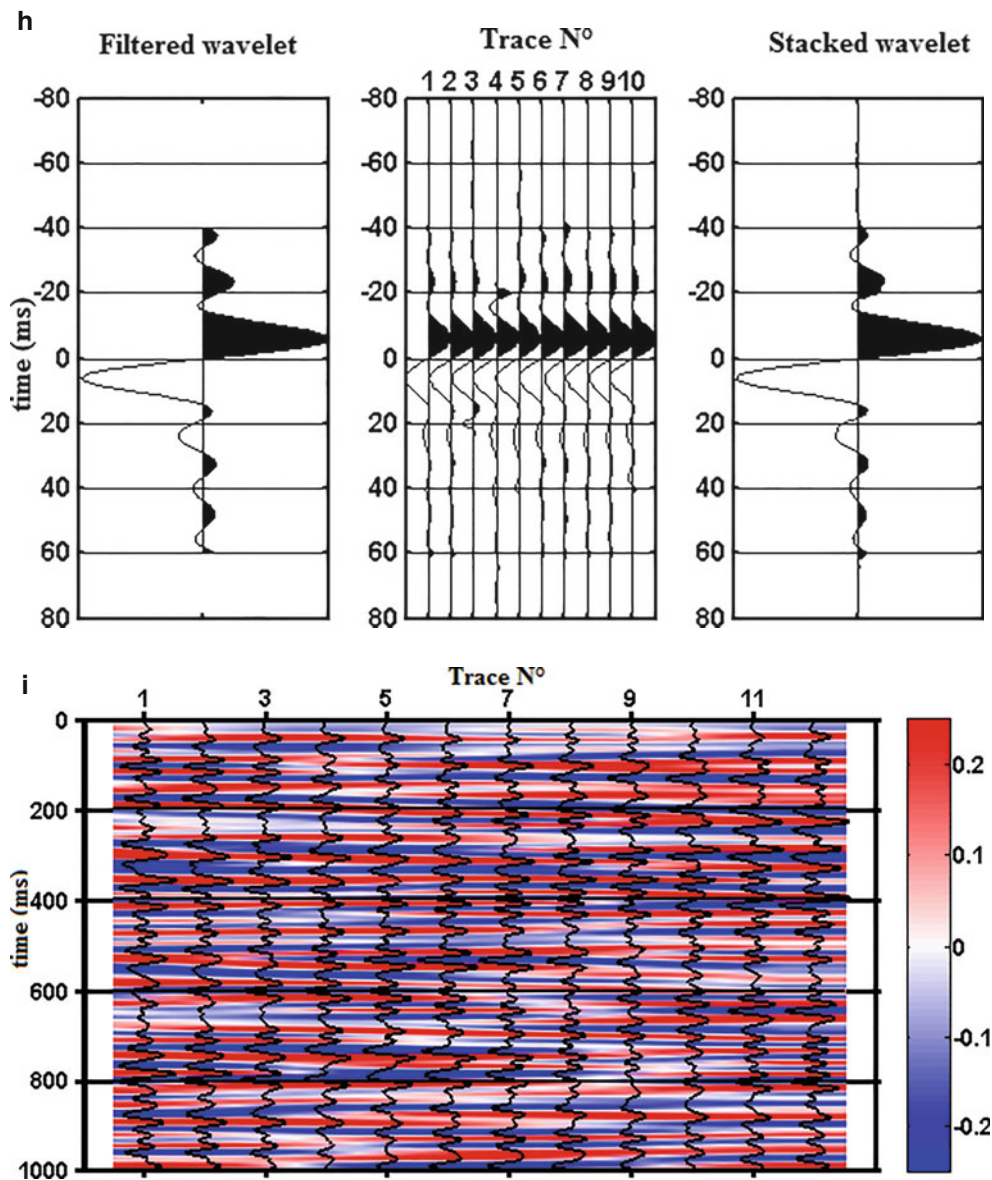


Signal Analysis, Fig. 2 Examples of seismic data analysis: (a) original seismic trace responses of earth, (b) comparison between original and noise traces, (c) one trace extracted from a, (d) one trace extracted from noise data, (e) comparison between original and noised wavelet,

(f) Fourier transform of the wavelet in e, (g) fast Fourier transform of the wavelet in e, (h) example of filtered and stacked wavelet, (i) complete seismic trace



Signal Analysis, Fig. 2 (continued)



Signal Analysis, Fig. 2 (continued)

$$s(t) = S_0 + \sum_{n=1}^{\infty} [A_n \cos(n\omega_0 t) + B_n \sin(n\omega_0 t)] \quad (9)$$

with $\omega_0 = \frac{2\pi}{T_0}$, and $S_0 = \frac{1}{T_0} \int_{(T_0)} s(t) dt$, $A_n = \frac{2}{T_0} \int_{(T_0)} s(t) \cos(n\omega_0 t) dt$, $B_n = \frac{2}{T_0} \int_{(T_0)} s(t) \sin(n\omega_0 t) dt$ is the fundamental pulsation, $n\omega_0$ is the n rank harmonic, and S_0 is the average value of the signal $s(t)$. Other complex presentations of the Fourier series can be found in the literature too. The main limitation of the Fourier series is that it is applied only for periodic functions where the signal variables should be bounded too. For that reason, the Fourier transform is

developed from Fourier series by supposing that all non-periodic functions are periodic functions with a period that tends to infinity as explained in section “[The Fourier Transform](#)”.

The Fourier Transform

The Fourier transform is the generalization of the Fourier series decomposition to all deterministic signals; thus, it provides the spectral presentation of all signals all over the frequency range (Huang et al. 2015). In other terms, it

demonstrates the variation of the amplitude (or energy) of the signal as a function of its frequency. In order to explain that in details, let $s(t)$ be now a deterministic signal; its Fourier transform is a complex function of the frequency, and it is given by Eq. (10).

$$S(f) = FT[s(t)] = \int_{-\infty}^{+\infty} s(t)e^{-j2\pi ft} dt \quad (10)$$

It has been proven that if the transform given by Eq. (10) exists, then the inverse Fourier transform can be obtained by Eq. (11). The modulus of the Fourier transform of a signal is called the spectrum.

$$s(t) = TF^{-1}[S(f)] = \int_{-\infty}^{+\infty} S(f)e^{j2\pi ft} df \quad (11)$$

Table 2 summarizes some of the most useful properties of Fourier transform for signal analysis in earth sciences in general and for seismic processing in particular (Robinson 1981).

Convolution and Deconvolution Principle and the Fourier Transform

The convolution product of a given signal $s(t)$ by another signal $h(t)$ is given by Eq. (12).

$$s(t) * h(t) = \int_{-\infty}^{+\infty} s(k)h(t-k)dk \quad (12)$$

The main utility of the convolution is that it is used to extract the transfer function of systems, since, the output signal of convolution of the input direct impulse function with the transfer function of the system is the transfer function itself; this is called the impulse response of the system. The value of the output signal at time t is thus obtained by summing the past values of the excitation signals, weighted by

system response. It is proven that $TF[a(t) * b(t)] = A(f) \cdot B(f)$.

Filtering Concept

Filtering is a kind of signal analysis, where the frequency spectrum and the phase of the signal are changed based on our need; therefore, the temporal form of the signal is consequently changed (Huang et al. 2015). The changes can be either eliminating (or at least weakening) the unwanted frequencies of the noises or isolating the useful frequencies of the main signal (Li et al. 2018). Thence, an ideal filter can be described by zero attenuation in the frequency band that we want to keep and an infinite attenuation in the band that we want to eliminate. In practice, it is impossible to obtain such a perfect filter, but the best approach is to keep the attenuation lower than a given value A_{max} in the bandwidth and greater than A_{min} in the attenuated band. The bandwidth for filters of seismic data is defined based on regional studies, and they are generally between 20 and 80 Hz (Chakraborty and Okaya 1995). Based on the filter template, which defines the attenuation and the prohibited zones, four types of filters can be constructed: (1) low-pass filter, (2) high-pass filter, (3) band-pass filter, and (4) band-cut filter. By sizing the filter window, only a few analytical functions can be used, which their characteristics can be suitable for the realization of the template. Examples of such functions that will set the physical properties of the filter are Butterworth, Tchebycheff, Bessel, and Cauer.

Modulation Concept

The principle of modulating a signal is mainly used for the signal transmission; it is based on adapting the signal to be transmitted through the transmission channel, wherein modulation of the signal amplitude necessitates a frequency transposition of signal; it is achieved by a simple multiplication, where the necessary bandwidth to transmit a signal of f frequency is $2f$. In order to retrieve the original signal, demodulation process should be involved.

Signal Analysis, Table 2 Properties of the Fourier transform

Domain	$s(t)$	$S(f)$
Linearity	$a \cdot s(t) + b \cdot r(t)$	$a \cdot S(f) + b \cdot R(f)$
Translation	$s(t - t_0)$	$e^{-j2\pi ft_0} S(f)$
	$e^{+j2\pi ft_0} s(t)$	$S(f - f_0)$
Inflection	$s^*(t)$	$S^*(-f)$
Derivation	$\frac{d^n s(t)}{dt^n}$	$(j2\pi f)^n S(f)$
Dilatation	$s(a \cdot t)$ with $a \neq 0$	$\frac{1}{ a } S\left(\frac{f}{a}\right)$
Convolution	$s(t) * r(t)$	$S(f) \cdot R(f)$
	$s(t) \cdot r(t)$	$S(f) * R(f)$
Duality	$S(t)$	$s(-f)$

Digitization

Since the outside world is analog by default, the necessity to analog-to-digital conversion is growing up due the computational limitation of the digital world; the analog conversion to digital is essentially based on three main steps. The first step is the sampling process in order to convert the analog signal into discrete; then the second step is the quantization, which is the association an amplitude value for each time step; and the last step is coding, where each value is associated to its

appropriate code. These steps are gathered in one process named sampling-holding process as detailed in the next subsection.

Sampling-Holding

In practice, the analog to digital conversion is not done instantly; there should be a blocking time of the analog signal for a period of T to have an error-free conversion. The sample-holding is therefore used to achieve this step, where the blocking effect is generally modeled by a window function shifted by a period of $\tau/2$. The mathematical function of sampling and holding process is given by Eq. (13), and its graphical presentation is given by Fig. 1g.

$$\begin{aligned} y(t) &= \sum_{k \rightarrow -\infty}^{+\infty} \text{rect}\left(\frac{t - \frac{\tau}{2} - kT_e}{\tau}\right) \\ &= \text{rect}\left(\frac{t - \frac{\tau}{2}}{\tau}\right) * \sum_{k \rightarrow -\infty}^{+\infty} \delta(t - kT_e) \end{aligned} \quad (13)$$

The process of sampling-and-blocking can therefore be seen as multiplying the signal $s(t)$ by $y(t)$. The Fourier transform of the sampled signal is therefore can be given by Eq. (14).

$$S_e(f) = \frac{\tau}{T_e} \text{sinc}(\tau f) \cdot \sum_{k \rightarrow -\infty}^{+\infty} S(f - kf_e) \cdot e^{-j\pi f \tau} \quad (14)$$

Quantification

The quantization principle consists in adding to any real value x any other value x_q that belongs to a finite set of values according to a certain law: rounding up, rounding up closer, etc. The gap between each value x_q is called no quantization. Rounding off the start value inevitably results in a quantization error called quantization noise (Li et al. 2018).

Coding

Coding consists of associating a composite code with a set of discrete values of binary elements. The most famous codes: natural binary code, shifted binary code, code DCB, and gray code.

FFT: Fast Fourier Transform

In order to be able to apply Fourier Transform with large data, an algorithm is used to obtain the transformation in a much faster time under pre-defined conditions. This algorithm is called the fast Fourier transform. The transformation in discrete domains is given by Eq. (15).

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot W_N^{nk}, \quad (15)$$

$$\text{with } W_N = e^{-\frac{2j\pi}{N}}, \quad W_N^{2nk} = e^{-\frac{2j\pi nk}{N/2}} = W_{N/2}^{nk}, \quad \text{and } W_N^{nk+N/2} = e^{-\frac{2j\pi(nk+N/2)}{N}} = -W_N^{nk}$$

The necessary condition to use the FFT is that the number of sample should be written as a power of 2: $N = 2^m$; then by separating the odd and the even indices, the repeated calculation operations will be separated as follows: $x_1[n] = x[2n]$ and $x_2[n] = x[2n + 1]$. By exploring this property, we finally find the FFT as given by Eq. (16). The calculation rate will be minimized from N^2 to $N \log_2(N)$.

$$\begin{cases} X[k] = X_1[k] + W_N^k \cdot X_2[k] \\ X\left[k + \frac{N}{2}\right] = X_1[k] - W_N^k \cdot X_2[k] \end{cases} \quad \text{for } 0 \leq k \leq \frac{N}{2} - 1 \quad (16)$$

Example from Geophysics

In the example given in Fig. 2, we will show a seismic signal and some signal analysis techniques that can be applied to it in order to extract the desired information; the signal represents a seismic trace that shows a typical response (reflection) of earth rock to artificially submitted energy signal (Efi and Kumar 1994).

Conclusions

In this chapter, we have briefly reviewed the theory of signal analysis from mathematical perspectives, where definitions and most important functions and their mathematical properties have been given in order to allow beginner students to construct a general idea about this topic. Moreover, examples of signal analysis applications in geophysics have been demonstrated; they will permit future earth scientists in general and geophysicists in particular to link the theoretical information delivered in this chapter with practical applications in their field of study.

Cross-References

- [Compositional Data](#)
- [Fast Fourier Transform](#)
- [Inversion Theory.](#)
- [Mathematical Geosciences,](#)
- [Power Spectral Density](#)
- [Quality Control](#)
- [Random Variable](#)
- [Spatial Analysis](#)

- [Spectral Analysis](#)
- [Time Series Analysis](#)
- [Wavelets in Geosciences.](#)

Bibliography

- Benedetto A (2010) Water content evaluation in unsaturated soil using GPR signal analysis in the frequency domain. *J Appl Geophys* 71(1): 26–35. <https://doi.org/10.1016/j.jappgeo.2010.03.001>
- Chakraborty A, Okaya D (1995) Frequency-time decomposition of seismic data using wavelet-based methods. *Geophysics* 60(6). <https://doi.org/10.1190/1.1443922>
- Dimri V (2005) Fractals in geophysics and seismology: an introduction. In: Dimri VP (ed) *Fractal behaviour of the earth system*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-26536-8_1
- Efi F-G, Kumar P (1994) Wavelet analysis in geophysics: an introduction. *Wavelet Anal Appl* 4:1–43. <https://doi.org/10.1016/B978-0-08-052087-2.50007-4>
- Herrera RH, Han J, van der Baan M (2013) Applications of the synchrosqueezing transform in seismic time-frequency analysis. *Geophysics* 79(3). <https://doi.org/10.1190/geo2013-0204.1>
- Li F, Zhang B, Verma S, Marfurt KJ (2018) Seismic signal denoising using thresholded variational mode decomposition. *Explor Geophys* 49(4). <https://doi.org/10.1071/EG17004>
- Liu W, Cao S, Chen Y (2016) Seismic time–frequency analysis via empirical wavelet transform. *IEEE Geosci Remote Sens Lett* 13(1): 28–32. <https://doi.org/10.1109/LGRS.2015.2493198>
- Robinson EA (1981) Signal processing in geophysics. In Bjørnø L (eds) *Underwater acoustics and signal processing*. NATO advanced study institutes series (series C — mathematical and physical sciences), vol 66. Springer, Dordrecht. https://doi.org/10.1007/978-94-009-8447-9_56
- Wang T, Zhang M, Yu Q, Zhang H (2012) Comparing the applications of EMD and EEMD on time–frequency analysis of seismic signal. *J Appl Geophys* 83:29–34. <https://doi.org/10.1016/j.jappgeo.2012.05.002>. ISSN 0926-9851
- Huang W, Wang R, Yuan Y, Gan S, Chen Y (2015) Signal extraction using randomized-order multichannel singular spectrum analysis. *Geophysics* 82(2). <https://doi.org/10.1190/geo2015-0708.1>

Signal Processing in Geosciences

E. Chandrasekhar¹ and Rizwan Ahmed Ansari²

¹Department of Earth Sciences, Indian Institute of Technology Bombay, Mumbai, India

²Department of Electrical Engineering, Veermata Jijabai Technological Institute, Mumbai, India

Definition

Geophysical signal processing is an important branch of geophysics that deals with various signals and systems. It serves as a backbone for efficient analysis and interpretation of the complex geophysical data. Concepts of signal processing are encountered at various stages, right from the data

acquisition to processing to interpretation. Over the years, geophysical signal analysis techniques have steadily progressed from the use of basic integral transforms to present-day nonlinear signal processing techniques such as wavelet analysis, fractal and multifractal analysis, and empirical mode decomposition technique, which have found their forays in geosciences, paving the way for effectively unravelling the hidden information from the nonlinear and nonstationary geophysical data. The use of these novel signal analysis techniques in geophysical data analysis is always important and essential for a better understanding of the dynamics of the systems that are generating such complex geophysical data.

Introduction

The word “signal” can be best defined as the time or space variation of any physical quantity. While geophysical data constitute time signals (e.g., radiometric data, ionospheric total electron content (TEC) data), and space signals (e.g., well-log data), some are also classed into spatiotemporal signals, which vary as a function of both time and space (e.g., Earth’s magnetic and gravitational field). Also, there are some geophysical signals, which are causal (e.g., seismic data, well-log data, etc.) and noncausal (e.g., Earth’s upper atmospheric data, magnetic field, gravity field, to name a few) (Causal signals are those, which do not occur before the onset of any excitation. Noncausal signals do not require any excitation to occur. They are continuously generated naturally, since time immemorial.). If the statistical properties, such as mean, variance, etc. of a signal, do not vary as a function of space/time, they are defined as stationary signals. Otherwise, they are defined as nonstationary signals. Most geophysical signals are nonstationary in nature and a very few, namely, geomagnetic solar quiet day (Sq) variations, and the TEC data, may be called as quasi-stationary, as they generally exhibit some known periodicity in their temporal behavior. Initially, while the statistical methods helped to understand the behavior of geophysical signals by facilitating to estimate the trends and cyclicities present in the data, the well-known mathematical transforms, such as the Fourier transform and Hilbert transform, to name a few, have found their role in solving the potential field problems, thereby significantly improving the understanding and interpretation of the anomalous in situ geophysical signals. However, since all the above techniques are most effective in analyzing stationary or quasi-stationary signals, they lack in extracting some important information, such as time-frequency localization, from the nonlinear geophysical signals. Before discussing about nonlinearity in the signals, it would be more prudent, if we understand the difference between linearity and nonlinearity in signals, in the first place. If any observational data (y) can

be modeled in the form of, say, $y = mx + c$, where, x is any variable, then that data exhibits linear behavior. On the other hand, if any observational data best fits to a model having exponents, say, $y = x^m$ ($m \neq 1$), then such data exhibits nonlinear behavior. Most observational geophysical data, such as gravity data, magnetic data, seismic and seismological data, well-log data, geomagnetic data, ionospheric data, and weather data, to name a few are all nonlinear in nature. Nonlinear signals exhibit a power-law behavior with time and thus to correctly understand their behavior with time, suitable nonlinear mathematical techniques need to be employed. Basically, the aim is to correctly understand the dynamics of the nonlinear system that is generating such nonlinear signals. For example, the seismological data generated at different places on the Earth are different, which are a result of different subsurface dynamics associated with them (Telesca et al. 2004). Therefore, to correctly understand such nonlinear dynamics of the subsurface, responsible for generating such diverse signals, we must employ nonlinear signal analysis techniques. With the advent of wavelet analysis and other nonlinear and data-adaptive techniques, such as fractal and multifractal analyses and empirical mode decomposition technique, significant improvements have been made in the analysis of geophysical signals.

In this chapter, we first discuss the theory behind the basic integral transforms (mentioned here, as linear signal processing techniques) and their applications in geophysics. We next discuss the theory and geophysical application of continuous and discrete wavelet transform, the fractal and multifractal analysis techniques, and the fully data-adaptive empirical mode decomposition (EMD) technique. Finally, we discuss how the EMD technique facilitates to quantitatively estimate the degree of subsurface heterogeneity, using well-log data of two different wells. For all the above techniques, adequate references are provided for the benefit of the reader to further explore and understand the geophysical applications of all these techniques.

Linear Signal Processing Techniques in Geophysics

In this section, we briefly define the basic integral transforms like the Fourier transform, the Z-transform, and the Hilbert transform and discuss their applications in geophysical signal analysis. Although there are many other integral transforms in the literature, it is not possible to discuss all of them here, and thus we stick only to the well-known above three transforms, which have seen many applications in geophysics.

The Fourier Transform

If a time-varying signal is denoted as $f(t)$, then its Fourier transform, $F(\omega)$, is mathematically expressed as

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t} dt \quad (1)$$

Equation (1) explains that the time signal treated with a complex exponential function and summed up over the entire time signal will result in the frequency content of that signal (If the kernel is a negative exponential function in the Fourier transform, it should be positive in the inverse Fourier transform and vice versa.). The greatness of this transformation lies in the fact that the Fourier transform is the only efficient technique that can delineate all the frequencies present in the signal in the entire frequency range $-\infty$ to $+\infty$. $F(\omega)$ is called the power spectrum of $f(t)$. $F(\omega)$ is a complex function, implying that the Fourier transform of a real function will give not only the amplitude spectrum ($|F(\omega)| = A(\omega) = \sqrt{R(\omega)^2 + I(\omega)^2}$) but also the phase spectrum ($\theta(\omega) = \tan^{-1}\left(\frac{I(\omega)}{R(\omega)}\right)$), where $R(\omega)$ and $I(\omega)$, respectively, denote the real and imaginary parts of $F(\omega)$. Accordingly, Eq. (1) can also be written as

$$F(\omega) = |F(\omega)| e^{i\theta(\omega)}$$

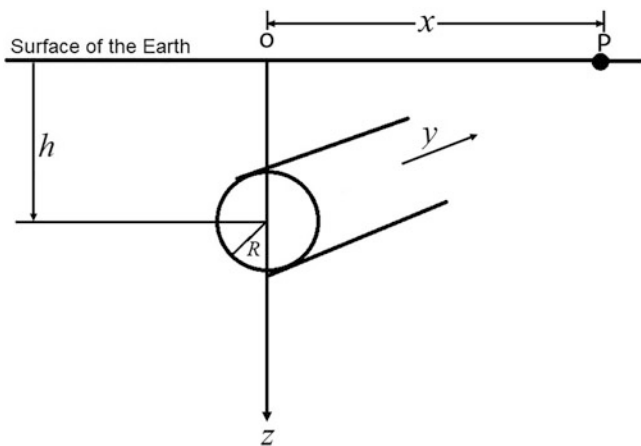
By applying a reverse operation, what is known as the inverse Fourier transform, it is also possible to get back the original signal, using the formula

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega)e^{i\omega t} d\omega \quad (2)$$

All types of signals, which satisfy the Dirichlet's conditions (see Hayes 1999), are Fourier transformable. Several properties of the Fourier transform can be found in Haykins (2006). Equation (1) can be applied to a variety of time, space, or spatiotemporal signals in geophysics to determine the parameters of the subsurface anomalous bodies, such as their depth of location, width, etc.

Application of the Fourier Transform to Determine the Subsurface Anomaly Parameters

As an example of the application of the Fourier transform in geophysics, we describe below a methodology to determine the parameters of a horizontal cylinder buried in the subsurface, by analyzing the gravity effect over it, using the Fourier transform. Figure 1 shows the schematic of the horizontal cylinder, having a radius R with depth to the center of the body as h , in the cartesian coordinate system.



Signal Processing in Geosciences, Fig. 1 A schematic representation of the buried horizontal cylinder (see text for the definitions of various parameters).

The gravity anomaly over a horizontal cylinder, as a function of horizontal distance, x , is mathematically expressed as

$$g(x) = \frac{2\pi\gamma\sigma R^2 h}{x^2 + h^2} \quad (3)$$

where γ is the universal gravitational constant, σ refers to the density contrast, and x designates the distance on the surface of the earth along a profile, over which, the gravity measurements are made.

From Eq. (1), the Fourier transform of Eq. (3) is given by

$$G(\omega) = 2\pi\gamma\sigma R^2 h \int_{-\infty}^{\infty} \frac{e^{-i\omega x}}{x^2 + h^2} dx \quad (4)$$

Since the integrand in Eq. (4) has poles at $x = \pm ih$, it needs to be solved using complex integration. Accordingly, considering the path of integration around $x = +ih$, the contour integration of Eq. (4) yields

$$\oint_{-\infty}^{\infty} \frac{e^{-i\omega x}}{x^2 + h^2} dx = \frac{\pi e^{\omega h}}{h} \quad (5)$$

Substituting Eq. (5) in Eq. (4), we get

$$G(\omega) = 2\pi^2 R^2 \gamma \sigma e^{\omega h} \quad (6)$$

Equation (6) at $\omega = 0$ gives

$$G(0) = 2\pi^2 R^2 \gamma \sigma \quad (7)$$

From Eq. (7) we have

$$R = \frac{1}{\pi} \sqrt{G(0)/2\gamma\sigma} \quad (8)$$

Similarly, Eq. (3) at $x = 0$ gives

$$g(0) = \frac{2\pi\gamma\sigma R^2}{h} = \frac{G(0)}{\pi h} \quad (9)$$

Therefore, from Eq. (9), we have

$$h = \frac{G(0)}{\pi g(0)} \quad (10)$$

Equations (8) and (10) explain the role of the Fourier transform in conveniently delineating the subsurface parameters, R and h of the buried anomalous body from its gravity effect. Similar examples of delineating various other parameters of various subsurface anomalous bodies using their respective geophysical anomalous effects can be found in Odegard and Berg (1965), Sharma and Geldart (1968), and Bhimasankaram et al. (1977a, b, 1978).

The Z-Transform

The Z-transform notation of a discrete signal, say, $[y_n] = y_0, y_1, y_2, y_3, \dots, y_{n-1}$, sampled at equal time intervals (denoted as subscripts) is generally expressed in the form of a polynomial in z as

$$Y(z) = y_0 + y_1 z + y_2 z^2 + y_3 z^3 + \dots + y_{n-1} z^{n-1} \quad (11)$$

For noncausal signals, Eq. (11) can be written as

$$Y(z) = y_{-n} z^{-n} + y_{-3} z^{-3} + y_{-2} z^{-2} + y_{-1} z^{-1} + y_0 + y_1 z + y_2 z^2 + y_3 z^3 + \dots + y_n z^n \quad (12)$$

where the coefficients of z represent the amplitudes of the signal at different times. Both in Eqs. 11 and 12, the z can be visualized as a unit delay operator. That means, the z operated on the sample at time t is actually the value of the signal after t units of time are elapsed. Accordingly, we can define

$$z[y_n] = y_{n-1}. \text{ Similarly, } z^2[y_n] = y_{n-2}; z^3[y_n] = y_{n-3}, \text{ etc.}$$

Mathematically, the unit delay operator, z , is expressed as $z = e^{i\omega}$. While the magnitude of this unit delay operator and all the powers of z is always 1, graphically, a single unit delay represents the time taken by the signal in completing one circle in the anticlockwise direction in the argand diagram, when ω goes from 0 to 2π . Similarly, a two-unit delay (i.e., z^2 or $e^{2i\omega}$) represents the time taken by the signal in completing the path 0 to 2π twice in the anticlockwise

direction, and a three-unit delay (i.e., z^3 or $e^{3i\omega}$) represents the time taken by the signal in completing the path 0 to 2π thrice in the anticlockwise direction and so on. This way the time-delayed representation of the signal at different delayed times can be truly expressed in z-transform notation, in the form of a polynomial in z .

Relation Between the Fourier Transform and Z-Transform Equation (11) can be written as

$$Y(z) = \sum_{t=0}^{n-1} y_t z^t \quad (13)$$

If we substitute, $z = e^{i\omega}$ in the above equation, we get

$$Y(\omega) = \sum_{t=0}^{n-1} y_t e^{i\omega t} \quad (14)$$

The analog form of Eq. (14) with extended limits, written as

$$Y(\omega) = \int_{-\infty}^{\infty} y(t) e^{i\omega t} dt \quad (15)$$

defines the Fourier transform.

An interesting question that arises as a consequence is: If the simple process of attaching powers of z to the discrete time signal defines the Z-transform, then the inverse Z-transform should amount to determining the coefficients of z . Let's look into this.

From Eq. (15), the inverse Fourier transform is written as

$$y(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} Y(\omega) e^{-i\omega t} d\omega \quad (16)$$

We know that the integration of $e^{in\omega}$ (or z^n) over the integral $-\pi \leq \omega < \pi$ is nonzero for $n = 0$ and is equal to 0 for all $n \neq 0$ i.e.

$$\begin{aligned} \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{in\omega} d\omega &= \frac{1}{2\pi} \int_{-\pi}^{\pi} (\cos n\omega + i \sin n\omega) d\omega \\ &= \begin{cases} 1 & \text{for } n=0 \\ 0 & \text{for } n \neq 0 \end{cases} \end{aligned} \quad (17)$$

Therefore, accordingly, changing the limits of integration and in terms of discretized function, Eq. (16) can be written as

$$\begin{aligned} y(t) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} (\dots + y_{-2} e^{-2i\omega} + y_{-1} e^{-i\omega} + y_0 + y_1 e^{i\omega} \\ &\quad + y_2 e^{2i\omega} + \dots) e^{-i\omega t} d\omega \end{aligned} \quad (18)$$

Invoking the Dirac delta function, the left-hand side of Eq. (18) can be written as

$$y(t) = \sum_l y_l \delta(t-l) \quad (19)$$

Therefore, by virtue of Eq. (17), all terms in Eq. (18) would vanish except the term with the coefficient of z with power zero, actually contributing to the integral. Thus, with the help of Eq. (19), we can write Eq. (18) as

$$y_t = \frac{1}{2\pi} y_t \int_{-\pi}^{\pi} d\omega = y_t \quad (20)$$

This proves that the inverse Z-transform facilitates to identify the coefficients of powers of z .

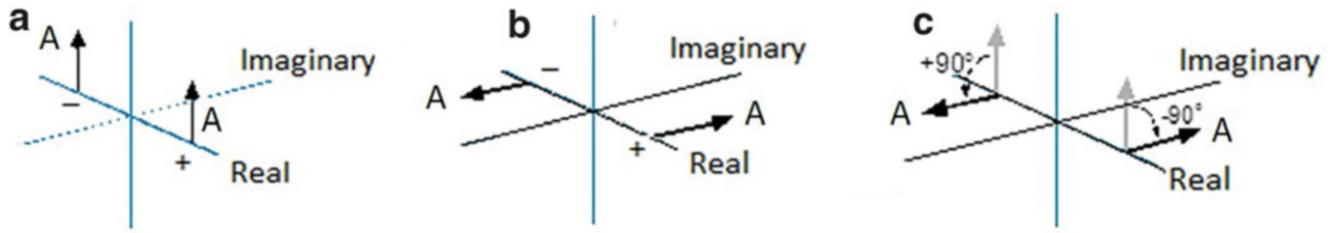
Z-transform has a lot of applications in designing the digital filters and in determining their stability by calculating the transfer functions and poles and zeros of digital filters (see Hayes 1999; Haykins 2006). Thus, the Z-transform has indirect application in geophysics, in the sense that they will be effectively used in the filtering and windowing operations of geophysical signals.

The Hilbert Transform

We earlier have seen that the Fourier transform operating on a given signal changes the time domain signal to frequency domain, thereby facilitating to determine its frequency content. The Hilbert transform on the other hand, while keeping the time signal in the same time domain, will change its phase content. In other words, Hilbert transform facilitates to characterize the signal based on its phase information.

The Hilbert transform is a linear time-invariant filter, whose transfer function results in changing the phase of each frequency in the input signals by an amount of $\pi/2$ radians. Let us understand the concept of Hilbert transform further.

Let us examine the spectral amplitudes of cosine and sine waves, as shown in Fig. 2. It can be seen from Fig. 2a that the spectral amplitude for cosine wave is symmetric and will appear only on the real plane for both positive and negative frequencies. In contrast, for sine wave (Fig. 2b), the spectral amplitude is asymmetric and will appear only on the imaginary plane for both positive and negative frequencies. Accordingly, the phase is $+90^\circ$ for the positive frequencies



Signal Processing in Geosciences, Fig. 2 Schematic representation of spectral amplitudes of (a) cosine and (b) sine waves and (c) and their rotational transformation. “A” designates the amplitude of cosine and sine waves

and -90° for the negative frequencies. From Fig. 2c, we can easily observe that if we rotate the negative frequency components of the cosine by $+90^\circ$ and the positive frequency components by -90° , we can transform a cosine function into a sine function. This rotational transformation is called Hilbert transformation. In other words, the Hilbert transform operation involves rotation of all negative frequency components of a signal by a $+90^\circ$ phase shift and all positive frequency components by a -90° phase shift (see Fig. 2c). Thus, it is equivalent to multiplying the negative phasor by $+i$ and the positive phasor by $-i$ in the frequency domain. It is important to note here that the Hilbert transform affects only the phase of the signal and has no effect on the amplitude.

If $g(t)$ is a real-valued time signal and $\hat{g}(f)$ denotes its Hilbert transform in the frequency domain, then $\hat{g}(f)$ is defined as

$$\hat{g}(f) = \begin{cases} -i g(f) & f > 0 \\ +i g(f) & f < 0 \end{cases} \quad (21)$$

Equation (21) implies that, while the Hilbert transform of a cosine function results in a sine function, the Hilbert transform of a sine function yields a negative cosine function. A close observation of Eq. (21) explains that it can be written as

$$\hat{g}(f) = -i \operatorname{sgn}(f) g(f) \quad (22)$$

where, $\operatorname{sgn}(f)$ denotes the signum function. Equation (22) defines the frequency domain representation of the Hilbert transform, representing the multiplication of two functions, namely, $-i \operatorname{sgn}(f)$ and $g(f)$. Therefore, according to the convolution theorem, Eq. (22) transforms to convolution of respective time functions in the time domain, given as (Convolution theorem states that the convolution of two functions in time domain is equal to multiplication of their respective Fourier spectra in frequency domain)

$$\hat{g}(t) = \frac{1}{\pi t} * g(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{g(\tau)}{t - \tau} d\tau \quad (23)$$

where $1/\pi t$ defines the inverse Fourier transform of the function $-i \operatorname{sgn}(f)$. Eq. (23) represents the time domain form of Hilbert transform. The Hilbert transform has a lot of applications in geophysics, particularly in solving the potential field problems involving delineation of subsurface anomaly parameters from the measured gravity or magnetic field in any area (Mohan et al. 1982; Sundararajan 1983; Sundararajan and Srinivas 2010).

Another Definition of Hilbert Transform

From Equation (1), we can have

$$F(\omega) = R(\omega) - iI(\omega) \quad \text{and} \quad F(-\omega) = R(\omega) + iI(\omega)$$

$$\begin{aligned} \text{where, } R(\omega) &= \int_{-\infty}^{\infty} f(t) \cos \omega t dt \quad \text{and} \quad I(\omega) \\ &= \int_{-\infty}^{\infty} f(t) \sin \omega t dt \end{aligned}$$

Therefore, $R(\omega)$ and $I(\omega)$ can be written as

$$R(\omega) = \frac{F(\omega) + F(-\omega)}{2} \quad \text{and} \quad I(\omega) = i \left(\frac{F(\omega) - F(-\omega)}{2} \right)$$

Equation (2) can be written as

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^0 F(\omega) e^{i\omega t} d\omega + \int_0^{\infty} F(\omega) e^{i\omega t} d\omega$$

Put $\omega = -\omega$ in the first integral; we have

$$f(t) = \frac{1}{2\pi} \int_0^{\infty} F(-\omega) e^{-i\omega t} d\omega + \int_0^{\infty} F(\omega) e^{i\omega t} d\omega$$

Using the above relations for $R(\omega)$ and $I(\omega)$, the above equation can be written as

$$f(t) = \frac{1}{\pi} \int_0^{\infty} R(\omega) \cos \omega t + I(\omega) \sin \omega t d\omega \quad (24)$$

From the above definition, Hilbert transform of Eq. 24 can be written as

$$\hat{f}(t) = \frac{1}{\pi} \int_0^{\infty} R(\omega) \sin \omega t - I(\omega) \cos \omega t d\omega \quad (25)$$

Equation (25) is another definition of Hilbert transform, which is widely used in the Hilbert analysis of geophysical signals.

Determination of Subsurface Anomaly Parameters from Gravity Field Using Hilbert Transform

Let us consider the mathematical expression for the gravity anomaly over a horizontal cylinder, as a function of horizontal distance, x , which is given in Eq. (3), whose Fourier transform is given in Eq. (6). We find that $R(\omega) = 2\pi^2 R^2 \gamma \sigma e^{\omega h}$ and $I(\omega) = 0$ in Eq. (6). Using Eq. (25), the Hilbert transform of Eq. (3) is given by

$$\hat{g}(x) = \frac{1}{\pi} \int_0^{\infty} R(\omega) \sin \omega x - I(\omega) \cos \omega x d\omega \quad (26)$$

Substituting $R(\omega)$ and $I(\omega)$ in Eq. (26), we have

$$\hat{g}(x) = 2\pi R^2 \gamma \sigma \int_0^{\infty} e^{\omega h} \sin \omega x d\omega \quad (27)$$

From standard integrals, the solution to Eq. (27) is given by

$$\hat{g}(x) = \frac{2\pi R^2 \gamma \sigma x}{x^2 + h^2} \quad (28)$$

If we plot $g(x)$ and $\hat{g}(x)$ as function of x , then at the point of intersection, say, x_1 , we have $g(x_1) = \hat{g}(x_1)$. Accordingly, we have

$$\begin{aligned} \frac{2\pi \gamma \sigma R^2 h}{x_1^2 + h^2} &= \frac{2\pi R^2 \gamma \sigma x_1}{x_1^2 + h^2} \\ \Rightarrow x_1 &= h \end{aligned} \quad (29)$$

This indicates that the depth to the center of the horizontal cylinder is simply the distance, x_1 , from the origin in the $g(x)$ and $\hat{g}(x)$ plot. Secondly, at the origin, we have

$$\begin{aligned} \hat{g}(0) &= 0; g(0) = \frac{2\pi \gamma \sigma R^2}{h} = \frac{2\pi \gamma \sigma R^2}{x_1} \\ \Rightarrow R &= \sqrt{\frac{g(0)x_1}{2\pi \gamma \sigma}} \end{aligned} \quad (30)$$

Thus, the subsurface anomaly parameters, R and h of the horizontal cylinder (Fig. 1) could be easily derived from the Hilbert transform. Similarly, several other examples of the application of Hilbert transform to analyze geophysical signals can be found in Mohan et al. (1982), Sundararajan (1983), and Sundararajan and Srinivas (2010).

Derivation of Time Domain Representation of the Hilbert Transform from the Fourier Transform

Substituting $R(\omega)$ and $I(\omega)$ (see section “Another Definition of Hilbert Transform”) in Equation (25), we get

$$\begin{aligned} \hat{f}(t) &= \frac{1}{\pi} \int_0^{\infty} \left(\int_{-\infty}^{\infty} f(\tau) \cos \omega \tau d\tau \right) \sin \omega t \\ &\quad - \left(\int_{-\infty}^{\infty} f(\tau) \sin \omega \tau d\tau \right) \cos \omega t d\omega \end{aligned} \quad (31)$$

$$= \hat{f}(t) = \frac{1}{\pi} \int_0^{\infty} \int_{-\infty}^{\infty} f(\tau) (\sin \omega(t - \tau)) d\tau d\omega \quad (32)$$

Regrouping the terms and interchanging the order of integration, we have

$$= \hat{f}(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} f(\tau) d\tau \int_0^{\infty} (\sin \omega(t - \tau)) d\omega \quad (33)$$

From standard integrals, we know that

$$\int_0^{\infty} \sin \omega(t - \tau) d\omega = \frac{1}{t - \tau} \quad (34)$$

Substituting Eq. (34) in Eq. (33), we get

$$\hat{f}(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{f(\tau)}{t - \tau} d\tau \quad (35)$$

Note the similarity of Eqs. 23 and 35. Thus Eq. (35) could also be seen as the relation between the Fourier transform and the Hilbert transform.

Role of Nonlinear Signal Processing Techniques in Geophysics

The mathematical transformations that we discussed so far are linear signal analysis techniques. However, since most geophysical data are nonlinear and nonstationary in nature, analyzing them with nonlinear signal analysis techniques will be more appropriate. In this section, we discuss about the role of some nonlinear signal analysis techniques, such as wavelet analysis and nonlinear data-adaptive techniques such as fractal and multifractal analyses and empirical mode decomposition technique.

Wavelet Analysis

The conventional signal analysis technique, Fourier transform, can only provide information about the frequencies and their respective phases present in the signal under investigation. While this is fine, we cannot infer the time variation of frequencies present in the signal with the Fourier transform. In other words, with the Fourier transform, it is impossible to find out, when in time, the frequencies of our interest are occurring in the signal? This is largely due to the limitation of the Fourier transform, as it has cosine and sine functions as the basis functions, which are infinitely supported and cannot be defined in the $\mathcal{L}^2(\mathbb{R})$ domain. As a result, the time localization of frequency components is not possible in Fourier domain. But, why such an information is important in the first place? It is because the frequency variations in the nonlinear signals (rapid or slow) occurring at irregular times manifest the dynamical behavior of the system as a function of time. Therefore, knowledge about the time localization of frequency components is always important for better understanding and characterization of the dynamics of the system generating such nonlinear signals. Therefore, to extract such useful information from nonlinear signals, we need to employ suitable mathematical techniques in their analysis.

Wavelet analysis is one such technique, which provides such vital information. Wavelet analysis facilitates to detect the time-varying frequency components in nonlinear and

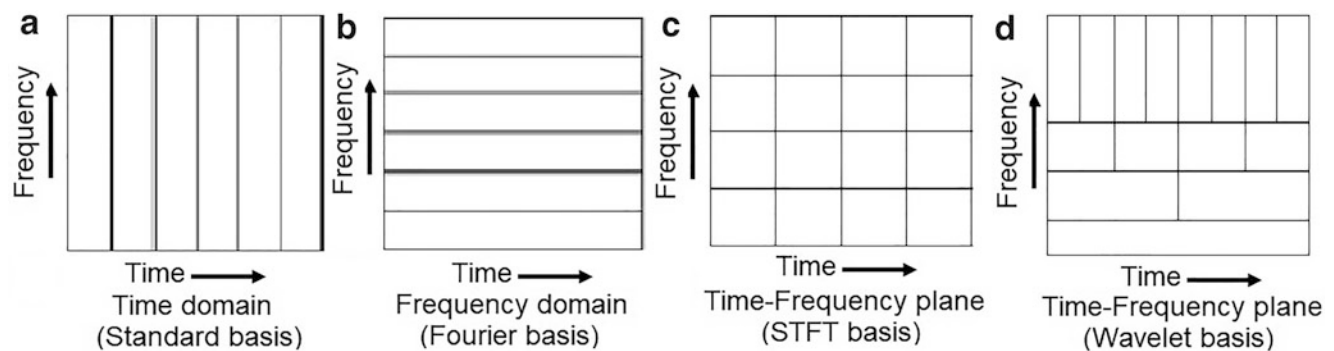
nonstationary signals with required resolution. In other words, if some frequencies of interest occur at irregular times in nonlinear signals, the use of wavelet analysis is ideally suited to detect their temporal behavior. This also facilitates to treat noises in the data as independent entities for their easy identification and processing, thereby enhancing the signal-to-noise ratio in the data.

A simple understanding about how wavelets help in the effective characterization of the signal can be seen as follows:

- From a time domain signal, the time information can be easily obtained but not the frequency information (Fig. 3a). Hence, the time localization is easy in time domain signals.
- Similarly, from the frequency spectrum, the frequency information can be easily obtained but not the time information (Fig. 3b). Hence, the frequency localization is easy to obtain from the frequency spectra.
- If we apply a moving window of fixed length to any signal and compute the Fourier transform of it within that window at every location of its movement, then we get, what is known as the short-time Fourier transform (STFT) or windowed Fourier transform (Gabor 1946) of that signal. In STFT, the idea is to break up the signal into smaller portions with a fixed window length and Fourier analyze each portion of the signal to determine its frequency content. Mathematical representation of the STFT of a time function $f(t)$ and a window function $g(t)$ is given by (Gabor 1946)

$$\text{STFT}(\tau, s) = \int_{-\infty}^{\infty} f(t)g(t-\tau)e^{-i2\pi st} dt$$

In STFT, although the short lengths of data are considered at each step, the main transformation again is FT, whose basis functions, as mentioned above, are again the infinitely supported sine and cosine functions and thus the problem of time-frequency localization with good resolution still persists.



Signal Processing in Geosciences, Fig. 3 Schematic representation of comparison of signals in (a) time domain, (b) Fourier domain, (c) Short-time Fourier transform (STFT) domain, and (d) wavelet domain

Furthermore, because the window length is fixed at each step, the STFT partitions the entire time-frequency plane with cells of equal size and thus similar time-frequency resolution is obtained, as shown in Fig. 3c (see Graps 1995; Mallat 1999).

Since a fixed window length needs to be used in STFT to analyze the entire signal to estimate all the frequencies present in it, the resolution of the frequencies is the same at all locations in the time-frequency plane, which may not be the desired resolution most of the times. This is one of the shortcomings of STFT technique. Further, the Fourier transformation is not ideally suited for analyzing short lengths of data (Burg 1975).

- However, in the case of wavelet analysis, the higher frequencies (smaller scales) are well localized in time, but their frequency localization is poor. On the other hand, the lower frequencies (larger scales) have a very good frequency localization but poor time localization. This can be better visualized in Fig. 3d.

This explains that the wavelet analysis facilitates to visualize the data at different frequency resolutions. If we analyze the signals with larger (smaller) scales of the wavelet function, we get a good resolution of the low (high) frequencies present in the data. Wavelet analysis involves use of a wavelet function, called analyzing wavelet or “mother wavelet”. The chosen mother wavelet is treated with a signal as an inner product in translated (shifted in time) and dilated (compressed or expanded) fashion to determine the time-frequency localization in the data.

It is important to note the following basic difference between the Fourier transform and wavelet transform. In the Fourier transform of a nonlinear signal, the Fourier coefficients of the transformed signal represent the contribution of respective sine and cosine function at each frequency, as the sine and cosine functions are the basis functions for the Fourier transform. In contrast, in the case of wavelet transform, the mother wavelet function itself is the basis function, which is compressed and dilated, while analyzing the signal at different time locations of the signal, thereby facilitating to extract the presence of high and low frequencies in the signal respectively, with good resolution. For example, to identify the sharp discontinuities or spikes or singularities present in the signal, the wavelet function at smaller scales (i.e., in the compressed form) could be used, and to identify the low-frequency features present in the signal, larger scales of the same wavelet function (i.e., in the dilated form) could be used. The scale and frequency are inversely related. Since the mother wavelet is the basis function in wavelet transform and since many wavelets can be defined and used to analyze a variety of nonlinear signals, the wavelet transform has many basis functions.

Basic Theory of Wavelet Transform

A wavelet is a small wave having a finite length in time (in technical terms known as *compactly supported*) and represented as a real or complex-valued function, say, $\psi(t)$, satisfying the following conditions:

$$\int \psi(t) dt = 0 \quad (36)$$

$$\int \psi^2(t) dt = 1 \quad (37)$$

Equation (36), known as *admissibility condition*, implies that the average value of the wavelet function in the time domain must be zero and that the function $\psi(t)$ must be oscillatory (so that the sum of its positive and negative excursions is zero). Eq. (37), known as *regularity condition*, implies that $\psi(t)$ must be finite in length and have finite energy. In other words, this condition signifies the fast decay of the wavelet. Further details about these two conditions can be found in Daubechies (1992) and Burrus et al. (1998).

The wavelet function, $\psi(t)$, signifying the time-frequency localization is defined as

$$\psi_{(\tau,s)}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-\tau}{s}\right) \quad (38)$$

where $s > 0$ indicates the scale (or dilation parameter) and τ indicates the translation parameter. s is analogous to frequency, in the sense that larger scales (low frequencies) provide overall information of the signal and smaller scales (high frequencies) provide detailed information of the signal. The scale-frequency relation is given by $f = \frac{f_c}{s\Delta t}$, where f_c indicates the central frequency of the wavelet and Δt refers to the sampling interval and f indicates the pseudo frequency of the wavelet corresponding to the scale, s . τ refers to the time location of the wavelet window as it is slid over the signal along the time axis of the signal and thus apparently refers to time information in the transformed domain.

If two functions $f(t)$ and $g(t)$ are square integrable in $\mathbb{R}$ (i.e., $f(t), g(t) \in \mathcal{L}^2(\mathbb{R})$), then their inner product is defined as

$$\langle f(t)g(t) \rangle = \int_{\mathbb{R}} f(t) \cdot g^*(t) dt \quad (39)$$

where “*” indicates the complex conjugate. According to (39), the wavelet transform of a function $f(t)$ can be defined as the inner product of the mother wavelet $\psi(t)$ and $f(t)$, given by

$$\text{WTf}(\tau, s) = \frac{1}{\sqrt{s}} \int f(t) \cdot \psi\left(\frac{t-\tau}{s}\right) dt \quad (40)$$

Equation (40) explains that the wavelet transformation gives a measure of the similarity between the signal and the

wavelet function. Such a measure at any particular scale s_0 and translation τ_0 is identified by a wavelet coefficient. The larger the value of this coefficient, the higher the similarity between the signal and the wavelet at τ_0, s_0 and vice versa. If a large number of high wavelet coefficients occur by using a particular wavelet, then that indicates the higher degree of suitability of that wavelet to study the signal under investigation.

Two types of wavelet transforms are continuous wavelet transform (CWT) and discrete wavelet transform (DWT). A brief explanation of both these techniques is provided below.

The Continuous Wavelet Transform (CWT)

In CWT (Eq. 40), the inner product of the signal and the wavelet function is computed for different segments of the data by continuously varying τ and s . Because the wavelet window can be scaled (compressed or dilated) at different levels of analysis, the time localization of high-frequency components of the signal and frequency localization of low-frequency components of the signal can be effectively achieved. Because the wavelet is translated along the time axis of the signal in a continuous fashion, the CWT is ideally suited to identify the jumps and singularities present in the data (Jaffard 1991). In CWT, the wavelet window is first placed at the beginning of the signal. The inner product of the signal and wavelet is computed, and the CWT coefficients are estimated. Next, the wavelet is slightly shifted along the time axis of the signal, and the CWT coefficients are again computed. This process is repeated till the end of the signal is reached. Figure 4a depicts the pictorial representation of the above operation. Next, the wavelet window is dilated (or increased in scale by stretching) by a small amount, and the above process is repeated at that scale, at all translations. Likewise, the CWT coefficients are calculated for several scales and translations. Finally, all the CWT coefficients corresponding to all translations and dilations are expressed in the form of a contour plot in the time-scale plane, known as the *scalogram*. Figure 4b depicts an example of a scalogram, obtained by performing wavelet analysis of gamma ray log data using the Gaus1 wavelet (see Chandrasekhar and Rao 2012). It can be seen in Fig. 4b that the CWT can clearly delineate and distinguish the thin and thick interfaces between different lithologies (seen as alternating sequences of low and high wavelet coefficients) at small and high wavelet scales as a function of depth, representing the high and low frequency features present in the data respectively.

According to Eq. (40), the scalogram signifies that any given one-dimensional time signal is transformed into a two-dimensional time-scale (read as time-frequency) plane, thereby increasing the degrees of freedom to understand the given signal effectively. Thus, the wavelet scalograms facilitate to identify what frequencies are present at what times in

any given timeseries. CWT has been applied to a variety of problems in various branches of science and engineering, such as geophysical fluid dynamics (Farge 1992; Farge et al. 1996), geomagnetism (Alexandrescu et al. 1995; Kunagu et al. 2013; Chandrasekhar et al. 2013, 2015), material science (Gururajan et al. 2013), and electromagnetic induction studies (Zhang and Paulson 1997; Zhang et al. 1997; Garcia and Jones 2008). In the case of well-logging, the wavelet analysis has found its wide applications in effectively describing the inter-well relationship (Jansen and Kelkar 1997), determining the sedimentary cycles (Prokoph and Agterberg 2000), reservoir characterization (Panda et al. 2000; Vega 2003), and the space localization and the depths to the tops of reservoir zones (Chandrasekhar and Rao 2012 and references therein; Hill and Uvarova 2018, Zhang et al. 2018). Chandrasekhar and Rao (2012) have also explained the optimization of suitable mother wavelets through histogram analysis of CWT coefficients.

Translation-Invariance Property of Wavelets in CWT Unlike in the Fourier transform, one of the important properties of wavelets in CWT is their translation-invariance property. It explains that a small amount of shift in the wavelet function will also result in the same amount of shift in the wavelet transformed signal. Mathematically, this can be proved as follows.

Let $f_{t_0} = f(t - t_0)$ be the time-shifted form of the signal $f(t)$, by a small shift, t_0 . The CWT of f_{t_0} is given by

$$\text{CWT}f_{t_0}(\tau, s) = \frac{1}{\sqrt{s}} \int f(t - t_0) \cdot \psi\left(\frac{t - \tau}{s}\right) dt \quad (41)$$

Put $t' = t - t_0$ in (41). Then we have

$$\text{CWT}f_{t_0}(\tau, s) = \frac{1}{\sqrt{s}} \int f(t') \cdot \psi\left(\frac{t' - (\tau - t_0)}{s}\right) dt' \quad (42)$$

$$= \text{CWT}f(\tau - t_0, s) \quad (43)$$

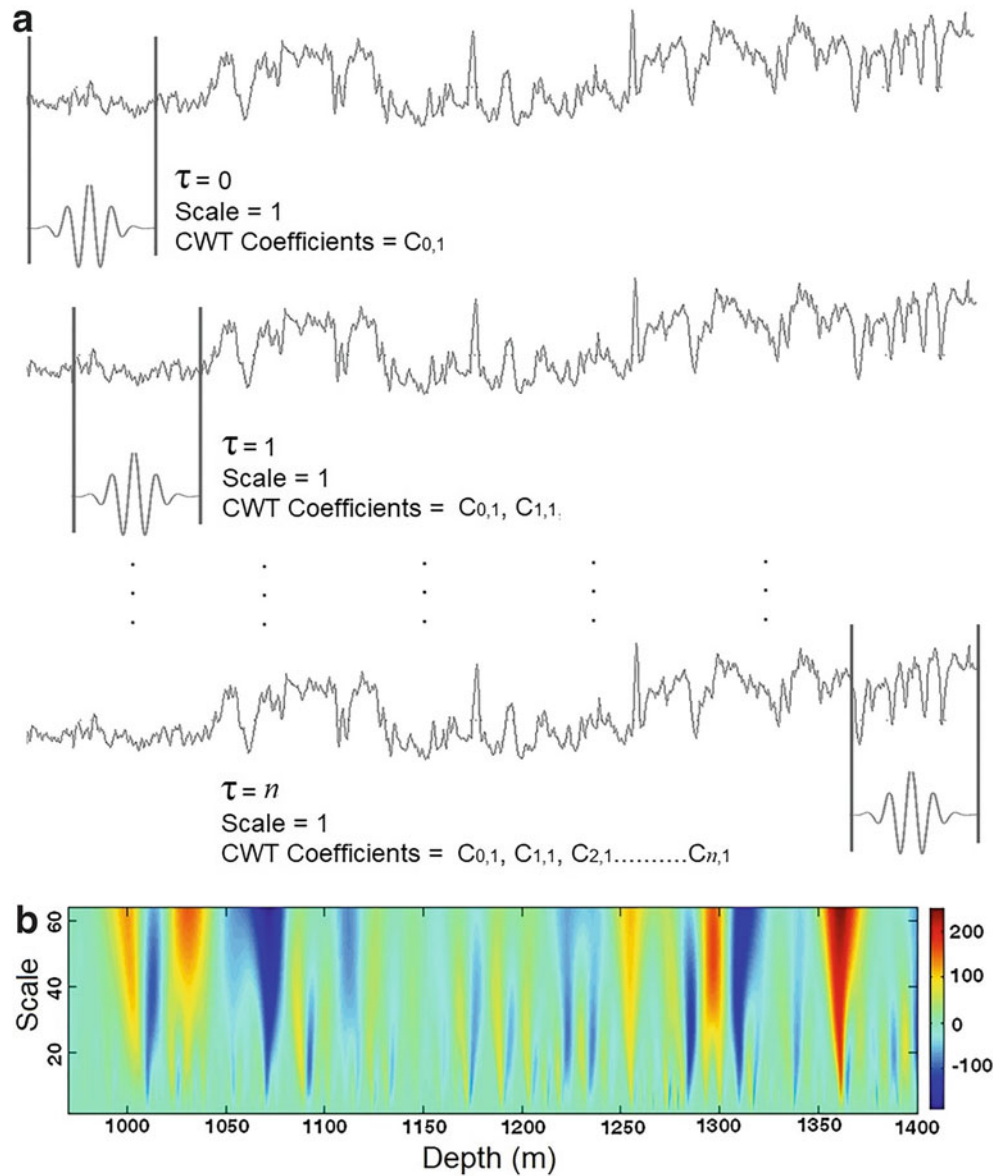
Equation (43) explains that, since the output is also shifted in time by the same amount as in the input, the CWT is translation-invariant. The translation-invariance property plays a vital role in identifying the stable features in shifted images in pattern recognition (Ma and Tang 2001). For various other properties of wavelets, the reader is referred to Mallat (1989, 1999).

Discrete Wavelet Transform (DWT)

In CWT, since the analysis of the entire signal is done by translating and dilating the wavelet in a continuous fashion, the CWT has proven to be very effective in identifying the sudden jumps and singularities in the data (Alexandrescu et al. 1995; Chandrasekhar et al. 2013). Further, in the

Signal Processing in Geosciences, Fig. 4 (a)

A pictorial representation of the CWT operation. First, the CWT coefficients, when the wavelet at translation $\tau = 0$ and dilation $s = 1$, are computed. Next, the coefficients, when the wavelet at translation $\tau = 1$ and dilation $s = 1$, are obtained. Likewise, for the same scale of the wavelet, the coefficients corresponding to all other translations, $C_{2,1}, C_{3,1}, \dots, C_{n,1}$, are obtained. Next, the above steps are repeated for different translations and dilations of the wavelet. Then, a complete scalogram depicting the time-scale representation of the signal under study is obtained (after Chandrasekhar and Dimri 2013). (b) An example of a scalogram obtained by performing CWT on gamma ray log data with Gaus1 wavelet (after Chandrasekhar and Rao 2012). The color scale represents the values of wavelet coefficients in the scalogram



scalogram, the region, where the $WTf(\tau, s)$ is nonzero, is known as the region of support in the τ - s plane. However, such an analysis in continuous fashion also produces a lot of redundant information, and thus the entire support of $WTf(\tau, s)$ may not be required to recover the signal $f(t)$, particularly, when it could be possible to represent any signal with a fewer wavelet coefficients. Therefore, to reduce the redundancy, the discrete wavelet transform (DWT) has been introduced (Daubechies 1988; Mallat 1989). Unlike in CWT, the translation and dilation operations in DWT are done in discrete steps by considering, $s = s_0^k$; $\tau = l\tau_0 s_0^k$, where k and l are integers, known as scale factor and shift factor, respectively. Thus, the DWT representation of Eq. (38) is given by (Daubechies 1988)

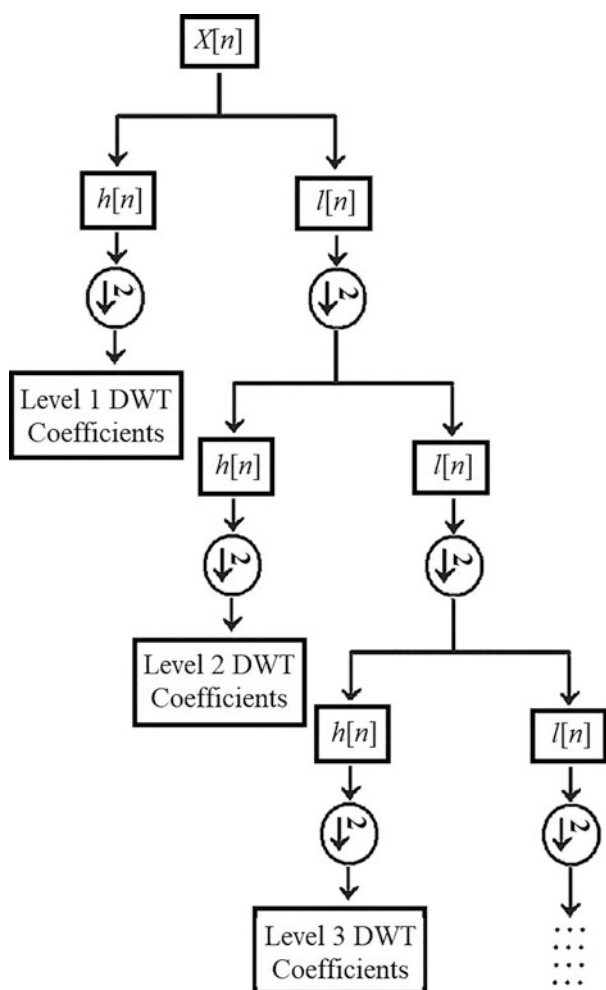
$$\psi_{k,l}(\tau, s) = \frac{1}{\sqrt{s_0^k}} \psi\left(\frac{t - l\tau_0 s_0^k}{s_0^k}\right) \quad (44)$$

The translation parameter, $l\tau_0 s_0^k$, depends on the chosen dilation rate of the scale (Daubechies 1988). For all practical purposes, generally, the values for s_0 and τ_0 are chosen as 2 and 1, respectively, so that the dyadic sampling is obtained both on frequency (scale) and time axes. Since the DWT samples both the time and scale in a dyadic fashion, it is not translation-invariant (see Ma and Tang 2001).

The DWT is computed using Mallat's algorithm (Mallat 1989), which essentially consists of successive low-pass and high-pass filtering steps. First, the signal ($X[n]$) to be wavelet

transformed is decomposed into high-frequency ($h[n]$) and low-frequency components ($l[n]$). $h[n]$ and $l[n]$ are also, respectively, known as detailed and approximate coefficients. While the former are known as level 1 coefficients, the latter are again decomposed into the next level of detailed and approximate coefficients and so on. A pictorial representation of the entire DWT process described above is shown in Fig. 5. However, at this stage, it is important to understand what exactly happens at each level of such decomposition.

The successive decomposition of the signal into detailed and approximate coefficients implies that at the first level of decomposition, the frequency resolution of the detailed coefficients is enhanced, thereby reducing its uncertainty by half. The decimation by two (see Fig. 5) halves the time resolution, as the entire signal after the first level of decomposition is represented by only half the number of samples. Therefore, in the second level of decomposition, while the half-band low-pass filtering removes half of the frequencies, the



Signal Processing in Geosciences, Fig. 5 Schematic representation of the DWT operation in the form a tree. The number “2” in the circles indicates the decimation factor for down sampling. See text for more description of this tree (after Chandrasekhar and Dimri 2013)

decimation by two doubles the scale. With this approach, the time resolution in the signal gets improved at high frequencies, and the frequency resolution is improved at low frequencies. Such an iterative process is repeated until the desired levels of resolution in frequency are reached. However, the important point to note here is that at every wavelet operation, only half of the spectrum is covered. Does this mean an infinite number of operations are needed to compute the DWT? The answer is no. The DWT process needs to be carried out up to a reasonable level, when the required time and frequency resolutions are achieved. The scaling function of the wavelet is used to control this. Let's summarize the above description as follows: In DWT, the given signal is analyzed using the combination of a wavelet function (denoted by $\psi(t)$) and a scaling function (denoted by $\phi(t)$). While the wavelet function helps to calculate the detailed coefficients (high frequency components), the scaling function helps to calculate the approximate coefficients (low frequency components). Finally, the DWT of the original signal is obtained by concatenating all the detailed and approximate coefficients. The correctness of the reconstructed signal will suggest, whether the required levels of decomposition in DWT were properly achieved or not. Thus, the DWT can be viewed as a filter bank operation. More details about DWT and design of wavelets can be found in several literatures such as Daubechies (1988, 1992), Mallat (1989, 1999), Sharma et al. (2013), and Sengar et al. (2013), to name a few.

The discretization of translation and dilation parameters in DWT gives rise to a two-dimensional sequence, $d_{j,k}$, commonly referred to as the DWT of the function $f(t)$. In other words, DWT amounts to mapping of $f(t)$; $t \in \mathbb{R}$ into a two-dimensional sequence, $d_{j,k}$. This implies, any signal, $f(t)$, $\in \mathcal{L}^2(\mathbb{R})$ can be represented as an orthogonal wavelet series expansion if the sampling on the τ - s plane is adequate, thereby reducing the redundancy in the estimation of wavelet transform coefficients (Sharma et al. 2013). Because of such a significant reduction in the redundant computations, DWT offers a wide range of applications, particularly in field of image processing. In geosciences, it is widely used in the fields of remote sensing and geomorphology. The next section provides a brief account of the application of multiresolution analysis of remote sensing images based on DWT.

Multiresolution Analysis of Remote Sensing Images

Remote sensing is a well-known technique used to classify different regions on the earth, based on the amount of electromagnetic radiation reflected from such regions. Changes in the amount and properties of the electromagnetic radiation in any region are valuable sources of data for interpreting the geological and geophysical properties of the region. The amount of electromagnetic radiation depends on the frequency of operation, instantaneous field of view or spatial resolution, radiometric resolution, etc. Data acquired in the

form of images correspond to localized quantization levels of radiometric resolution in the spectral band of each pixel. Although various techniques have been developed to process and analyze the data (see Gonzalez and Woods 2007), it is difficult to analyze and assess the local changes of the intensity in individual pixels in an image. As discussed earlier, given the poor localization properties of the sinusoid basis functions used in the Fourier transform, the Fourier transform is not very effective in extracting the localized features of the image. To overcome this problem, a multiresolution analysis technique, based on DWT, can be implemented, which can provide a coarse-to-fine and scale-invariant decomposition for the interpretation of images (Mallat 1989).

According to Mallat's algorithm, an input image is analyzed with high-pass and low-pass filters using a wavelet function and its associated scaling function, and down-sampling (Fig. 5). Four images (each one with half the size of the original image) are obtained corresponding to high frequencies in the horizontal direction and low frequencies in the vertical direction (HL), low frequencies in the horizontal direction and high frequencies in the vertical direction (LH), high frequencies in both directions (HH), and low frequencies in both directions (LL). The last image (LL) is a low-pass version of the original image and is called as the approximation image. Figure 6a shows a three-level decomposition where LL1 (approximation sub-band at level 1) is further decomposed into LL2, LH2, HL2, and HH2. This procedure is repeated for the approximation image at each resolution 2^j (for dyadic scales).

Wavelets are viewed as the projection of the signal on a specific set of $\psi(t)$ and $\phi(t)$ functions in the vector space, defined as

$$\psi(t) = \sum_{n \in \mathbb{Z}} h[n] \phi(2^s t - \tau) \quad (45)$$

$$\phi(t) = \sum_{n \in \mathbb{Z}} l[n] \phi(2^s t - \tau) \quad (46)$$

where $h[.]$ are high-pass filter coefficients and $l[.]$ are low-pass filter coefficients of filter bank. s and τ , respectively, denote the scaling and translating indexes.

The signal decomposition into different frequency bands is obtained by successive high-pass and low-pass filtering of the time domain signal. The original signal $X[n]$ is first passed through a half-band high-pass filter $h[n]$ and a low-pass filter $l[n]$. After the filtering, half of the samples can be eliminated according to Nyquist's criteria; the signal can therefore be down-sampled by 2, by eliminating every other sample. This constitutes one level of decomposition and can mathematically be expressed as follows

$$\begin{aligned} o_h[t] &= \sum_n X[n] \cdot h[2t - n] \text{ and } o_l[t] \\ &= \sum_n X[n] \cdot l[2t - n] \end{aligned} \quad (47)$$

where o_h and o_l are the outputs of the high-pass and low-pass filters respectively, after down-sampling by two. Eq. (47) is iteratively repeated to obtain DWT coefficients at other levels of decomposition. Figure 6b shows an image of Kuwait City, acquired by the Indian Resourcesat-1 satellite with a spatial resolution of 5.8 m x 5.8 m (609 x 609 pixels). This resolution is well suited for texture analysis, since a spatial resolution of this order is not adequate to extract individual buildings or narrow roads but groups of them, which render a visible checked pattern in dense urban areas. Figure 6c shows the approximate and detailed sub-bands of the original image, where different horizontal, vertical, and diagonal details of urban features can be visualized.

Several methods, including wavelet, curvelet, and contourlet transforms have been used in multiresolution analysis for image enhancement (Ansari and Buddhiraju 2015), texture analysis (Ansari and Buddhiraju 2016), slum identification (Ansari et al. 2020a), and change detection (Ansari et al. 2020b) applications using remotely sensed images.

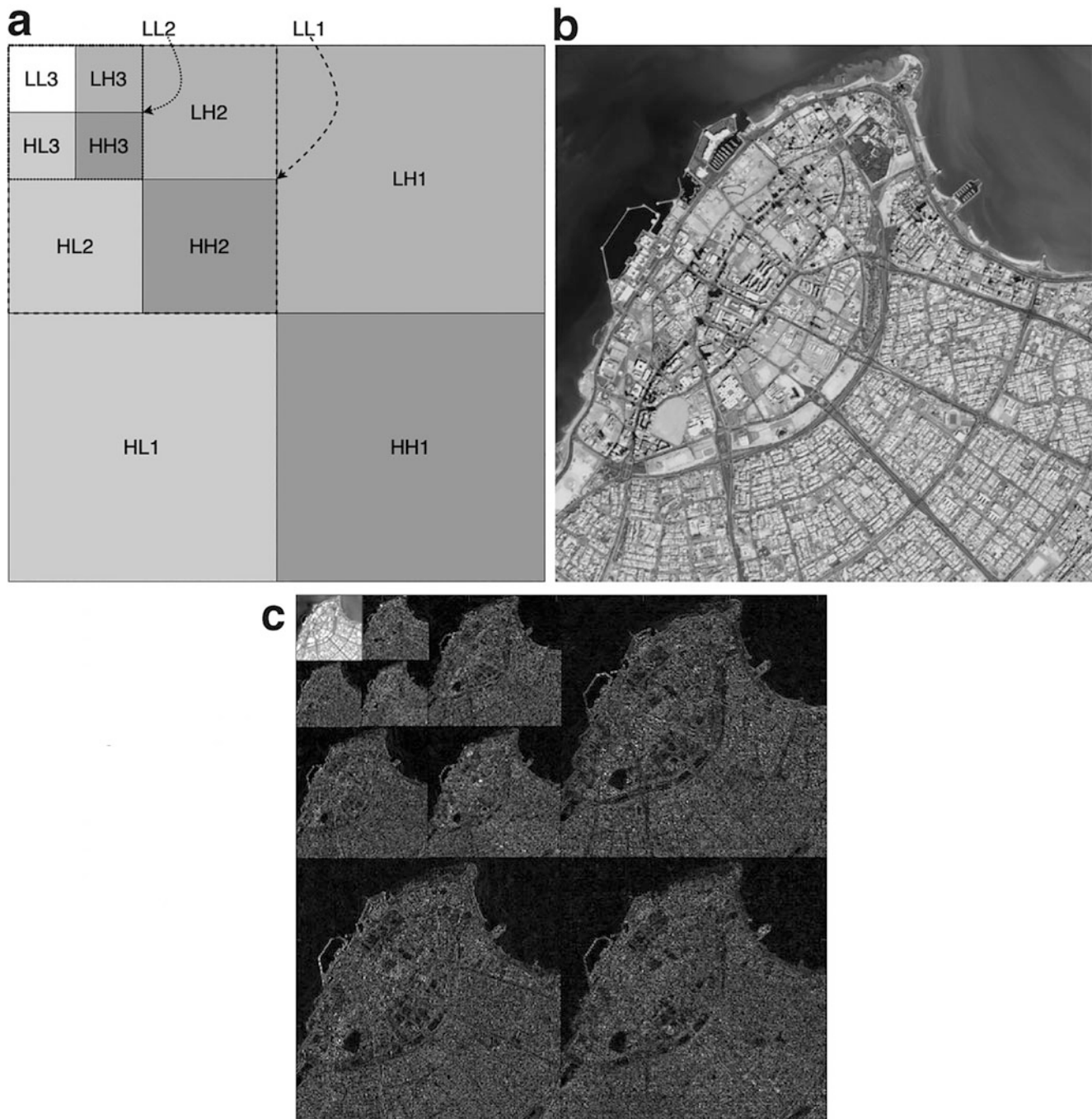
Fractal and Multifractal Analyses

Benoit B. Mandelbrot, who invented fractal theory, first coined the term, *fractal*, to open a new discipline of mathematics that deals with non-differentiable curves and geometrical shapes (Mandelbort 1967). The fractal theory provides a framework to study irregular shapes as they exist, instead of approximating them using regular geometry, which usually is done for the sake of mathematical convenience. A prime characteristic of fractal objects is that their measured metric properties, namely, the length, area, or volume, are a function of the scale lengths of measurement. Mandelbrot, in his landmark paper, entitled, "How long is the coast of Britain?" (Mandelbort 1967), illustrates this property by explaining that finer details of coastline are not well realized at larger scale lengths (lower spatial resolution) but become apparent and thus well recognized only at smaller scale lengths (higher spatial resolution). This implies that the measured length of any irregular or wriggled line increases with a decrease in scale length. This further implies that in fractal geometry, the Euclidean concept of "length" becomes a continuously varying process rather than a fixed entity.

Two types of fractals that we largely deal with in nature are classed as "self-similar fractals" and "self-affine fractals." A fractal is defined as self-similar, if its geometry is retained even after magnification. Mathematically it can be expressed in Cartesian reference frame as

$$x = \lambda^x x' \text{ and } y = \lambda^y y' \quad (48)$$

Examples include cauliflower, a cobweb, fern leaves, cantor set, and Koch curve. On the other hand, a self-affine fractal requires different scaling on either axis, to retain its geometry, in the sense that the variation in one axis scales differently



Signal Processing in Geosciences, Fig. 6 Figure showing (a) the structure of three-level decomposition of DWT, (b) original map of Kuwait City as test data, and (c) three-level wavelet decomposition of original image shown in (b)

from that on the other axis. Mathematically it can be expressed in Cartesian reference frame as

$$x = \lambda^\alpha x' \text{ and } y = \mu^\beta y' \quad (49)$$

where, λ , α , μ and β represent some non-zero constants. Examples include irregular curves and nonlinear timeseries of any dynamical system. The fractal dimension, D , for a self-similar set in $\mathbb{R}^n$ defines N nonoverlapping copies of itself

with each copy scaled down by a ratio $r < 1$ in all coordinates. Mathematically, it is expressed as

$$N = r^{-D} \Rightarrow D = \frac{\ln N}{\ln(1/r)} \quad (50)$$

Although a straight line, a square, and a cube satisfy the self-similar conditions described by Eq. (48), they are not fractals, since they have integer dimensions. Therefore, any

self-similar objects, which have noninteger dimensions only are said to be self-similar fractals. For more details on the estimation of fractal dimension for fractal structures, the reader is referred to Fernandez-Martinez et al. (2019). Fractal dimensions for self-affine irregular shapes and curves can be calculated using several techniques, some of which, we shall briefly discuss below.

It is important to note here that the concept of fractals is not limited to the study of geometrical shapes alone. It has rather wide applications in the study of different signals having a variation of some physical properties as a function of time. Since most geophysical timeseries represent the outcome of the behavior of any dynamical system, responsible for generating such nonstationary and nonlinear signals, they are expected to show self-affine behavior. In the case of timeseries analysis, different magnification factors are needed (along the vertical and horizontal axes), to compare the statistical properties of the rescaled and original timeseries, since these two axes represent different physical parameters: amplitude and time. Let us explain this further, as given below.

Consider a nonlinear data of length l_2 (Fig. 7a). A portion of the same data having length l_1 is expanded to a length, equal to l_2 (Fig. 7b). Now calculate the probability distribution of the data corresponding to window lengths l_1 and l_2 , and plot them as shown in Fig. 7c. s_1 and s_2 indicate the half-widths of the probability distribution curves corresponding to the data of lengths l_1 and l_2 , respectively. The relationship

between s_1, s_2 and l_1, l_2 is expressed by a parameter called, the *scaling exponent*, denoted by α , is given by

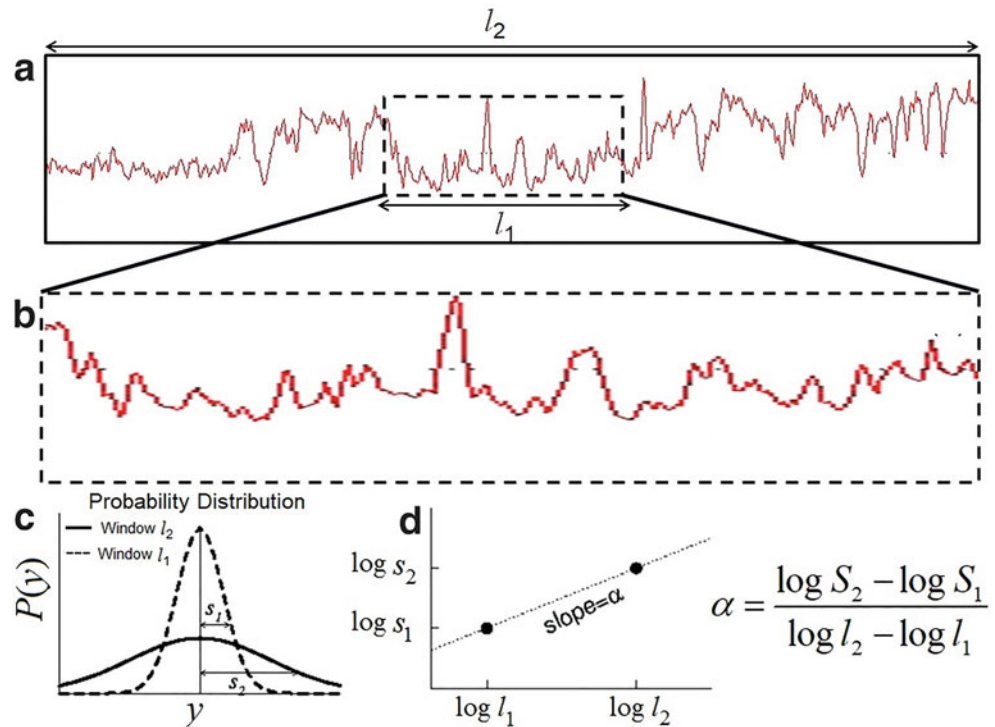
$$\frac{s_2}{s_1} = \left(\frac{l_2}{l_1}\right)^\alpha \quad (51)$$

The scaling exponent is estimated as shown in Fig. 7d. Likewise, α is calculated for different portions of the data with varying window lengths. If the α value is observed to be constant, indicating a steady increasing trend with the increasing window lengths, then that data is said to possess *monofractal* behavior. Alternatively, if the α values show different trends, with increasing window lengths, then that data is said to possess *multifractal* behavior.

Most observational data (generated by the dynamics of the unknown systems producing such nonlinear and nonstationary data) cannot be characterized by a single exponent throughout its length, and thus more than one scaling exponent are obtained for different segments within the same data based on the temporal variations in the statistical properties of the data. Such signals show multifractal behavior. Examples include seismological data, ionospheric data, geomagnetic data, and weather data, to name a few. While it is true that sometimes the nature of the data itself warrants its multifractal behavior, in most other cases, the effect of various types of noises that grossly vitiate the data, make it “behave” like a multifractal.

Signal Processing in

Geosciences, Fig. 7 Schematic representation of estimation of scaling exponent for a nonlinear data. (a) example nonlinear signal, (b) a portion of the signal expanded to be equal to the length of the original signal, (c) estimation of half-widths s_1 and s_2 of probability distribution function, $P(y)$ for the data corresponding to l_1 and l_2 windows, and (d) calculation of the scaling exponent, α



A few techniques to determine the scaling exponents for nonlinear and nonstationary data are (i) rescaled range (R/S) analysis, (ii) detrended fluctuation analysis (DFA) and (iii) multifractal DFA, and (iv) wavelet transform modulus maxima (WTMM) method. Let us briefly discuss each of them. For a full review of several other fractal and multifractal techniques, the reader is referred to Lopes and Betrouni (2009).

Rescaled Range (R/S) Analysis

This is also known as R/S technique, where R and S , respectively, denote the range and standard deviation of the integrated series generated from the given signal. This method basically relies on the fact that R and S bear a power-law relationship, defined as

$$\frac{R(k)}{S(k)} = k^H \quad (52)$$

where k designates the time length of the segment of the input signal, say $x(t)$. The integrated timeseries, y_k , of $x(t)$ is estimated as

$$y_k = \sum_{i=1}^k [x(i) - \bar{x}] \quad k = 1, 2, 3, \dots, N \quad (53)$$

The range, $R(k)$, defines the difference between the maximum and minimum of y_k , given by

$$R(k) = \max(y_0, y_1, y_2 \dots y_k) - \min(y_0, y_1, y_2 \dots y_k) \quad (54)$$

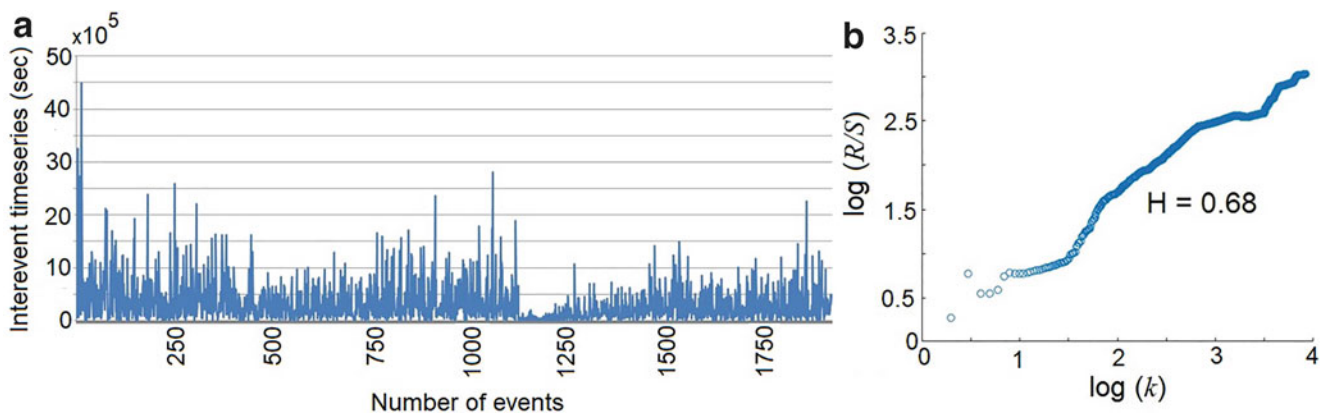
The standard deviation, $S(k)$, of the integrated series y_k , over the period k , is given by

$$S(k) = \frac{1}{k} \sqrt{\sum_{i=1}^k [y_i - \bar{y}]^2} \quad (55)$$

H in Eq. (52) is known as the *Hurst exponent*. The slope of the curve in the plot of the logarithm of the ratio of $R(k)$ to $S(k)$ Vs. the logarithm of k defines H . Figure 8a shows an example of interevent timeseries data generated from the earthquake magnitude (Mw) data (source: <http://www.isc.ac.uk/iscbulletin/search/catalogue/interactive/>) in the range, 4.0–4.9 of Japan region corresponding to a period between 1990 and 2016. Figure 8b shows the least-squares estimation of Hurst exponent value of 0.68, calculated using R/S analysis technique for the same data set, signifying the presence of persistent long-range correlations in the data. The Hurst exponent, H , is understood in the following way. (i) If $H > 0.5$, then the system behavior is called persistent. This suggests, an increasing (decreasing) trend in the past implies an increasing (decreasing) trend in the future. This is true for any natural phenomena, as in Fig. 8b. (ii) If $H < 0.5$, then the system behavior is called anti-persistent. This suggests, an increasing (decreasing) trend in the past implies a decreasing (increasing) trend in the future. Finally, (iii) if $H = 0.5$, then the correlation of the past and future increments vanishes at all times as required for an independent process, and thus the data represents the pure white noise series. More details about the Hurst exponent can be read from Bodruzzaman et al. (1991) and Gilmore et al. (2002).

Detrended Fluctuation Analysis (DFA) and Multifractal DFA (MFDFA)

DFA and MFDFA help to study the intrinsic self-similarities in nonstationary and nonlinear signals, by determining the scaling exponents in a modified least-square sense. In other words, they help to understand the order in the chaos. To estimate the scaling exponents through both these techniques, first the given signal is converted to a self-similar process by integration. This would help to bring out the power-law behavior in the signal, by making it unbounded, and detects the underlying self-similarities in the data over various



Signal Processing in Geosciences, Fig. 8 The test data (a) of interevent timeseries of earthquake magnitude (Mw) in the range, 4.0–4.9 of Japan region corresponding to a period between 1990 and 2016 and (b) Hurst exponent value calculated using R/S analysis technique

window lengths (Goldberger et al. 2000). Next, the integrated series is divided into short windows of equal length, and the local trend (a least-squares fit of chosen order) of the data corresponding to each window is calculated and removed from each data point of the respective window, and a fluctuation function corresponding to each window is calculated. This exercise is repeated for various chosen window lengths of the data. It is important to note here that after estimating the fluctuation functions with one window length, the successive window lengths are incremented by a factor of $\sqrt[3]{2}$, with the maximum window length not exceeding $N/4$, where N is the total number of data points in the signal under investigation (Peng et al. 1994). The fractal scaling exponent then signifies the slope of the linear least-squares regression between the logarithm of fluctuations and the logarithm of window lengths. In MFDFA, the above procedure is carried out repeatedly for different orders (also known as statistical moments) of fluctuation functions in a modified least-squares sense. The entire description of DFA and MFDFA techniques in simple mathematical steps is described below (see Kantelhardt 2002; Kantelhardt et al. 2002; Subhakar and Chandrasekhar 2016; Chandrasekhar et al. 2016):

1. First determine an integrated series $y(m)$ of N -point input data sequence, say $x(i)$, by

$$y(m) = \sum_{i=1}^m (x(i) - \bar{x}); m = 1, 2, 3, \dots, N \quad (56)$$

where, $\bar{x}$ denotes the mean of the N -point data sequence.

2. Divide the m -length integrated series into various m/k nonoverlapping windows of equal length, each consisting of k number of samples.

3. Calculate the least-squares fit of preferred order to the data points in each window. This represents the local trend, $y_k(m)$.
4. Detrend the integrated series, $y(m)$, by subtracting the local trend, $y_k(m)$, from each data point of the corresponding window.
5. For a window of length, k , the average fluctuation, $F(k, n)$, of the detrended series is calculated by

$$F(k, n) = \sqrt{\frac{1}{k} \sum_{i=s+1}^{nk} [y(i) - y_k(i)]^2} \quad (57)$$

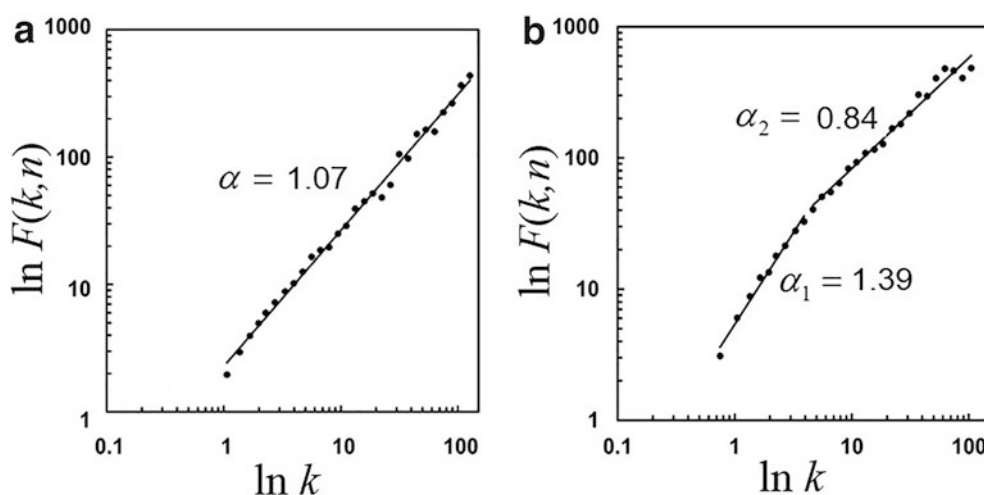
where $s = (n - 1)k$; $N_k = \text{int}(m/k)$ and n denotes the window number.

6. Equation (57) is calculated for various window lengths, k . The power-law relation between $F(k, n)$ and k is given by $F(k, n) = k^\alpha$. The scaling exponent, α is estimated by

$$\alpha = \ln F(k, n) / \ln k \quad (58)$$

If α bears a linear relation between $\ln F(k, n)$ and $\ln k$, such that it steadily increases with increase in window length, for the entire data, then the data is said to possess *monofractal* behaviour. If α bears more than one linear relation between $\ln F(k, n)$ and $\ln k$, corresponding to a group of window lengths, as k increases, then the data is said to possess *multifractal* behavior. In other words, the data possessing multifractal behavior shows persistent short-range correlations up to a certain group of shorter window lengths and persistent long-range correlations for a group of larger window lengths. Fig. 9 depicts the examples of two different nonlinear mono and multifractal signals characterized by a single scaling exponent (Fig. 9a) and two scaling exponents (Fig. 9b), respectively.

Signal Processing in Geosciences, Fig. 9 Examples of two different nonlinear mono and multifractal signals characterized by (a) single scaling exponent and (b) two scaling exponents respectively (after Subhakar and Chandrasekhar 2015)



7. In MF DFA calculations, Eq. (57) is calculated by operating the windows in both forward and backward directions. For forward operations, n is considered as $n = 1, 2, 3, \dots, N_k$ and for backward operations, n is considered as, $n = N_k, N_k - 1, N_k - 2, \dots, 1$. Considering both forward and backward operations, the generalized form for overall average fluctuations is expressed as q^{th} order fluctuation function as (Kantelhardt et al. 2002).

$$F_q(k) = \left\{ \frac{1}{2N_k} \sum_{n=1}^{2N_k} [F^2(k, n)]^{\frac{q}{2}} \right\}^{\frac{1}{q}} \approx k^{h(q)} \quad (59)$$

Equation (59) is iteratively calculated for various k and q values to provide a power-law relation between $F_q(k)$ and $k^{h(q)}$. The Hurst exponent, $h(q)$, defined as

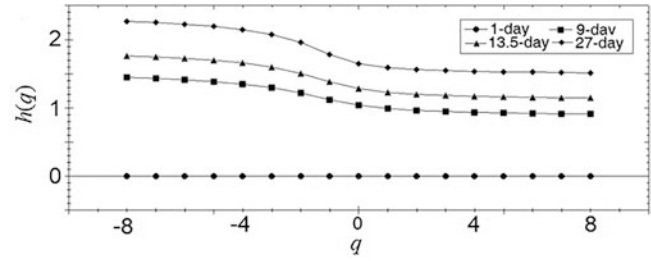
$$h(q) = \ln F_q(k) / \ln k \quad (60)$$

expresses the slope of the linear least-squares regression between the logarithm of the overall average fluctuations $F_q(k)$ and the logarithm of k , for corresponding q . q signifies the statistical moment, defined as mean ($q = 1$), variance ($q = 2$), skewness ($q = 3$), kurtosis ($q = 4$), etc. However, for better characterization of the signal, Eq. (60) is calculated for further upper and lower bounds of q , such that q can be negative also. Generally, q values considered in the range -8 to 8 are sufficient to meaningfully characterize any non-linear signal. In the range, $-q$ to $+q$, for the case of $q = 0$ (as the exponent becomes divergent at this value), Eq. (59) will be transformed to a logarithmic averaging procedure, given by (Kantelhardt et al. 2002).

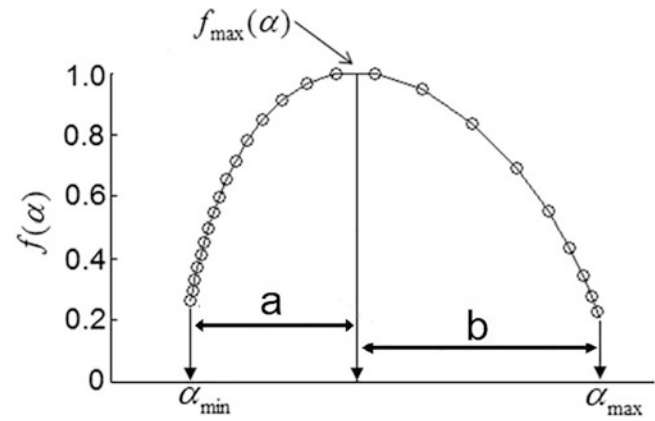
$$F_0(k) = \exp \left[\frac{1}{4N_k} \sum_{n=1}^{2N_k} \ln \{F^2(k, n)\} \right] \approx k^{h(0)} \quad (61)$$

The behavior of multifractal Hurst exponent, $h(q)$, depends on the average fluctuations, $F_q(k)$. It bears a nonlinear relation with the order of the fluctuation function, q , such that the average fluctuations have high (low) $h(q)$ for negative (positive) q (Kantelhardt et al. 2002). If $h(q)$ varies (remains constant) for various q , then the signal is said to have multifractal (monofractal) behavior. This can be easily observed in Fig. 10, which depicts the q Vs $h(q)$ plot, obtained using synthetic data consisting of several periodicities (after Chandrasekhar et al. 2016). In Fig. 10, while the 1-day periodicity signal depicts monofractal behavior, the signals of all other periodicities exhibit multifractal behavior.

8. Finally, estimate the multifractal singularity spectrum, defining the relation between the singularity spectrum, $f(\alpha)$ and strength of the singularity, α , defined by



Signal Processing in Geosciences, Fig. 10 The q Vs $h(q)$ plot, obtained using synthetic data consisting of several periodicities, obtained using Eqs. 59, 60, and 61 (after Chandrasekhar et al. 2016)



Signal Processing in Geosciences, Fig. 11 A schematic representation of an example of multifractal (MF) singularity spectrum, depicting several parameters that aid in clear understanding of the spectrum. See the text for better understanding of the MF spectrum (after Chandrasekhar et al. 2016)

$$f(\alpha) = q[\alpha - h(q)] + 1 \text{ and } \alpha = h(q) + qh'(q) \quad (62)$$

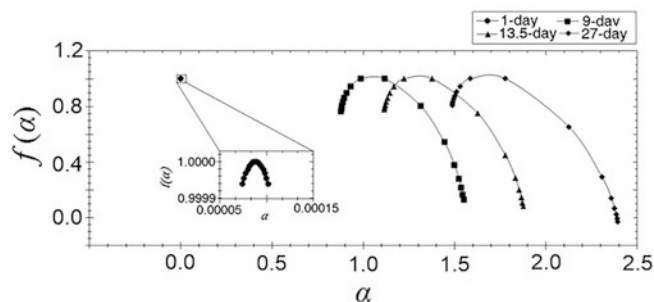
α is also known as Hölder exponent. The reader is referred to Kantelhardt et al. (2002) for more details on α and $f(\alpha)$. Fig. 11 shows an example of $f(\alpha)$ versus α plot, known as multifractal singularity spectrum, depicting different parameters required to interpret the data. $\alpha_{\max} - \alpha_{\min}$ defines the width of the multifractal spectrum. The broader the width of the singularity spectrum, the stronger the multifractal behavior of the nonlinear data and vice versa (Kantelhardt et al. 2002). Also, in Fig. 11, the parameters **a** and **b** signify the distances of the extrema from the center of the spectrum. If **b** > **a**, then the spectrum is said to be right-skewed and if **a** > **b**, then the spectrum is said to be left-skewed. The right-skewed spectrum signifies the presence of finer structures in the data, and the left-skewed spectrum signifies the presence of coarser structures in the data. One easy way to understand this concept is as follows: If the multifractal singularity spectrum of the ECG signal of a heart patient is right-skewed, then the heart functioning is said to be good, indicating the

systematic heartbeat of the patient. However, if the spectrum is left-skewed, then the heart functioning is said to be not good, indicating the presence of large-scale and irregular fluctuations in the heartbeat, suggesting medical attention for that patient. This way, multifractal analysis helps to understand and interpret the factors contributing to the dynamics of the system, generating the nonlinear data.

Another important point to understand about the multifractal singularity spectrum is its positioning in the α Vs. $f(\alpha)$ plot. Calculating the multifractal spectra of a synthetic time series (the same synthetic data used for calculating $h(q)$ for Fig. 10), Chandrasekhar et al. (2016) have explained (see Fig. 12) that the positioning of multifractal singularity spectrum in the α Vs. $f(\alpha)$ plot corresponding to any periodicity in any given data is dictated by the number of cycles of that periodicity present in the data. If a fewer number of cycles of any periodicity are present in the data, then the multifractal spectrum corresponding to the data of that periodicity will be positioned at higher values of α and vice versa.

Applications The Hurst exponent, $h(q)$ (Eq. 60), plotted as a function of q , and the multifractal singularity spectra (Eq. 62), depicting the $f(\alpha)$ versus α plot, are the two important diagnostic information used in multifractal analysis to understand the fractal/multifractal behavior of the nonlinear and nonstationary data.

Telesca et al. (2004) applied the MFDFA algorithm on seismological data recorded for several years at three tectonically active zones in Italy. Multifractal studies helped them to easily demonstrate the differences in the subsurface dynamics in all the different zones. They also have compared their results with those of R/S analysis performed on the same data sets. Telesca et al. (2012) have also studied the environmental influences such as ocean effect and coast effect, on the magnetotelluric data recorded at different sites in Taiwan, using the MFDFA technique.



Signal Processing in Geosciences, Fig. 12 The α Vs $f(\alpha)$ plots, depicting the comparison of the position locations of the multifractal singularity spectra of different periodicities for the same synthetic data used in Fig. 10, obtained using Eq. (62) (after Chandrasekhar et al. 2016)

Subhakar and Chandrasekhar (2015, 2016) applied both DFA and MFDFA algorithms on geophysical well-log data of Bombay offshore basin (the same data used by Chandrasekhar and Rao (2012) for wavelet analysis) and identified the depths to the tops of formations and compared their results of DFA with those of wavelet analysis. Subhakar and Chandrasekhar (2016) also distinguished the oil and gas zones in the subsurface hydrocarbon reservoir zones, based on the multifractal behavior of well-log data. Further, they also explained whether the multifractal behavior in well-log signals is due to the presence of long-range correlations or the broad probability distribution in the data (Subhakar and Chandrasekhar 2015). Chandrasekhar et al. (2016) applied the MFDFA algorithm to ionospheric total electron content (TEC) data of a particular longitude zone covering both the hemispheres to understand the hemispherical differences in the spatiotemporal behavior of TEC and also the multifractal behavior of TEC recorded during the geomagnetically quiet and disturbed periods. Multifractal studies facilitated them to clearly demonstrate the semi-lunar tidal influences on TEC data as a function of latitude. Chandrasekhar et al. (2015) studied the multifractal behavior of geomagnetic field corresponding to geomagnetic quiet and disturbed days. They suggest that more geomagnetic data sets corresponding to different seasons would provide more comprehensive information on the multifractal behavior of the extraterrestrial source characteristics influencing the geomagnetic field variations during geomagnetic disturbed days. For other applications of multifractal analysis in the field of geomagnetism, the reader is referred to Hongre et al. (1999), Yu et al. (2009), and the references therein.

Wavelet-Based Multifractal Formalism: The Wavelet Transform Modulus Maxima (WTMM) Method

This method allows the determination of multifractal behavior of nonlinear dynamical systems using wavelets. This technique relies on the wavelet coefficients estimated from CWT. In this method, if the chosen mother wavelet has n vanishing moments, it will nullify all the polynomial trends of the order $n-1$ in the data (Daubechies 1992), leading to a natural detrending process (Vanishing moments of a wavelet signify the ability of the wavelet to suppress the polynomial trends in the given data. Higher vanishing moments imply the wavelet's ability to suppress the higher order trends in the data and vice versa.). This facilitates to accurately identify the singularities in the data. The “modulus” of the wavelet coefficients, obtained along the lines of maxima, can easily bring out the multifractal behavior of the data. (Lines of maxima are the curves formed by joining the points of local maxima at different scales in the wavelet scalogram (cf. Section “The Continuous Wavelet Transform (CWT)”). The convergence of lines of maxima at the lowest scale on the time axis

indicates the presence of singularities in the signal at those times). Mathematical details of the WTMM technique are as follows.

The multifractal behavior of a function, $f(t)$, around any point is characterized by the scaling properties of its local power law, given by

$$\langle |f(t+k) - f(t)| \rangle \approx k^{\alpha(t)} \quad (63)$$

The exponent $\alpha(t)$ describes the local degree of singularity or regularity around t . The collection of all the points that share the same singularity exponent is called the singularity manifold of exponent α and is a fractal set of fractal dimension $D(\alpha)$. The relation between $D(\alpha)$ and α signifies the multifractal singularity spectrum that fully describes the distribution of the singularities in $f(t)$.

The CWT of a function is defined by

$$Wf(u, s) = \langle f, \psi_{u,s} \rangle = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} f(t) \cdot \psi^*\left(\frac{t-u}{s}\right) dt \quad (64)$$

The presence of the local singularity of $f(t)$ in a scalogram can be checked by the decay of $|Wf(u, s)|$ at t_0 on the time axis. The wavelet transform modulus maxima at a point (u_0, s_0) describes that $|Wf(u, s)|$ is locally maximum at $u = \pm u_0$, such that

$$\frac{\partial Wf(u_0, s_0)}{\partial u} = 0 \quad (65)$$

The modulus of the wavelet coefficients along the lines of maxima bears a power-law relationship with the scale corresponding to each identified singularity in the signal (Mallat 1999). Let (u_p, s) be the position of all local maxima of $|Wf(u, s)|$ at a fixed scale s . The global partition function, $Z(q, s)$, signifying the measure of the sum at a power q of all these wavelet modulus maxima is given by (Arneodo et al. 2008).

$$Z(q, s) = \sum_p |Wf(u_p, s)|^q \quad \text{where, } Z(q, s) \sim s^{\tau(q)} \quad (66)$$

q defines the statistical moment. The multifractal singularity spectrum from the wavelet transform local maxima can be obtained using $Z(q, s)$. The modulus of the wavelet transform and the set of maxima locations corresponding to scale s is given by (Mallat 1999)

$$|Wf(u, s)| \sim S^{(\alpha+1/2)} \text{ when } s \rightarrow 0^+ \quad (67)$$

From Eq. (66), the scaling exponent $\tau(q)$ can be obtained by

$$\log_2 Z(q, s) \sim \tau(q) \log_2 s \quad (68)$$

Finally, the multifractal singularity spectrum from the WTMM method is obtained by (Bacry et al. 1993).

$$D(\alpha) = \min_{q=R} [q(\alpha + 1/2) - \tau(q)] \quad (69)$$

For further illustration of the WTMM, the reader is referred to Arneodo et al. (2002, 2008) for numerous examples, Kantelhardt et al. (2002), Audit et al. (2002), and Salat et al. (2017) for comparison with the MDFA method. Using the cumulants of WTMM coefficients, Bhardwaj et al. (2020) studied the multifractal behavior of the ionospheric total electron content (TEC) data and established the consistent results obtained by Chandrasekhar et al. (2016) using MF DFA on the same data sets.

Empirical Mode Decomposition (EMD) Technique

Since most geophysical data are nonstationary and nonlinear in nature, the use of wavelet transformation, instead of the Fourier transform, has been proved to be one of the efficient techniques for characterizing the spatiotemporal nature of a variety of signals. However, the inherent problem with wavelet analysis lies with the identification of a suitable mother wavelet that best suits to analyze the signals. Besides wavelet analysis and the data-adaptive fractal and multifractal studies, another efficient and fully data-adaptive signal analysis technique, called the empirical mode decomposition (EMD) technique, can unravel the hidden information from the nonlinear signals. EMD technique, combined with Hilbert spectral analysis, is known as Hilbert-Huang transform (see Huang et al. 1998; Huang and Wu 2008). The EMD technique decomposes the data into oscillatory signals of different wavelengths, known as intrinsic mode functions (IMFs), just the similar way the discrete wavelet transform (DWT) decomposes the signal into detailed and approximate wavelet coefficients, representing the high- and low-frequency components of the signal respectively. However, the advantage of EMD technique over the DWT is that in the former, the decomposition of the signal into IMFs is easily done through some simple sifting operations, which are totally data-adaptive, unlike in the latter, where a similar operation necessitates an arduous task of choosing suitable wavelet and scaling functions. Although the EMD technique lacks a thorough mathematical framework, it is proved to be best suited to analyze nonlinear and nonstationary signals (Huang et al. 1998; Huang and Wu 2008; Gairola and Chandrasekhar 2017) and also acts as an efficient filter bank (Flandrin et al. 2004).

Methodology to Estimate the IMFs by EMD Technique

The end result of the fully data-adaptive EMD technique is the decomposition of different frequency components of the

signal (from highest to lowest) till the signal becomes monotonic, which is called the residue and represents the overall trend of the signal. The sequential steps to determine IMFs are as follows:

1. Identify the local maxima and minima of the given signal, $x(t)$, with a cubic-spline interpolation of them to form upper and lower envelopes, respectively, and calculate their mean, m_1 .
2. Calculate the first proto-mode, x_1 given by $x_1 = x(t) - m_1$.
3. If x_1 still contains maxima and minima, repeat steps 1 and 2 to get $x_{11} = x_1 - m_{11}$. Accordingly, after k iterations, we shall have, $x_{1k} = x_{1(k-1)} - m_{1k}$. According to Huang et al. (1998), the stopping criteria for this iterative process are (i) the number of extrema and number of zero crossings must be equal to or differ at most by 1. (ii) m_1 should be zero for the entire signal. Condition (i) ensures that the signal under investigation should not contain any local fluctuations and condition (ii) ensures that the Hilbert transform estimates the correct instantaneous frequencies from the data.
4. Choose the stopping criterion as the normalized squared difference of the data after two successive sifting operations, given by

$$D_k = \frac{\sum_{i=1}^N (x_{k-1}(i) - x_k(i))^2}{\sum_{i=1}^N (x_{k-1}^2(i))}$$

where N denotes the total number of data samples. Generally, D_k should be as small as possible, say $D_k \leq 0.1$, for correct estimation of each IMF after k iterations. Once the stopping criterion is satisfied after the first set of k iterations, then the first IMF (IMF-1), signifying the highest frequency present in the data is determined. It is important to note here that the condition, $D_k \leq 0.1$, is rather arbitrary and largely depends on the data quality. Higher D_k limits can be set for poor quality data and vice versa (Gairola and Chandrasekhar 2017).

5. Calculate the residue, r_1 , by subtracting IMF-1 data from the original signal and repeat steps 1–4 above to determine IMF-2.
6. Repeat steps 1–5 to calculate the successive IMFs until the residue becomes monotonic, when no further IMFs could be extracted from the data. The original signal can be obtained by synthesizing all the IMFs and the residue, given by $x(t) = \sum_{i=1}^{n-1} \text{IMF}_i + r_n$, where n denotes the n^{th} number of iteration, at which r_n becomes monotonic.

Figure 13 depicts an example of a nonlinear test signal and the IMFs obtained by EMD technique. EMD technique has

found several applications in a variety of specializations in geophysics, such as magnetotellurics (Cai et al. 2009; Cai 2013; Neukirch and Garcia 2014), well-logging (Gairola and Chandrasekhar 2017, Gaci and Zaourar (2014), seismics (Battista et al. 2007; Xue et al. 2013; Jayaswal et al. 2021), fault detection in mechanical devices (Gao et al. 2008), atmospheric sciences (McDonald et al. 2007), and sequence stratigraphy (Zhao and Li 2015), to name a few. Dätig and Schlurmann (2004) discussed the limitations of HHT, while applying it to study irregular water waves. Gaci and Zaourar (2014) and Gairola and Chandrasekhar (2017) have also discussed the use of IMFs in quantifying the degree of subsurface heterogeneity from nonlinear signals.

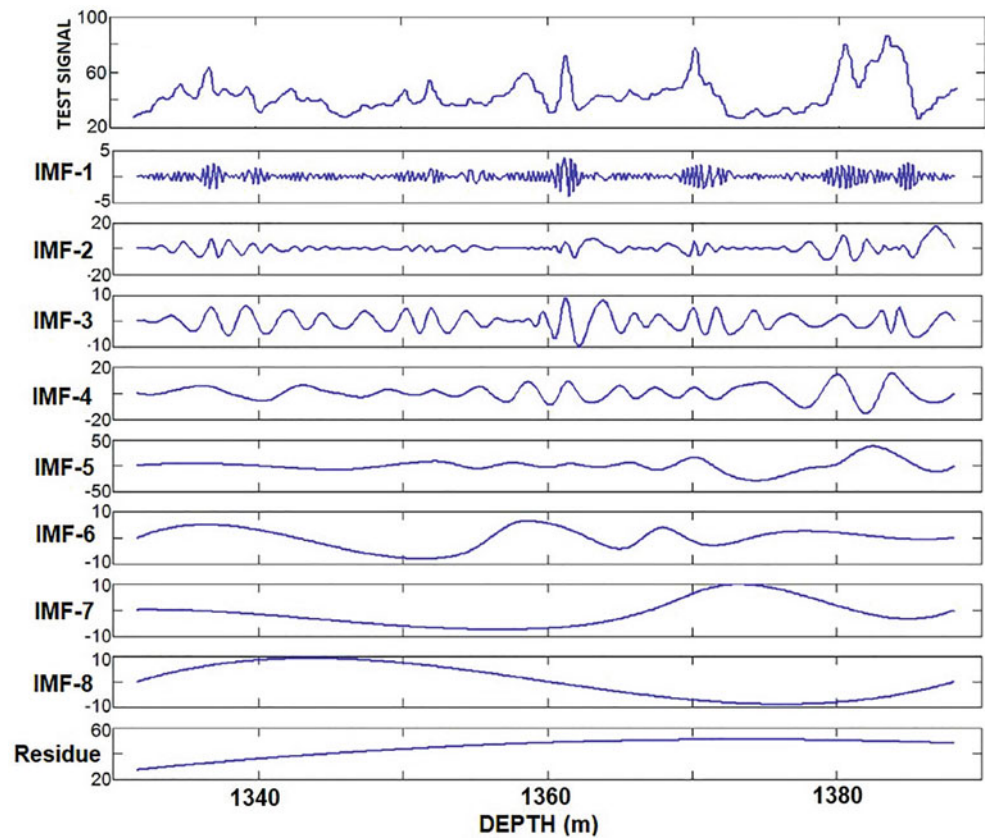
The EMD technique suffers with the problem of spilling over of frequencies from one IMF into other IMFs, called *mode mixing*. Mode mixing mainly arises due to overshoot and undershoot in cubic spline interpolation of upper and lower envelopes, while estimating the IMFs. To address this problem, a minor variant of the EMD technique, called the ensemble EMD (EEMD), technique has been developed (Wu and Huang 2009; Torres et al. 2011; Gaci 2016). Basically, in EEMD, different white noise series are added into the signal at several sifting operations. Since the noise added at each operation is different, the resulting IMFs do not display any correlation with their adjacent ones, and thus spilling of frequencies from one IMF to another one can be arrested. However, the principle of all these techniques remains the same as that of EMD technique described above. Also, the applications of EEMD technique in geosciences have been very limited as of now.

Application of IMFs in Characterizing the Degree of Sub-surface Heterogeneity If two nonlinear signals are generated from two different heterogeneous systems, then generally it is natural to believe that the signal producing more number of IMFs must have been generated from a system having a higher degree of heterogeneity than the other, thereby qualitatively understanding the heterogeneity of the two systems generating such signals. However, while this may be true, sometimes, it is not an easy task to assess the heterogeneity in different systems, when signals generated from two different dynamical systems produce equal number of IMFs. In such situations, it becomes difficult to establish, which system is more heterogeneous than the other. A better understanding of the dynamics of such systems, generating such nonlinear signals, can be best achieved by carrying out the heterogeneity analysis (see Gaci and Zaourar 2014).

Heterogeneity analysis of the IMFs of nonlinear signals provides a quantitative estimate of the degree of heterogeneity of the dynamics of the system producing such signals. We earlier have explained that the EMD technique acts as a filter bank (Flandrin et al. 2004), just the similar way the discrete wavelet transform does (Mallat 1999). Therefore, according

Signal Processing in

Geosciences, Fig. 13 Schematic representation of a nonlinear test signal and the IMFs generated from it. Note that the first IMF (IMF-1) designates the highest frequency and the last IMF (IMF-8) depicts the lowest frequency present in the signal. The bottom most curve depicts the residue, which is a monotonic function (see the text), which cannot be decomposed into any further IMFs (after Gairola and Chandrasekhar 2017)



to the filter bank theory, there must exist a nonlinear relationship between the IMF number (m) and the respective mean wavelength (I_m) of each IMF, given by $I_m = k\rho^m$, where ρ designates the heterogeneity index and k is the constant (Gaci and Zaourar 2014). As the above relation between I_m and m explains, the estimated ρ values bear an inverse relation with the heterogeneity of the subsurface. Accordingly, smaller (larger) ρ values designate higher (lesser) heterogeneity of the system. The two-step procedure to determine ρ is as follows:

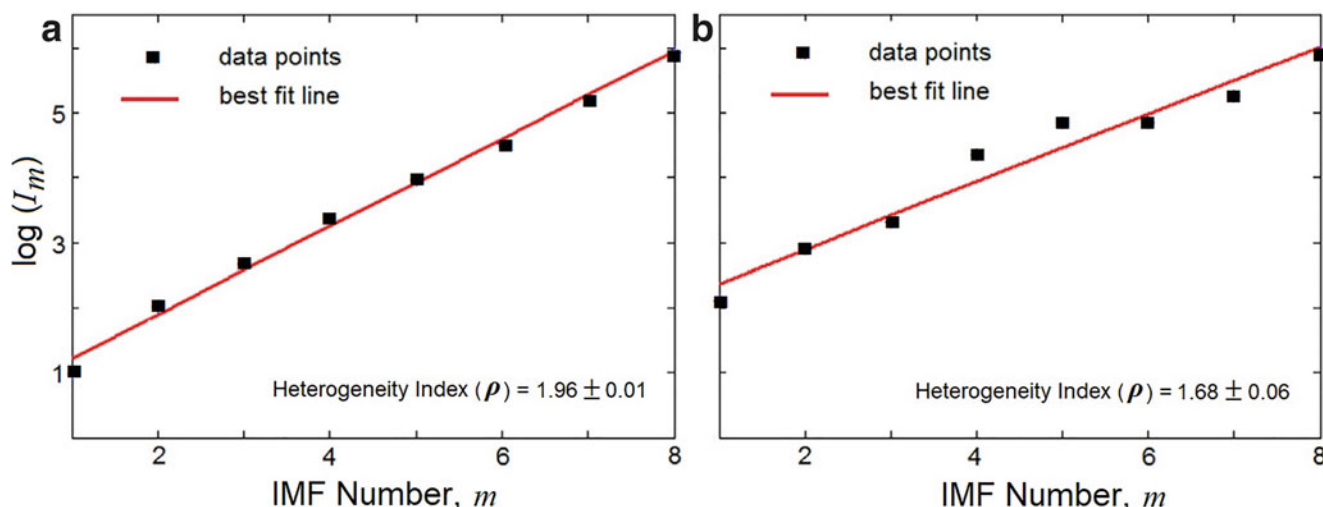
1. For each IMF, calculate the mean wavelength (I_m) defined as the ratio of the total number of points to the total number of peaks.
2. Draw a plot between the IMF number, m , and the logarithm of I_m . The antilogarithm of the slope of this line defines the heterogeneity index, ρ .

Figure 14 depicts an example of heterogeneity indices determined from the geophysical well-log data of two different wells having an equal number of IMFs (see Gairola and Chandrasekhar 2017). Since the value of the heterogeneity index of the data of the first well (Fig. 14a) is greater than that

of Fig. 14b, the subsurface of the former well is less heterogeneous than the latter. This has also been confirmed with the results of multifractal analysis of the same data sets (see Subhakar and Chandrasekhar 2016).

Summary

Fundamentals of signal processing techniques and their applications in various problems of geoscience research have been discussed. Starting from the basics of integral transforms to advanced levels of nonlinear signal analysis techniques, their theory and applications have been discussed. Particularly, the application of linear integral transforms, namely, the Fourier transform, Z-transform, and the Hilbert transform in geophysics, and the applications of nonlinear signal analysis techniques, such as wavelet analysis (CWT and DWT), fractal and multifractal analysis, empirical mode decomposition technique, and all of their roles in diverse fields of geosciences, such as solid earth geophysics, ionospheric studies, and geomagnetism, has been discussed. Exhaustive literature on all of the above techniques has also been provided for the benefit of the reader.



Signal Processing in Geosciences, Fig. 14 Linear regression between the logarithm of mean wavelength of each IMF and the respective IMF number for geophysical well-log data of two different wells. Since the heterogeneity index of the well-log data corresponding to the

first well (a) is greater than that of the second well (b), the subsurface of the former well is less heterogeneous than the latter (after Gairola and Chandrasekhar 2017)

Cross-References

- Fast Fourier Transform
- Fast Wavelet Transform
- Hilbert Space
- Multifractals
- Z-transform

References

- Alexandrescu M, Gilbert D, Hulot G, Le Mouél JL, Saracco G (1995) Detection of geomagnetic jerks using wavelet analysis. *J Geophys Res* 100:12557–12572
- Ansari RA, Buddhiraju KM (2015) k-means based hybrid wavelet and curvelet transform approach for denoising of remotely sensed images. *Remote Sens Lett* 6(12):982–991
- Ansari RA, Buddhiraju KM (2016) Textural classification based on wavelet, curvelet and contourlet features. In: 2016 IEEE international geoscience and remote sensing symposium (IGARSS). IEEE, pp 2753–2756
- Ansari RA, Malhotra R, Buddhiraju KM (2020a) Informal settlement identification using contourlet assisted deep learning. *Sensors* 20(9): 2733. <https://doi.org/10.3390/s200927332>
- Ansari RA, Buddhiraju KM, Malhotra R (2020b) Urban change detection analysis utilizing multiresolution texture features from polarimetric SAR images. *Remote Sens Appl Soc Environ* 20:100418. <https://doi.org/10.1016/j.rsase.2020.100418>
- Arneodo A, Audit B, Decoster N, Muzy J-F, Vaillant C (2002) Wavelet based multifractal formalism: applications to DNA sequences, satellite images of the cloud structure and stock market data. In: Bunde A, Kropp J, Schellnhuber HJ (eds) *The science of disasters: climate disruptions, heart attacks, and market crashes*. Springer, Berlin, pp 26–102
- Arneodo A, Audit B, Kestener P, Roux S (2008) Wavelet-based multifractal analysis. *Scholarpedia* 3(3):4103. revision #137211
- Audit B, Bacry E, Muzy J-F, Arneodo A (2002) Wavelet-based estimators of scaling behavior. *IEEE Trans Inf Theory* 48:2938–2954. <https://doi.org/10.1109/TIT.2002.802631>
- Bacry E, Muzy J-F, Arneodo A (1993) Singularity spectrum of fractal signals from wavelet analysis: exact results. *J Stat Phys* 70:635–674
- Battista BM, Knapp C, Goebel T, M.V. (2007) Application of the empirical mode decomposition and Hilbert-Huang transform to seismic reflection data. *Geophysics* 72(2):H29–H37. <https://doi.org/10.1190/1.2437700>
- Bhardwaj S, Chandrasekhar E, Seemala GK, Gadre VM (2020) Characterization of ionospheric total electron content using wavelet-based multifractal formalism. *Chaos, Solitons Fractals* 134:109653. <https://doi.org/10.1016/j.chaos.2020.109653>
- Bhimasankaram VLS, Mohan NL, Rao SVS (1977a) Analysis of gravity effect of two dimensional trapezoidal prism using Fourier transforms. *Geophys Prospect* 25:334–341
- Bhimasankaram VLS, Nagendra R, Rao SVS (1977b) Interpretation of gravity anomalies due to finite incline dikes using Fourier transforms. *Geophysics* 42:51–60
- Bhimasankaram VLS, Mohan NL, Rao SVS (1978) Interpretation of magnetic anomalies of dikes using Fourier transforms. *GeosExploration* 16:259–266
- Bodruzzaman M, Cadzow J, Shiavi R, Kilroy A, Dawant B, Wilkes M, (1991) Hurst's rescaled-range (R/S) analysis and fractal dimension of electromyographic (EMG) signal. *IEEE Proc. Southeastcon '91*, Williamsburg, VA, 2, 1121–1123
- Burg JP (1975) Maximum entropy spectral analysis, PhD thesis, Stanford University, Stanford (Un published)
- Burrus CS, Gopinath RA, Guo H (1998) Introduction to wavelets and wavelet transforms. A Primer, Upper Saddle River, Prentice Hall
- Cai JH (2013) Magnetotelluric response function estimation based on Hilbert-Huang transform. *Pure Appl Geophys* 170:1899–1911
- Cai JH, Tang JT, Hua XR, Gong YR (2009) An analysis method for magnetotelluric data based on the Hilbert-Huang transform. *Explr Geophys* 40:197–205
- Chandrasekhar E, Dimri VP (2013) Introduction to wavelets and fractals. In: Chandrasekhar E et al (eds) *Wavelets and fractals in earth system sciences*. CRC Press, Taylor and Francis, pp 1–28

- Chandrasekhar E, Rao VE (2012) Wavelet analysis of geophysical well-log data of Bombay offshore basin, India. *Math Geosci* 44(8): 901–928. <https://doi.org/10.1007/s11004-012-9423-4>
- Chandrasekhar E, Prasad P, Gurijala VG (2013) Geomagnetic jerks: a study using complex wavelets. In: Chandrasekhar E et al (eds) *Wavelets and fractals in earth system sciences*. CRC Press, Taylor and Francis, pp 195–217
- Chandrasekhar E, Subhakar D, Vishnu Vardhan Y (2015) Multifractal analysis of geomagnetic storms. *Proc. of the 17th annual conference of the Intl. Assoc. Math. Geosci.* (Editors: Helmut Schaben et al.), ISBN: 978-3-00-050337-5 (DVD), pp 1122–1127
- Chandrasekhar E, Prabhudesai SS, Seemala GK, Shenvi N (2016) Multifractal detrended fluctuation analysis of ionospheric Total electron content data during solar minimum and maximum. *J Atmos Sol Terr Phys* 149:31–39
- Dätig M, Schlurmann T (2004) Performance and limitations of the Hilbert-Huang transformation (HHT) with an application to irregular water waves. *Ocean Eng* 31:1783–1834
- Daubechies I (1988) Orthonormal bases of compactly supported wavelets. *Commun Pure Appl Math* XLI:909–996. John Wiley & Sons Inc.
- Daubechies I (1992) Ten lectures on wavelets, 2nd ed., SIAM, Philadelphia. CBMS-NSF regional conference series in applied mathematics 61
- Farge M (1992) Wavelet transforms and their applications to turbulence. *Annu Rev Fluid Mech* 24:395–457
- Farge M, Kevlahan N, Perrier V, Goirand E (1996) Wavelets and turbulence. *Proc IEEE* 84(4):639–669
- Fernandez-Martinez M, Guirao JLG, Sanchez-Granero MA, Segovia JET (2019) Fractal dimension for fractal structures (with applications to finance). *Springer Nature, Cham.*, 221 pages. <https://doi.org/10.1007/978-3-030-16645-8>
- Flandrin P, Rilling G, Gonçalves P (2004) Empirical mode decomposition as a filter Bank. *IEEE Sig Proc Lett* 11:112–114
- Gabor D (1946) Theory of communication. *J Instit Elect Eng* 93: 429–441
- Gaci S (2016) A new ensemble empirical mode decomposition (EEMD) denoising method for seismic signals. *Energy Procedia* 97:84–91
- Gaci S, Zaourar N (2014) On exploring heterogeneities from well logs using the empirical mode decomposition. *Energy Procedia* 59:44–50
- Gairola GS, Chandrasekhar E (2017) Heterogeneity analysis of geophysical well-log data using Hilbert Huang transform. *Physica A* 478: 131–142. <https://doi.org/10.1016/j.physa.2017.02.029>
- Gao Q, Duan C, Fan H, Meng Q (2008) Rotating machine fault diagnosis using empirical mode decomposition. *Mech Syst Signal Process* 22: 1072–1081
- Garcia X, Jones AG (2008) Robust processing of magnetotelluric data in the AMT dead band using the continuous wavelet transform. *Geophysics* 73:F223–F234
- Gilmore M, Yu CX, Rhodes TL, Peebles WA (2002) Investigation of rescaled range analysis, the Hurst exponent, and long-time correlations in plasma turbulence. *Phys Plasmas* 9:1312–1317
- Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE (2000) Physio bank, physio toolkit, and physio net: components of a new research resource for complex physiologic signals. *Circulation* 101(23):e215–e220. <http://circ.ahajournals.org/cgi/content/full/101/23/e215>
- Gonzalez RC, Woods RE (2007) Digital image processing, 3rd edn. Prentice Hall, New York
- Graps A (1995) An introduction to wavelets. *IEEE Comput Sci Eng* 2: 50–61. <https://doi.org/10.1109/99.388960>
- Gururajan MP, Mitra M, Amol SB, Chandrasekhar E (2013) Phase field modeling of the evolution of solid–solid and solid–liquid boundaries: Fourier and Wavelet implementations. In: Chandrasekhar E et al (eds) *Wavelets and fractals in earth system sciences*. CRC Press, Taylor and Francis, pp 247–271
- Hayes MH (1999) Schaum's outline of theory and problems of digital signal processing. McGraw-Hill Companies, Inc. ISBN 0-07-027389-8, 436 pages
- Haykins S (2006) Signals and systems, 4th edn. Wiley. 816 pages
- Hill EJ, Uvarova Y (2018) Identifying the nature of lithogeochemical boundaries in drill holes. *J Geochem Explr* 184:167–178
- Hongre L, Sailhac P, Alexandrescu M, Dubois J (1999) Nonlinear and multifractal approaches of the geomagnetic field. *Phys Earth Planet Inter* 110:157–190
- Huang NE, Wu Z (2008) A review on Hilbert-Huang transform: method and its application to geophysical studies. *Rev Geophys* 46:RG2006
- Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, Yen NC, Tung CC, Liu HH (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series analysis. *Proc R Soc Lond Ser A Math Phys Eng Sci* 454:903–993
- Jaffard S (1991) Detection and identification of singularities by the continuous wavelet transform, Preprint, Lab. Math. Modelisation, Ecole Nat. Ponts-et-Chaussees, La Courtille, Noisy-le-Grand, France
- Jansen FE, Kelkar M (1997) Application of wavelets to production data in describing inter-well relationships, in: Society of Petroleum Engineers # 38876: annual technical conference and exhibition, San Antonio, TX, October 5–8, 1997
- Jayaswal V, Gairola GS, Chandrasekhar E (2021) Identification of the typical frequency range associated with subsurface gas zones: a study using Hilbert-Huang transform and wavelet analysis. *Arab J Geo Sci* 14:335. <https://doi.org/10.1007/s12517-021-06606-5>
- Kantelhardt JW (2002) Fractal and multifractal time series. In: Meyers R (ed) *Encyclopedia of complexity and system science*. Springer, Berlin/Heidelberg, pp 3754–3779
- Kantelhardt JW, Zschiegner SA, Koscielny-Bunde E, Havlin S, Bunde A, Stanley HE (2002) Multifractal detrended fluctuation analysis of nonstationary time series. *Physica A* 316:87–114. [https://doi.org/10.1016/S0378-4371\(02\)01383-3](https://doi.org/10.1016/S0378-4371(02)01383-3)
- Kunagu P, Balasis G, Lesur V, Chandrasekhar E, Papadimitriou C (2013) Wavelet characterization of external magnetic sources as observed by CHAMP satellite: evidence for unmodelled signals in geomagnetic field models. *Geophys J Int* 192:946–950. <https://doi.org/10.1093/gji/ggs093>
- Lopes R, Betrouni N (2009) Fractal and multifractal analysis: a review. *Med Image Anal* 13:634–649. <https://doi.org/10.1016/j.media.2009.05.003>
- Ma K, Tang X (2001) Translation-invariant face feature estimation using discrete wavelet transform. In: Tang YY et al (eds) *WAA 2001*, LNCS 2251, Springer, Berlin, Heidelberg, pp 201–210
- Mallat S (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans Pattern Anal Mach Intell* 11: 674–693
- Mallat S (1999) A wavelet tour of signal processing. Academic, San Diego
- Mandelbort BB (1967) How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science* 156:636–638
- McDonald AJ, Baumgärtner AJG, Frasher GJ, George SE, Marsh S (2007) Empirical mode decomposition of atmospheric wave field. *Ann Geophys* 25:375–384
- Mohan NL, Sundararajan N, Rao SVS (1982) Interpretation of some two-dimensional magnetic bodies using Hilbert transforms. *Geophysics* 47:376–387
- Neukirch M, Garcia X (2014) Nonstationary magnetotelluric data processing with instantaneous parameter. *J Geophys Res Solid Earth* 119:1634–1654. <https://doi.org/10.1002/2013JB010494>
- Odegard ME, Berg JW (1965) Gravity interpretation using the Fourier integral. *Geophysics* 30:424–438
- Panda MN, Mosher CC, Chopra AK (2000) Application of wavelet transforms to reservoir – data analysis and scaling. *SPE J* 5:92–101
- Peng C-K, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL (1994) Mosaic organization of DNA nucleotides. *Phys Rev E* 49(2):1685–1689
- Prokoph A, Agterberg FP (2000) Wavelet analysis of well-logging data from oil source rock, egret member offshore eastern Canada. *Am Assoc Pet Geol Bull* 84:1617–1632

- Salat H, Murcio R, Arcaute E (2017) Multifractal methodology. *Physica A* 473:467–487
- Sengar RS, Cherukuri V, Agarwal A, Gadre VM (2013) Multiscale processing: a boon for self-similar data, data compression, singularities, and noise removal. In: Chandrasekhar E et al (eds) *Wavelets and fractals in earth system sciences*. CRC Press, Taylor and Francis, pp 117–154
- Sharma B, Geldart LP (1968) Analysis of gravity anomalies of two dimensional faults using Fourier transforms. *Geophys Prospect* 16: 76–93
- Sharma M, Vanmali AV, Gadre VM (2013) Construction of wavelets: principles and practices. In: Chandrasekhar E et al (eds) *Wavelets and fractals in earth system sciences*. CRC Press, Taylor and Francis, pp 29–92
- Subhakar D, Chandrasekhar E (2015) Detrended fluctuation analysis of geophysical well-log data. In: Dimri VP (ed) *Fractal solutions for understanding complex systems in Earth sciences*, 2nd edn. Springer International Publishing, Cham, pp 47–65. https://doi.org/10.1007/978-3-319-24675-8_4
- Subhakar D, Chandrasekhar E (2016) Reservoir characterization using multifractal detrended fluctuation analysis of geophysical well-log data. *Physica A* 445:57–65. <https://doi.org/10.1016/j.physa.2015.10.103>
- Sundararajan N (1983) Interpretation techniques in exploration geophysics using Hilbert transforms: Ph.D. thesis, Osmania University, Hyderabad, India
- Sundararajan N, Srinivas Y (2010) Fourier–Hilbert versus Hartley–Hilbert transforms with some geophysical applications. *J Appl Geophys* 71:157–161
- Telesca L, Lapenna V, Macchiato M (2004) Mono- and multi-fractal investigation of scaling properties in temporal patterns of seismic sequences. *Chaos, Solitons Fractals* 19:1–15. [https://doi.org/10.1016/S0960-0779\(03\)00188-7](https://doi.org/10.1016/S0960-0779(03)00188-7)
- Telesca L, Lovullo M, Hsu HL, Chen CC (2012) Analysis of site effects in magnetotelluric data by using the multifractal detrended fluctuation analysis. *J Asian Earth Sci* 54–55:72–77
- Torres ME, Colominas MA, Schlotthauer G, Flandrin P (2011) A complete ensemble empirical mode decomposition with adaptive noise. *IEEE Int Conf Acoustics Speech Signal Process (ICASSP)*, 2011, 4144–4147. <https://doi.org/10.1109/ICASSP.2011.5947265>
- Vega NR (2003) Reservoir characterization using wavelet transforms (Ph.D. dissertation), Texas A&M University
- Wu ZH, Huang NE (2009) Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Adv Adapt Data Anal* 1: 1–4
- Xue Y, Cao J, Tian R (2013) A comparative study on hydrocarbon detection using three EMD-based time frequency analysis methods. *J Appl Geophys* 89:108–115
- Yu ZG, Anh V, Estes R (2009) Multifractal analysis of geomagnetic storm and solar flare indices and their class dependence. *J Geophys Res* 114:A05214. <https://doi.org/10.1029/2008JA013854>
- Zhang Y, Paulson KV (1997) Enhancement of signal-to-noise ratio in natural-source transient magnetotelluric data with wavelet transform. *PAGEOPH* 149:405–419
- Zhang Y, Goldak D, Paulson KV (1997) Detection and processing of lightning-sourced magnetotelluric transients with the wavelet transform. *IEICE Trans Fund Elect Commun Comp Sci* E80: 849–858
- Zhang Q, Zhang F, Liu J, Wang X, Chen Q, Zhao L, Tian L, Wang Y (2018) A method for identifying the thin layer using the wavelet transform of density logging data. *J Pet Sci Eng* 160:433–441
- Zhao N, Li R (2015) EMD method applied to identification of logging sequence strata. *Acta Geophys* 63(5):1256–1275. <https://doi.org/10.1515/acgeo-2015-0052>

Simulated Annealing

Jaya Sreevalsan-Nair

Graphics-Visualization-Computing Lab, International Institute of Information Technology Bangalore, Electronic City, Bangalore, Karnataka, India

Definition

Physical annealing with solids is a thermodynamic process used in metallurgy, where heating and controlled cooling are used for modifying the physical properties of a material. The crystalline solid is first heated and then left to cool slowly until it reaches its most stable and regular crystal lattice configuration, i.e., with minimum lattice energy, and is free of crystal defects. Annealing gives superior structural integrity of solids if the cooling is done sufficiently slowly (Henderson et al. 2003). Thus, the physical state of a material, broadly represented as a physical system, is measured using its internal energy. The state of the material, i.e., the system, is stable when its internal energy is minimum. The state transition occurs during the cooling schedule in one of the two ways, namely, (i) when the material goes to a lower-energy configuration, by design, and (ii) when it goes to an upper-energy configuration under a specific condition. The condition for the latter is that the transition has an acceptance probability higher than a random number between 0 and 1.

Simulated annealing (SA) is akin to the annealing process where a computational system finds the global minimum as an optimum in the solution space through local searches. Thus, SA is defined as an optimization method that is a local search meta-heuristic (Henderson et al. 2003). Optimization methods are a class of computational methods that find a solution that either maximizes or minimizes an objective function, where the state of the system being optimized reaches the maximum or minimum value in its optimal state. Meta-heuristics are strategies of searching for a sufficiently good solution to an optimization problem. SA is also an iterative algorithm which uses local search strategy using hill-climbing method. Hill-climbing methods involve transitions of the states of the physical system that worsen the objective function. That said, SA is focused on finding an *approximate* globally optimal solution, as opposed to a *precise* locally optimal solution. Thus, the hill-climbing moves help in escaping the local extrema. In each iteration, the annealing process requires identifying two solutions in the solution space, where the probability of the system changing from one state to another is computed, and the solutions are compared to select the one that improves the value of the objective function is chosen, for the next iteration. Such transitions are referred to as *improving solutions*. In a few cases, a solution that

deteriorates the objective function value referred to as a *non-improving solution*, is chosen only to escape the local optima, in favor of the search for the global optima. This characteristic of SA has made it a popular optimization tool for both discrete optimization problems as well as continuous variable problems.

Overview

While SA has been found to be used prior to the formal usage of its current name, the documented introduction of the method, that also coined the term, was by Kirkpatrick et al. in 1983. This original method has since then been improved, and now, these resulting methods form a class of algorithms that mimics the physical annealing process (Kirkpatrick et al. 1983) and also includes the use of hill-climbing. The other class of approaches includes those that solve the Langevin equations, which are stochastic differential equations, and the global minimum is marked as the solution (Henderson et al. 2003). Langevin equation describes how a system goes through state transitions under the influence of deterministic but fluctuating random forces and is also used for describing *Brownian motion*.

Every instance of SA has the following elements (Bertsimas and Tsitsiklis 1993):

- A finite set S of states.
- A real valued objective/cost function J defined on S such that $S^* \subset S$, where the global minimum of J lie in S^* .
- A set of neighbors of i given as $\mathcal{N}(i)$ with symmetric neighborhood relationship, i.e., $j \in \mathcal{N}(i)$ implies that $i \in \mathcal{N}(j)$, where $j \in S - \{i\}$.
- The *cooling schedule*, which is a monotonously decreasing function $T : Z \rightarrow (0, \infty)$, where Z is a set of positive integers, i.e., $Z \subseteq \mathbb{Z}^+$, where $T(t)$ is the temperature at time t , and the initial temperature is $T(t_0) = T_0$, which is a positive value.
- An initial state $i_0 \in S$.

The iterative algorithm has a nested loop (Henderson et al. 2003) where the outer loop is governed by the temperature, and the inner loop is governed by the local neighborhood of a chosen state in a specific iteration. For every iteration of the outer loop, a specific number of iterations of the inner loop is implemented, and a decrement of temperature occurs. The outer loop is terminated when the stopping criterion is met, e.g., a certain temperature is reached. There is a “current state” associated with each iteration of the outer loop, whose neighborhood is searched for better solutions within the inner loop. The “current state” in the first iteration of the outer loop is i_0 . In the inner iteration, the neighbors of the chosen iteration are considered for a transition, if it is an improving solution except in specific cases of avoiding the local

minimum. The transitions lead to the change of the “current state” of the outer loop. The transitions are selected using hill-climbing, in practice. SA has been used for several applications, including traditional optimization problems, e.g., graph partitioning, graph coloring, and the traveling salesman problem (TSP) (Bertsimas and Tsitsiklis 1993), and industrial engineering/operations research problems, e.g., flow shop and job shop scheduling and lot sizing (Henderson et al. 2003).

The standard SA has a few disadvantages, including (a) the computationally intensive and inefficient approach in finding the optimal solution and (b) its inflexibility in certain problems (Ingber 1993). Thus, throughout the evolution of the SA, several directions have been pursued to improve the algorithm and its usability. These include generalizability of hill-climbing algorithm, convergence, efficient implementation, its extensions, etc. (Henderson et al. 2003). Genetic algorithms have been used in speeding up SA (Ingber 1993). One such method is simulated quenching, where the cooling schedule is a logarithmic function to allow faster cooling (Ingber 1993). The computational efficiency can be improved using parallel implementation using multiple processing units, such as CPU and GPU. Specific strategies are required to run an inherently sequential algorithm, such as SA, using parallel methods. One such implementation uses speculative computing to impose multi-point statistics for generating stochastic models subject to constraints (Peredo and Ortiz 2011).

A variant of SA is the adaptive simulated annealing (ASA), which relies on importance sampling of the parameter space, as opposed to deterministic methods. There are several strategies to implement ASA, of which one is about deriving parameters for current iteration from previous iterations.

SA can also be extended to multidimensional case where the objective function is defined on a subset of a k -dimensional Euclidean space (Fabian 1997). This implies the neighborhood search has to be extended to k -dimensional space, where both deterministic and random search strategies can be used.

Applications

SA has been applied for several applications in geoscience and related domains since its inception. One such application is to being a feasible alternative for multi-objective land allocation (MOLA) (Santé-Riveira et al. 2008). Rural land-use allocation is a complex problem given the multi-functionality of land units and noncontiguous fragmentation owing to urbanization. Thus, the conflicting multiple objectives for land-use allocation includes evaluating land units for both its utility type (e.g., crop type being cultivated), the contiguity of the units, and the compactness of resultant single-use landmasses. Additionally, SA is compared against

other similar optimization methods, such as integer programming, with respect to its run-time efficiency for large-scale problems, which is the case for regions with 2000–3000 land units. The SA has been found to be superior to alternative methods such as hierarchical optimization, ideal point analysis, and MOLA. The efficiency of SA has been found to be improved when provided with good a priori land use areas.

Optimization of reconfigurable satellite constellations (ReCon) taking into consideration the satellite design and orbits is a multidisciplinary problem (Paek et al. 2019). A ReCon is significant in agile earth observation, where it is used for both regular Earth observation as well as disaster monitoring. It is a constrained optimization problem which included four constraints, namely, (i) global observation mode (GOM) coverage, (ii) regional observation mode (ROM) revisit time, (iii) constellation mass, and (iv) reconfiguration time. The threefold objective in optimizing a ReCon include (i) minimization of revisit access time and maximization of area coverage, (ii) minimization of initial launch mass of ReCon, and (iii) minimization of reconfiguration time. The variables corresponding to the constraints are known to demonstrate nonlinear behavior, owing to which closed-form models are not available. Among the different SA methods used for this optimization problem, the point-based method has been found to outperform the gradient-based one. Given the nonlinearity behavior of the variables, genetic algorithms perform slightly better than SA in giving more optimal solutions, for this application.

SA can be used in conjunction with other data analytic methods. For an application of flash-flood hazard mapping, SA is used for eliminating redundant variables from the classifier models, which include boosted generalized linear model (GLMBoost), random forest classifier (RFC), and Bayes generalized linear model (BayesGLM) used in an ensemble (Hosseini et al. 2020). The classification done in this application is based on the severity of the flash flood and is used for forecasting. SA has been used in a novel and effective way for feature selection here.

Future Scope

In a similar way of integrating SA in a data analytic process, there is increased interest in the use of SA in deep learning, of late (Rere et al. 2015). While deep learning has been effective in classification and feature learning, it has its limitations in the ease of training models. SA can be used in improving the performance of convolutional neural networks (CNNs), in lieu of using optimal deep learning using meta-heuristics. This work has shown improvement in the performance of the CNNs with SA, specifically in the reduction of classification error, albeit at the cost of higher computation time.

Overall, it can be concluded that simulated annealing continues to be used as a powerful optimization tool in

applications in geosciences and related domains. It has the potential to combine with newer data analysis methods to give improved results.

Cross-References

- [Constrained optimization](#)
- [Deep learning in Geoscience](#)
- [Optimization in Geosciences](#)

References

- Bertsimas D, Tsitsiklis J (1993) Simulated annealing. *Stat Sci* 8(1):10–15
- Fabian V (1997) Simulated annealing simulated. *Comput Math Appl* 33(1–2):81–94
- Henderson D, Jacobson SH, Johnson AW (2003) The theory and practice of simulated annealing. In: *Handbook of metaheuristics*. Springer, New York, pp 287–319
- Hosseini FS, Choubin B, Mosavi A, Nabipour N, Shamshirband S, Darabi H, Haghighi AT (2020) Flash-flood hazard assessment using ensembles and bayesian-based machine learning models: application of the simulated annealing feature selection method. *Sci Total Environ* 711(135):161
- Ingber L (1993) Simulated annealing: practice versus theory. *Math Comput Model* 18(11):29–57
- Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220(4598):671–680
- Paek SW, Kim S, de Weck O (2019) Optimization of reconfigurable satellite constellations using simulated annealing and genetic algorithm. *Sensors* 19(4):765
- Peredo O, Ortiz JM (2011) Parallel implementation of simulated annealing to reproduce multiple-point statistics. *Comput Geosci* 37(8):1110–1121
- Rere LR, Fanany MI, Arymurthy AM (2015) Simulated annealing algorithm for deep learning. *Proc Comput Sci* 72:137–144
- Santé-Riveira I, Boullón-Magán M, Crecente-Maseda R, Miranda-Barros D (2008) Algorithm based on simulated annealing for land-use allocation. *Comput Geosci* 34(3):259–268

Simulation

Behnam Sadeghi^{1,2} and Julian M. Ortiz³

¹EarthByte Group, School of Geosciences, University of Sydney, Sydney, NSW, Australia

²Earth and Sustainability Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW, Australia

³The Robert M. Buchan Department of Mining, Queen's University, Kingston, ON, Canada

Definition

Simulation refers to the construction of computer-based models that represent natural phenomena. These models can

be stochastic in nature, where an underlying probabilistic framework defines the possible outcomes of the model, or they can refer to complex numerical modeling based on processes or physics principles. In this article, we refer to stochastic models built for geoscientific applications in a spatial context, where geostatistical methods are used to create multiple possible outcomes of the spatial distribution of different attributes. These outcomes are called realizations and are based on a random function model that is reproduced by a geostatistical method, or are defined algorithmically. In all cases, the goal is to properly characterize the spatial variability of one or more variables that represent the phenomenon.

Introduction

The application of geostatistical simulation has been significantly developed in geosciences to quantify and assess the variability at a local scale using conditional realizations, and model the uncertainty at unsampled locations (Journel 1974; Deutsch and Journel 1998; Chilès and Delfiner 2012; Mariethoz and Caers 2014; Sadeghi 2020) – see the ► “Uncertainty Quantification” chapter.

Geostatistical simulation methods can be used to characterize categorical and continuous variables. In the case of categorical variables, each location can take one and only one of K categories. For continuous variables, the variable can take a value in a continuous range, which depends on its statistical distribution. In both cases, the spatial continuity of the variable is characterized by variograms (see the ► “Variogram” chapter) or multiple-point statistics (see the ► “Multiple Point Statistics” chapter). Unlike interpolation methods where a single prediction is obtained at every location, and the predicted values do not reproduce the spatial variability of the phenomenon, geostatistical simulations honor the data at sample locations, reproduce the statistical distribution of the variable over the simulated domain, and impose the spatial continuity of the model over each realization, which allows the characterization of uncertainty over any transfer function that involves multiple locations in the domain simultaneously.

Methods

The simplest frameworks to create these stochastic models include the indicators approach, which can be used to characterize categorical and continuous variables, and the multi-Gaussian approach, which is the basis of several simulation methods for continuous variables, but can also be used for methods based on truncation rules to represent categorical variables. Algorithmically driven methods are based on

optimization (e.g., simulated annealing) and on reproduction of multiple-point statistics.

Indicator Approach

The indicator approach is based on coding the random variable as an indicator. This indicator variable can then be characterized by its indicator variogram or transition probabilities, and estimated with indicator kriging. The indicator kriging estimates the probability of prevalence of a given class in the case of categorical variables, or the probability of not exceeding a threshold, in continuous variables. These probabilities can be used to build the conditional distribution and draw simulated values in a sequential fashion (Alabert 1987; Journel 1989).

For a categorical variable S , where $S(u)$ can take one of K values $\{s_1, \dots, s_K\}$, K indicator variables are defined as:

$$I(u; k) = \begin{cases} 1, & \text{if } S(u) = s_k \\ 0, & \text{otherwise} \end{cases} \quad k = 1, \dots, K$$

The categorical random function S can be characterized by its statistical distribution, in particular its probability mass function, and by the direct and cross-indicator variograms (or transition probabilities):

$$\gamma_I(h; k, k') = \frac{1}{2} E\{[I(u; k) - I(u+h; k)] \cdot [I(u; k') - I(u+h; k')]\} \quad k, k' = 1, \dots, K$$

The probability of prevalence of any of the K categories at a particular location can be inferred by applying kriging (or cokriging) to the indicator variable for that particular category. For instance, the simple indicator kriging estimate at location u_0 is:

$$I(u_0; k)_{SK}^* = \left(1 - \sum_{j=1}^n \lambda_j\right) \cdot P_k + \sum_{j=1}^n \lambda_j \cdot I(u_j; k)$$

where P_k is the probability mass of category s_k , and the weights λ_j are determined by solving the simple indicator kriging system of equations:

$$\sum_{j=1}^n C_I(u_i, u_j; k) \cdot \lambda_j = C_I(u_i, u_0; k) \quad i = 1, \dots, n$$

where the indicator covariances C_I are inferred from the variogram γ_I that characterizes the indicator variable. The predicted conditional distribution can be built by combining the K estimates $I(u; k)_{SK}^*, k = 1, \dots, K$. This conditional

distribution can be used in a sequential framework for conditional simulation, as described later.

In the case of a continuous variable Z , where $Z(u)$ is characterized by its statistical distribution F_Z , a set of K indicator variables is defined as:

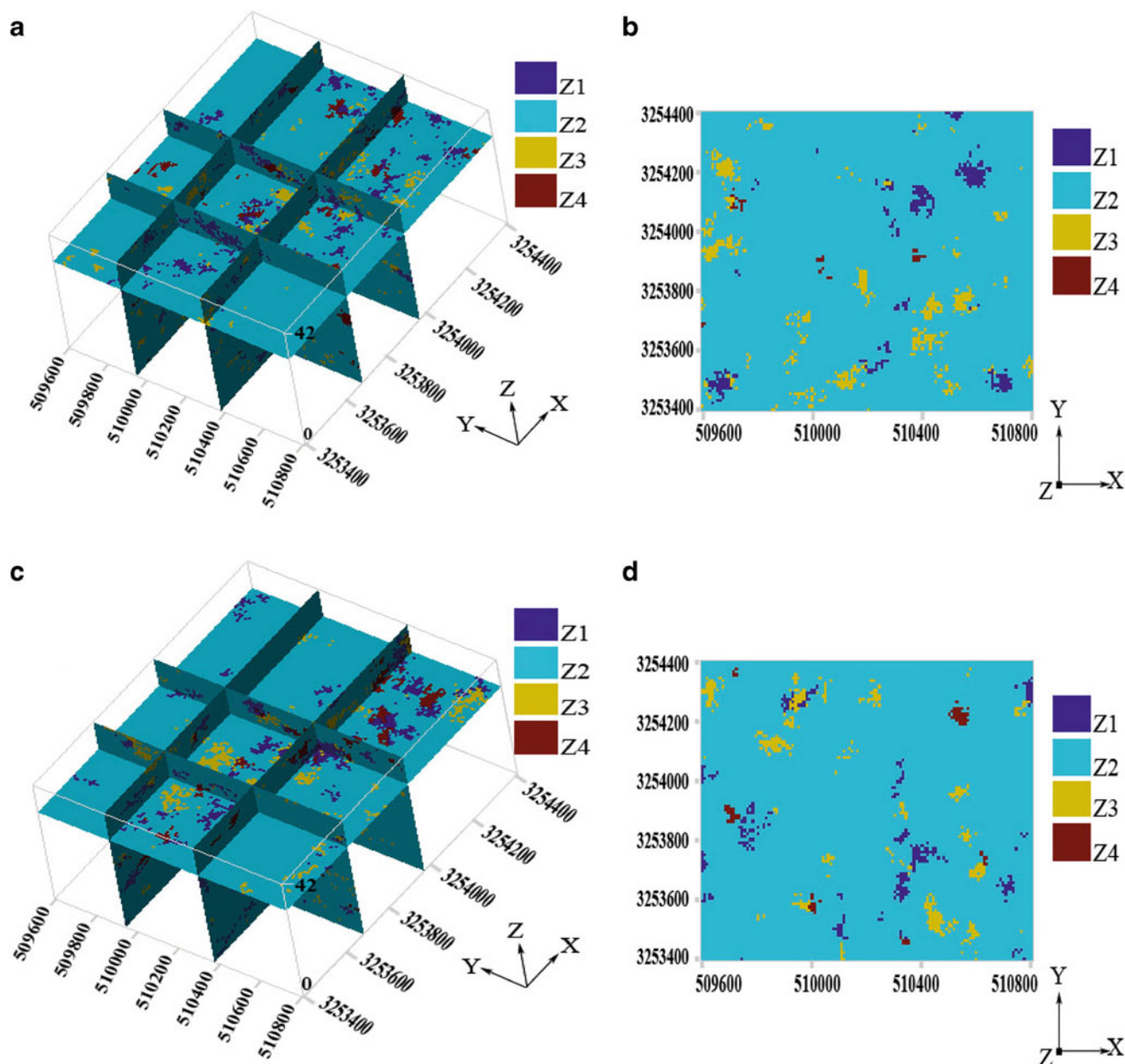
$$I(u; k) = \begin{cases} 1, & \text{if } Z(u) \leq z_k \\ 0, & \text{otherwise} \end{cases} \quad k = 1, \dots, K$$

where the thresholds z_k , $k = 1, \dots, K$ discretize the global distribution F_Z . The random function $Z(u)$ is characterized by

its spatial continuity through direct and cross-indicator variograms, and the conditional distribution at a particular location u_0 can be inferred by kriging (or cokriging) the indicators. The conditional cumulative distribution function is approximated by:

$$F_{Z \leq z_k}(u_0)^* = I(u_0; k)^*$$

As with the case of categorical variables (Fig. 1), these conditional distributions can be used in sequential simulation to create random fields that reproduce the spatial continuity of the variable.



Simulation, Fig. 1 Sequential indicator simulation (SISIM) representative realizations (#20, 40, 60, and 80) of 100 conditional realizations for four mineralized zones in Daralu copper deposit (SE Iran). (From Sojodehee et al. 2015)

Multi-Gaussian Approach

The multi-Gaussian approach relies on a distributional assumption to infer the properties of the random function (Goovaerts 1997). This greatly simplifies the inference and calculations. In most cases, this requires an initial transformation of the variable into a (standard) normal distribution that is assumed to behave as a multi-Gaussian variable:

$$Z = \varphi(Y), \text{ with } Y \sim N(0, 1)$$

The multi-Gaussian assumption for Y means that the joint

distribution of $Y(u), Y(u + h_1), \dots, Y(u + h_n)$ is a $(n + 1)$ -variate multi-Gaussian distribution for any n , that is:

$$f_{Y(u)\dots Y(u+h_n)}(y_0, \dots, y_n) = \frac{\exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right)}{\sqrt{(2\pi)^{n+1} |\boldsymbol{\Sigma}|}}$$

where

$$\mathbf{y} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_0 \\ \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{Y(u)}^2 & C_{Y(u)Y(u+h_1)} & \cdots & C_{Y(u)Y(u+h_n)} \\ C_{Y(u+h_1)Y(u)} & \sigma_{Y(u+h_1)}^2 & \cdots & C_{Y(u+h_1)Y(u+h_n)} \\ \vdots & \vdots & \ddots & \vdots \\ C_{Y(u+h_n)Y(u)} & C_{Y(u+h_n)Y(u+h_1)} & \cdots & \sigma_{Y(u+h_n)}^2 \end{pmatrix}$$

The vector $\boldsymbol{\mu}$ contains the means of the corresponding random variables, while $\boldsymbol{\Sigma}$ represents the covariances between pairs of random variables.

Under this assumption any conditional distribution is Gaussian in shape and its conditional mean and conditional variance can be inferred by simple kriging of the conditioning variables:

$$Y_{SK}^*(u) = \sum_{j=1}^n \lambda_j \cdot Y(u + h_j)$$

$$\sigma_{Y,SK}^2(u) = 1 - \sum_{j=1}^n \lambda_j \cdot C_{Y(u)Y(u+h_j)}.$$

As in the indicator framework, these conditional distributions can be used in sequential simulation to create random fields that reproduce the spatial continuity of the variable Y . These values can be back-transformed to the original units to create a random field of the original variable Z .

Sequential Simulation

The indicator or multi-Gaussian approaches facilitate the inference of the local distribution at a given location u , conditioned by a set of known values at locations $u + h_1, \dots, u + h_n$. A random field can be simulated by the recursive application of Bayes' law. Every newly simulated value

becomes a new conditioning point, leading to a set of spatially correlated points drawn from the joint multivariate (multi-point) distribution (Journel 1994).

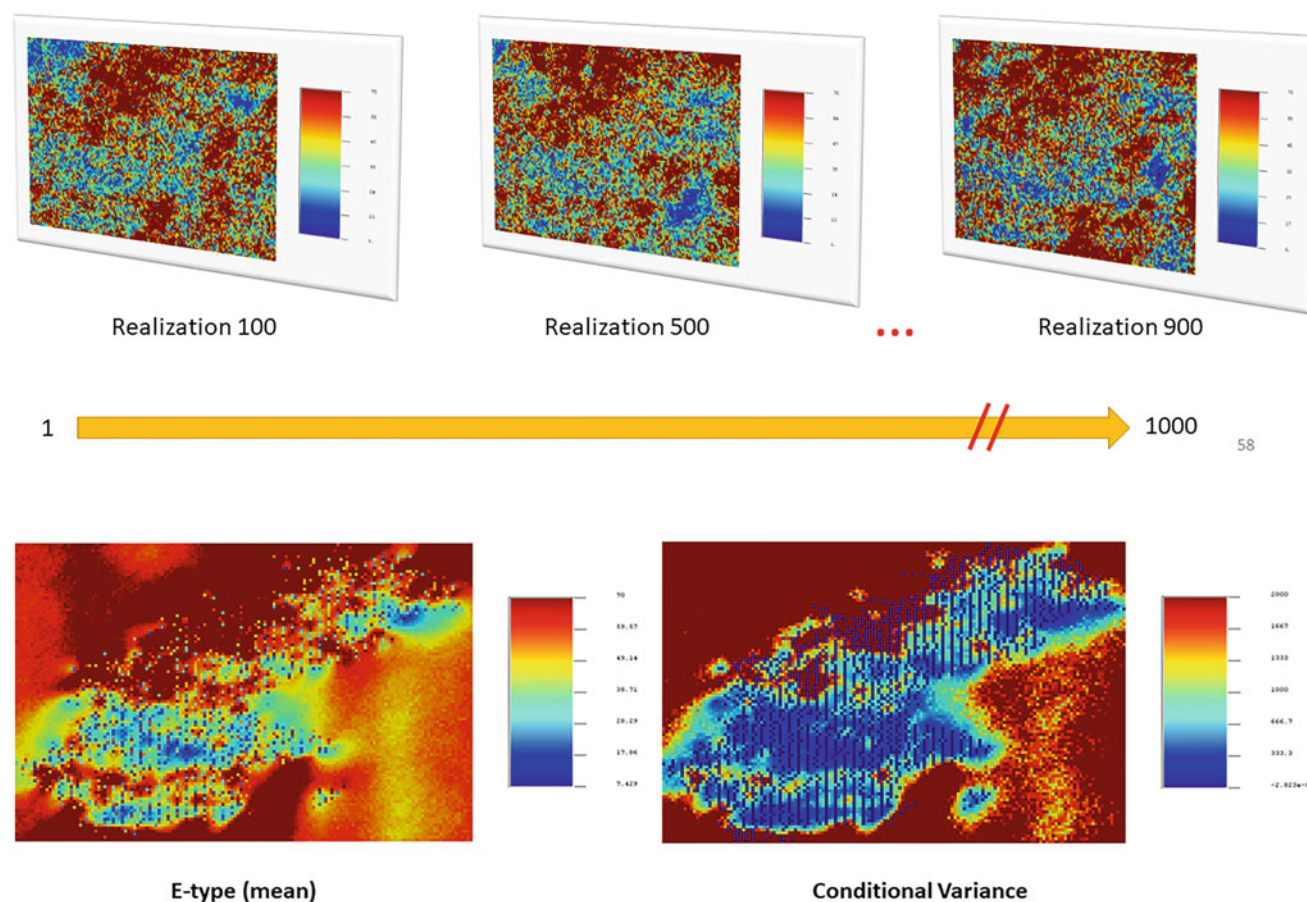
Simulation Methods

Although the sequential simulation framework provides a general method to simulate categorical or continuous variables, there exist other methods to achieve similar results.

Categorical variables can be simulated with sequential indicator simulation, or by truncation of one or more multi-Gaussian fields, as in truncated Gaussian simulation (Matheron et al. 1987), and pluri-Gaussian simulation (Armstrong et al. 2003). The key in the last two methods is to determine the spatial continuity of the multi-Gaussian fields in order to obtain the continuity of the indicator variables, once the truncation rule is applied.

Continuous variables can be simulated with sequential Gaussian simulation (SGSIM- Fig. 2) (Sadeghi 2020), but many other methods exist to create multi-Gaussian random fields, such as turning bands (TBSIM- Fig. 3), matrix decomposition (LU simulation), and FFT moving average.

Additionally, multiple-point simulation methods (Mariethoz and Caers 2014), some of which also rely on the sequential framework, can also be used to impose more complex relationships that are not properly reproduced with the sole use of variograms, both in the case of categorical and continuous variables.



Simulation, Fig. 2 Several representative realizations obtained by SGSIM, along with their Etype (mean at every location over the realizations), and conditional variance (variance at every location over the realizations)

Furthermore, most methods can be extended to the multi-variate case, where more than one attribute can be simulated simultaneously, preserving the direct and cross-spatial correlation, by generalizing the inference of the conditional distribution with simple cokriging.

In practice, there are many implementation details associated to the simulation of variables in real applications. Dealing with trends and nonstationary features and modeling multivariate relationships that depart from the multi-Gaussian assumption are the most relevant problems.

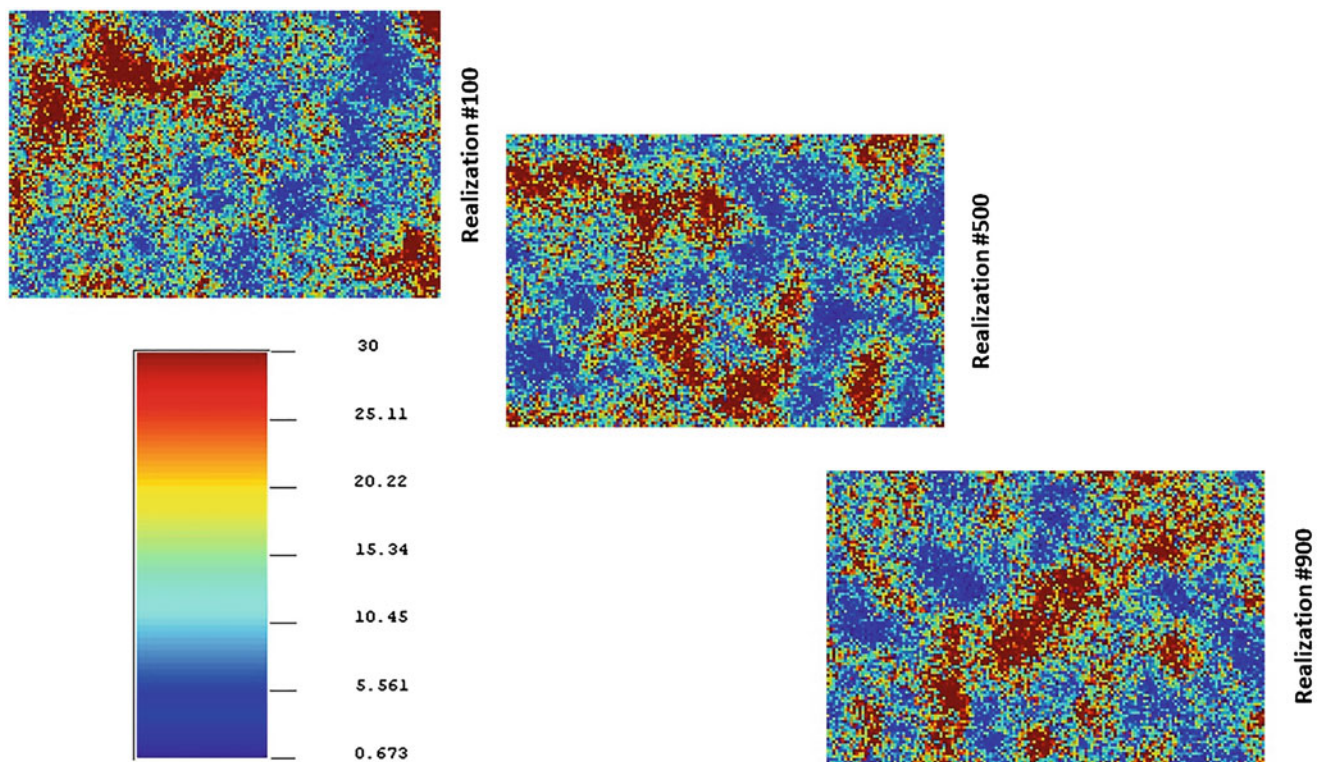
Use of Simulated Realizations

The different realizations generated by any of the above methods should:

1. Reproduce the conditioning sample data at their locations
2. Reproduce the global histogram (or proportions) of the variable in a stationary domain
3. Reproduce the spatial continuity of the variable (variogram or multiple-point statistics)

These realizations are not locally accurate, but aim at reproducing the in situ variability of the attribute in space. Therefore, these realizations can be used to characterize the uncertainty around the expected value at every location. This requires processing an ensemble (or a set) of realizations, in order to assess the expected variability in any response. This means that every realization is passed through a transfer function to generate a unique response. The set of responses obtained from processing an ensemble of realizations reflects the uncertainty about that response.

The ensemble of realizations can thus be interrogated to answer many different questions, such as the probability of one location to exceed a threshold z_C , or the probability of a volume (i.e., a collection of point locations) to exceed that same threshold z_C . More complex transfer functions are used in practical applications in environmental sciences, geological engineering, mining, petroleum, and other disciplines. The transfer function may be a complicated function of multiple locations, as in the case of the flow of petroleum in an oil reservoir, or be linked to a nonlinear function of the attribute, such as when calculating the metal recovery in a mine. It can



Simulation, Fig. 3 TBSIM several representative realizations (#100, #500 and #900)

also depend on the production sequence and time frame if a net present value is computed.

The variability captured by the simulations translates in uncertainty in prediction and design, which can be accounted for to obtain a robust and resilient design of the extractive approach in a mine or a petroleum reservoir.

Summary or Conclusions

Geostatistical simulation provides tools to characterize the variability of attributes at unsampled locations, and a quantification of the uncertainty related to any transfer function over the attributes. Simulation works by adding the lost variability that estimation methods smooth out, thus trading the local accuracy by the reproduction of the spatial variability. There are many different simulation methods to characterize categorical and continuous variables, which can also be extended to the simulation of multiple variables. The resulting realizations honor the data at sample locations, reproduce the global histogram, and reproduce the spatial continuity. These realizations must be treated as an ensemble to assess any response variable. There is no single best realization. All of the valid realizations must be processed through the transfer function to obtain multiple responses. The collection of responses characterizes the uncertainty after the transfer function. This is the

case in geosciences, environmental, mining, and petroleum applications, where the response may depend on multiple locations simultaneously and even change with time, as in the case of the net present value. Geostatistical simulation is a powerful tool to address the uncertainty in geoscience projects and obtain robust optima in decision-making that account for such uncertainty.

Cross-References

- [Multiple Point Statistics](#)
- [Uncertainty Quantification](#)
- [Variogram](#)

Bibliography

- Alabert FG (1987) Stochastic imaging of spatial distributions using hard and soft information. Master's thesis, Stanford University, Stanford
- Armstrong M, Galli A, Le Loc'h G, Geffroy F, Eschard R (2003) Plurigaussian simulations in geosciences. Springer, Berlin
- Chiles JP, Delfiner P (2012) Geostatistics modeling spatial uncertainty, 2nd edn. Wiley, New York
- Deutsch CV, Journel AG (1998) GSLIB: geostatistical software library and user's guide, 2nd edn. Oxford University Press, New York
- Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York

- Journal AG (1974) Geostatistics for conditional simulation of orebodies. *Econ Geol* 69:673–680
- Journal AG (1989) Fundamentals of geostatistics in five lessons. volume 8 Short course in geology. American Geophysical Union, Washington, DC
- Journal AG (1994) Modeling uncertainty: some conceptual thoughts. In: *Geostatistics for the next century*. Springer, Dordrecht
- Mariethoz G, Caers J (2014) Multiple-point geostatistics: stochastic modeling with training images. Wiley
- Matheron G, Beucher H, De Fouquet C, Galli A, Guerillot D, Ravenne C (1987) Conditional simulation of the geometry of fluvio-deltaic reservoirs. In: 62nd annual technical conference and exhibition of the society of petroleum engineers, Dallas. SPE 16753, 1987, pp 571–599
- Sadeghi B (2020) Quantification of uncertainty in geochemical anomalies in mineral exploration. PhD thesis, University of New South Wales
- Sojodehee M, Rasa I, Nezafati N, Vosoughi Abedini M, Madani N, Zeinedini E (2015) Probabilistic modeling of mineralized zones in Daralu copper deposit (SE Iran) using sequential indicator simulation. *Arab J Geosci* 8:8449–8459

Singular Spectrum Analysis

U. Radojčić¹, Klaus Nordhausen² and Sara Taskinen²

¹Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria

²Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland

Definition

Singular spectrum analysis (SSA) is a family of methods for time series analysis and forecasting, which seeks to decompose the original series into a sum of a small number of interpretable components such as trend, oscillatory components, and noise.

Introduction

Singular spectrum analysis (SSA) aims at decomposing the observed time series into the sum of a small number of independent and interpretable components such as a slowly varying trend, oscillatory components, and noise (Elsner and Tsonis 1996; Golyandina et al. 2001). SSA can be used, for example, for finding trends and seasonal components (both short and large period cycles) in time series, smoothing, and forecasting. When SSA is used as an exploratory tool, one does not need to know the underlying model of the time series.

The origins of SSA are usually associated with nonlinear dynamics studies (Broomhead and King 1986a,b), and over

time, the method has gained a lot of attention. As stated in (Golyandina et al. 2001), SSA has proven to be very successful in time series analysis and is now widely applied in the analysis of climatic, meteorological, and geophysical time series. Many variants of SSA exist. However, in the following, we focus on a specific SSA method, often referred to as *basic SSA*.

Singular Spectrum Analysis

We define $\mathbf{x} = \{x_t; t = 1, \dots, N\}$ for an observable time series. The SSA algorithm consists of the following four steps (Golyandina et al. 2018).

1. **Embedding:** First, the so-called $l \times k$ -dimensional *trajectory matrix*

$$\mathbf{T}(\mathbf{x}) = \begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_k \\ x_2 & x_3 & x_4 & \cdots & x_{k+1} \\ x_3 & x_4 & x_5 & \cdots & x_{k+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_l & x_{l+1} & x_{l+1} & \cdots & x_N \end{pmatrix} \quad (1)$$

is constructed. Here N is the time series length, l is so-called *window length* and $k = N - l + 1$. The trajectory matrix $\mathbf{T}$ is a linear map mapping $\mathbf{x} \in \mathbb{R}^N$ into a $l \times k$ -dimensional Hankel matrix, i.e., to $l \times k$ -dimensional matrix $\mathbf{X} = \mathbf{T}(\mathbf{x})$ with equal elements on the off-diagonals. Columns $\mathbf{x}_i = (x_i, \dots, x_{i+l-1})'$, $i = 1, \dots, k$, of $\mathbf{X}$ are called as *lagged vectors* of dimension l . Golyandina et al. (2001) suggests that l should be smaller than $N/2$ but sufficiently large so that l -lagged vector $\mathbf{x}_i$, $i = 1, \dots, k$, incorporates the essential part of the behavior in the initial time series $\mathbf{x}$.

2. **Decomposition:** Once the time series $\mathbf{x}$ is embedded into a trajectory matrix $\mathbf{X}$, it is decomposed into a sum of rank-1 matrices. Denote now $\mathbf{S} = \mathbf{X}\mathbf{X}'$ and write $\mathbf{S} = \mathbf{U}\mathbf{A}\mathbf{U}'$ for an eigendecomposition of $\mathbf{S}$. Here $\mathbf{A}$ is a $l \times l$ diagonal matrix with nonnegative eigenvalues, $\lambda_1 \geq \dots \geq \lambda_l$, as diagonal values and $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_l)'$, $\mathbf{u}_i \in \mathbb{R}^l$, is an $l \times l$ orthogonal matrix that includes the corresponding eigenvectors as columns. If we further write $\mathbf{v}_i = \mathbf{X}'\mathbf{u}_i / \sqrt{\lambda_i}$, $i = 1, \dots, d$, where $d = \text{rank}(\mathbf{X})$, then $\mathbf{X}$ can be decomposed into a sum of rank-1 matrices as follows:

$$\mathbf{X} = \sum_{i=1}^d \mathbf{X}_i = \sum_{i=1}^d \sqrt{\lambda_i} \mathbf{u}_i \mathbf{v}_i'. \quad (2)$$

Notice that $\mathbf{u}_i \in \mathbb{R}^l$ and $\mathbf{v}_i \in \mathbb{R}^k$ are left and right singular vectors of $\mathbf{X}$, respectively, and $\sqrt{\lambda_i} > 0$ are the

corresponding singular values. In SSA literature these ordered singular values are often referred to as the *singular spectrum*, thus giving the name to the method (Elsner and Tsonis, 1996). It is important to mention that the above method, where X is being decomposed into rank-1 components using singular value decomposition, is called basic SSA. In practice, also other decompositions can be used. For more details, see Golyandina et al., (2001), for example.

3. Grouping: In this step, the rank-1 components in decomposition (2) are grouped into predefined groups, where the group membership is described by a set of indices $I = \{i_1, \dots, i_p\} \subset \{1, \dots, d\}$, $p \leq d$. Then, the matrix corresponding to the group I is defined as $X_I = X_{i_1} + \dots + X_{i_p}$. If the set of indices $\{1, \dots, d\}$ is partitioned into m disjoint subsets $I_1, \dots, I_m$, $m \leq d$, then one obtains so-called *grouped decomposition* of matrix X :

$$X = X_{I_1} + \dots + X_{I_m}. \quad (3)$$

Especially, if $I_k = k$, for $k = 1, \dots, d$, the grouping is called elementary and the corresponding components in (3) are called elementary matrices. Furthermore, if only one group $I \subset \{1, \dots, d\}$ is specified, one proceeds by assuming that the given partition is $\{I, I^c\}$, where $I^c = \{1, \dots, d\} \setminus I$. In that case, X_I usually corresponds to the pattern of interest, while $X_{I^c} = X - X_I$ is treated as the residual.

4. Reconstruction: As final step, matrices X_{I_j} , $j = 1, \dots, m$, from decomposition (3) are transformed into new time series of length N . Reconstructed $\tilde{x}_{I_k}$ are obtained by sequentially averaging the elements of matrix X_{I_k} , $k = 1, \dots, m$, that lay on the off-diagonals, i.e.,

$$[\tilde{x}_{I_k}]_t = \frac{1}{|S_t|} \sum_{i+j=t+1} [X_{I_k}]_{i,j}, \quad (4)$$

where $|S_t|$ denotes the cardinality of the finite set $S_t = \{(i, j): i + j = t + 1\}$. In the literature, this process is known as *diagonal averaging*. If one applies the given reconstruction to all components in group decomposition (3) and, for simplicity of the notation, denotes $\tilde{x}_k := \tilde{x}_{I_k}$, $k = 1, \dots, m$, the resulting decomposition of initial time series x is given by:

$$x = \tilde{x}_1 + \dots + \tilde{x}_m. \quad (5)$$

In the case of basic SSA for univariate time series, the tuning parameters one needs to specify a priori are window length l and the partition of the set of indices. In more general SSA, the trajectory matrix, as well as its rank-1 decomposition, can be chosen more freely.

Separability

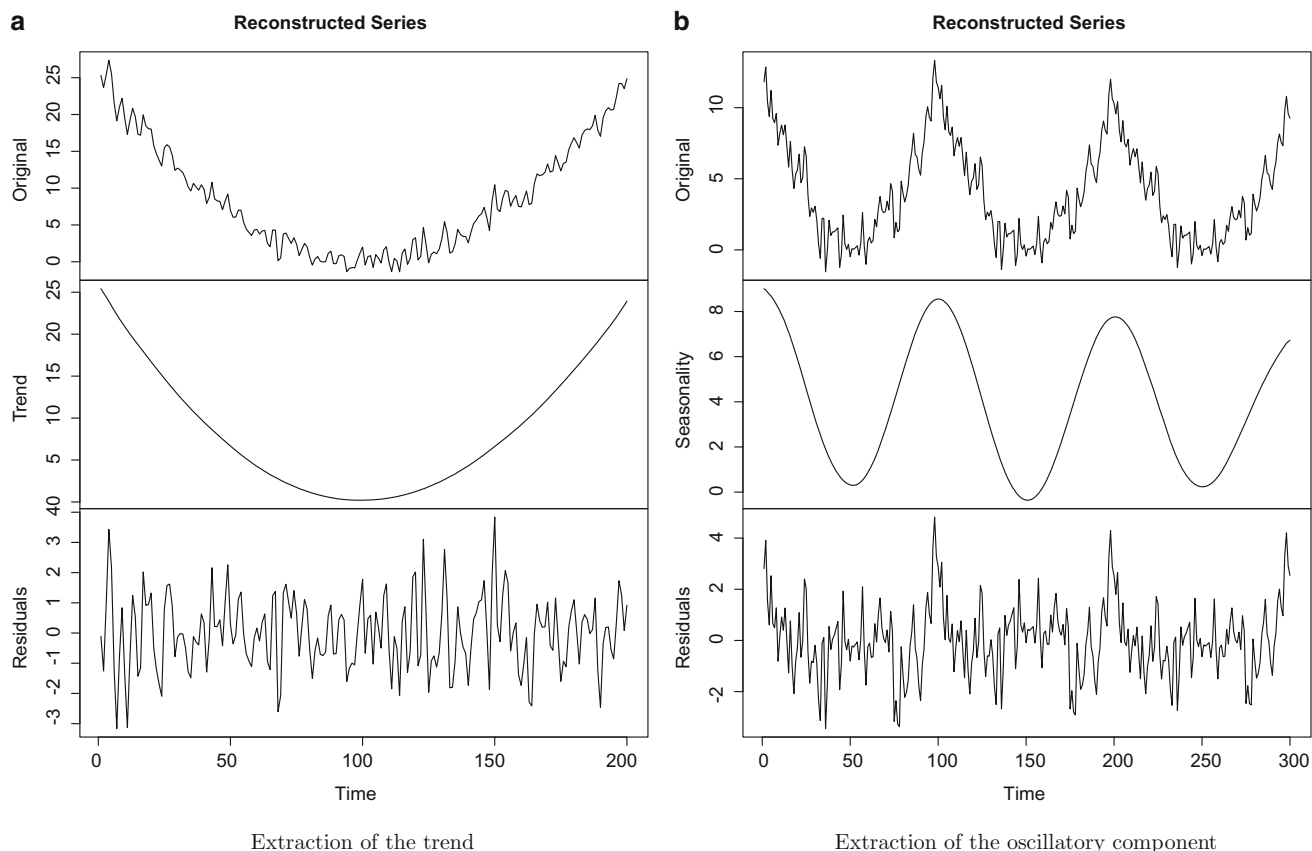
A central concept in SSA is separability, which to some extent ensures the validity of the method. Assume that one can decompose the time series into two univariate series, that is, $x = x_1 + x_2$. This representation is usually associated with a *signal plus noise* model, *trend plus the remainder* model, and other structured models. (Approximate) separability of the components x_1 and x_2 implies that there exists a grouping (see Step 4) such that reconstructed $\tilde{x}_1$ and $\tilde{x}_2$ are (approximately) equal to x_1 and x_2 , respectively, i.e., $\tilde{x}_i \approx x_i$, $i = 1, 2$. In basic SSA, the (approximate) separability corresponds to (approximate) orthogonality of the trajectory matrices.

Consider as an example a time series of length N . If N is large enough, the trend, which is a slowly varying smooth component, and periodic components are (approximately) separable and both are (approximately) separable from the noise. For illustration see Fig. 1, where a trend and a periodic component are separated from a noise component. In Fig. 1 (a), the trend is extracted using the first three significant elementary components, with window length size $l = 100$. In Fig. 1 (b), the seasonal component is extracted using the first six significant elementary components, with window length size $l = 150$. The extracted components are approximately equal to the theoretical trend and seasonal component.

In order to check the separability of the reconstructed components $\tilde{x}_1$ and $\tilde{x}_2$, with corresponding trajectory matrices $\tilde{X}_1$ and $\tilde{X}_2$, respectively, the normalized orthogonality measure given in (Golyandina et al., 2018) is:

$$\rho(\tilde{X}_1, \tilde{X}_2) = \frac{\langle \tilde{X}_1, \tilde{X}_2 \rangle_F}{\|\tilde{X}_1\|_F \|\tilde{X}_2\|_F},$$

where $\langle \cdot \rangle_F$ and $\|\cdot\|_F$ are Frobenius matrix inner product and norm, respectively. This orthogonality measure further induces dependence measure between two time series called *w-correlation* (Golyandina and Korobeynikov 2014). A large value of *w-correlation* between a pair of elementary components suggests that, in the grouping step, these should perhaps be in the same group. Furthermore, additive sub-series can be identified using the principle that the form of an eigenvector resembles the form of the sub-series produced by the eigenvector. For example, the eigenvectors produced by a slowly-varying component (trend) are slowly varying, and the eigenvectors produced by a sine wave (periodic component) are again sine waves with the same period (Golyandina and Korobeynikov 2014). Therefore, plots of eigenvectors are often used in the process of identification. For more details, see Golyandina and Korobeynikov (2014); Golyandina et al. (2018), for example.



Singular Spectrum Analysis, Fig. 1 Left: extraction of the polynomial trend in time series $x_t = (0.05(t-1)-5)^2 + y_t$, $t = 1, \dots, 200$, where y_t is $MA(0.9)$ process. Right: extraction of the oscillatory component in

time series $x_t = a_t y_t$, where $a_t = 1$, for $t = 1, \dots, 100$, $a_t = 0.9$ for $t = 101, \dots, 200$ and $a_t = 0.9^2$ for $t = 201, \dots, 300$. $y_t = (0.1(t-1)-5)^2 + z_t$, where z_t is $MA(0.9)$ process.

Prediction Using SSA

A class of time series especially suited for SSA are so-called *finite-rank* time series, where we say that the time series $\mathbf{x}$ is of finite-rank if its trajectory matrix is of rank $d < \min(k, l)$ and does not depend on window length l , for l large enough. In that case, under mild conditions, there exists one-to-one correspondence between the trajectory matrix of the time series $\mathbf{x}$ and a linear recurrent relation (LRR):

$$x_{t+d} = \sum_{i=1}^{d-1} a_i x_{t+i}, t = 1, \dots, N-d, \quad (6)$$

that governs the time series $\mathbf{x}$. Furthermore, the solution to LRR (6) can then be expressed as sums and products of polynomial, exponential, and sinusoidal components, whose identification leads to the reconstruction of trend and various periodic components, among others. If one applies obtained LRR (6) to the last terms of the initial time series $\mathbf{x}$, one obtains the continuation of $\mathbf{x}$ which serves as a prediction of the future.

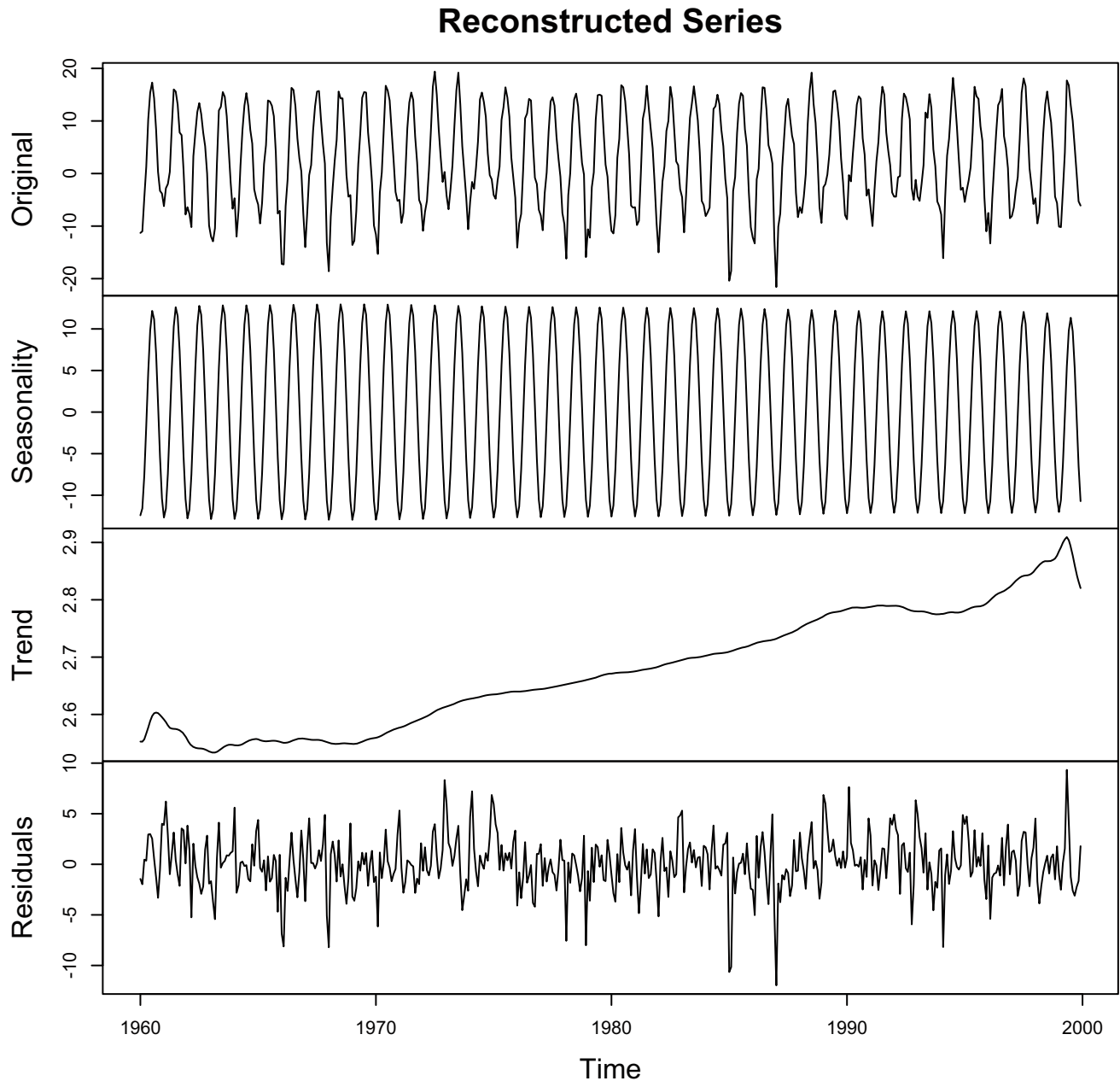
The real-data time series are not in general finite-rank time series. However, if time series $\mathbf{x}$ is a sum of a finite-rank signal $\mathbf{x}_1$ and additive noise, then SSA may approximately separate the signal component, and one can further use the methods designed for analysis and forecasting of the finite-rank series, thus obtaining the continuation (forecast) of a signal component $\mathbf{x}_1$ of $\mathbf{x}$ Golyandina and Korobeynikov (2014); Golyandina et al. (2001). Such a problem is known as forecasting the signal (trend, seasonality...) in the presence of additive noise. The confidence intervals for the forecast can be obtained using bootstrap techniques. For more details, see Golyandina et al. (2001); Golyandina and Korobeynikov (2014).

Example

To illustrate the SSA method, we use the average monthly temperatures (in °C) measured at the Jyväskylä airport, in Finland, from 1960 to 2000. To perform basic SSA as described above, we use the R (R Core Team 2020) package *Rssa* (Golyandina et al. 2015). As suggested in Golyandina

et al. (2001), for extracting a periodic component in short time series, it is preferable to take the window length l proportional to the period. Thus, we take $l = 48$. After investigating the singular values and paired scatter plots of the left singular vectors, we identify the first component as depicting a slowly varying trend, and the grouped fifth and sixth components to depict the yearly seasonality of the average monthly temperature. The original time series is shown together with its decomposition into seasonal component, trend component,

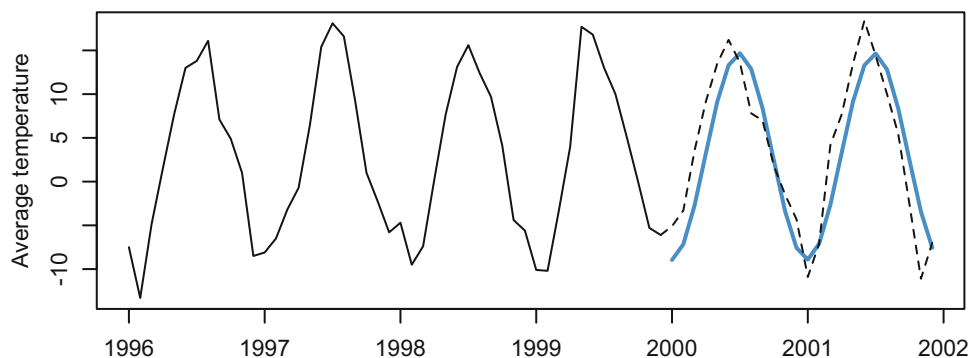
and residuals in Fig. 2. As seen in Fig. 2, the seasonal component is clearly dominating. However, there exists also a trend that looks almost linear. Eliminating the noise and using only the reconstructed trend and yearly seasonality, we predict the average monthly temperature for years 2000 and 2001 based on the components shown in Fig. 2. The predictions together with the true observed values are shown in Fig. 3. Figure 3 shows that SSA can be used to predict the temperatures in Jyväskylä quite nicely. For more details and



Singular Spectrum Analysis, Fig. 2 Top to bottom: Average monthly temperature ($^{\circ}\text{C}$) at the Jyväskylä airport from 1960 to 2000, the reconstructed seasonal component with 12-month period, reconstructed trend component, and residuals.

Singular Spectrum Analysis,

Fig. 3 The predicted average monthly temperature (°C) at the Jyväskylä airport for 2000 and 2001 (blue line). The dashed black line shows the true observed average temperatures.



guidelines for grouping and identification of trend and oscillatory components, we refer to Golyandina et al. (2001).

Extensions and Relations to Other Approaches

SSA has been extended to multivariate time series in which case it is known as multidimensional or multichannel SSA (M-SSA) and for the analysis of images where it is known as 2D singular spectrum analysis (2D-SSA).

While usually time series analysis is either performed in the time or the frequency domain, one can see SSA as a compromise, as a time-frequency method. For example, SSA can be seen as a method that chooses an adaptive basis generated by the time series itself while, for example, Fourier analysis uses a fixed basis of sine and cosine functions. For more detailed discussions and interpretations of SSA, we refer to Elsner and Tsonis (1996); Golyandina et al. (2001); Ghil et al. (2002).

Summary

Singular spectrum analysis (SSA) is a model-free family of methods for time series analysis, that can be used, for example, for eliminating noise in the data, identifying interpretable components such as trend and oscillatory components, forecasting, and imputing missing values. The method consists of the four main steps: embedding, decomposition, grouping, and reconstruction. The method is currently widely applied in the analysis of climatic, meteorological, and geophysical time series.

Cross-References

- [Singularity Analysis](#)
- [Time Series Analysis](#)
- [Time Series Analysis in the Geosciences](#)

References

- Broomhead D, King GP (1986a) Extracting qualitative dynamics from experimental data. *Physica D Nonlinear Phenomena* 20(2):217–236
- Broomhead D, King GP (1986b) On the qualitative analysis of experimental dynamical systems. *Nonlinear Phenomena and Chaos*:113–144
- Elsner JB, Tsonis AA (1996) *Singular spectrum analysis: a new tool in time series analysis*. Plenum Press, New York
- Ghil M, Allen MR, Dettinger MD, Ide K, Kondrashov D, Mann ME, Robertson AW, Saunders A, Tian Y, Varadi F, Yiou P (2002) Advanced spectral methods for climatic time series. *Rev Geophys* 40(1):3–1–3–41
- Golyandina N, Korobeynikov A (2014) Basic singular spectrum analysis and forecasting with R. *Comput Stat Data Anal* 71:934–954
- Golyandina N, Nekrutkin V, Zhigljavsky A (2001) *Analysis of time series structure: SSA and related techniques*. Chapman & Hall/CRC, Boca Raton
- Golyandina N, Korobeynikov A, Shlemov A, Usevich K (2015) Multivariate and 2D extensions of singular spectrum analysis with the Rssa package. *J Stat Softw* 67(2):1–78
- Golyandina N, Korobeynikov A, Zhigljavsky A (2018) *Singular spectrum analysis with R*. Springer, Berlin
- R Core Team (2020) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. <https://www.R-project.org/>

Singularity Analysis

Behnam Sadeghi^{1,2} and Frits Agterberg³

¹EarthByte Group, School of Geosciences, University of Sydney, Sydney, NSW, Australia

²Earth and Sustainability Research Centre, University of New South Wales, Sydney, NSW, Australia

³Central Canada Division, Geomathematics Section, Geological Survey of Canada, Ottawa, ON, Canada

Definition

Singular physical processes include various types of mineralization, volcanoes, floods, rainfall, and landslides (Wang et al. 2012). The singularity method is a model to “characterise

anomalous behavior of such processes mostly resulting in anomalous amounts of energy release or material accumulation within a narrow spatial–temporal interval” (Cheng 2007).

Introduction

One of the main objectives in the analysis of data in geochemical exploration programs is the selection and application of suitable classification models to distinguish between signals related to the mineralization effects and other processes or sources, as well as controls on variation in background geochemical distributions (Carranza 2009; Najafi et al. 2014; Zuo and Wang 2016; Grunsky and de Caritat 2020; Sadeghi 2020; see the entry ► “Exploration Geochemistry”). In environmental or urban geochemistry, the objective is generally the separation of geochemical patterns related to anthropogenic contamination from those derived from natural or geogenic processes (Guillén et al. 2011; Demetriades et al. 2015; Zuluaga et al. 2017; Albanese et al. 2018; Thiombane et al. 2019). The data addressed in these cases is commonly presented in the form of geochemical maps at various scales, sampling densities, and sampling geometries (Darnley 1990; Sadeghi 2020).

Various mathematical and statistical models have been applied to regional geochemical data to reveal subtle geochemical patterns related to the effects of different styles of mineralization or variations in parent lithologies. Conventional statistical methods such as exploratory data analysis (Tukey 1977; Chiprés et al. 2009), weights of evidence (Bonham-Carter et al. 1988, 1989), and probability plots (Sinclair 1991) used for anomaly detection or the isolation of geochemical signals and patterns related to the effects of mineralization on regolith geochemistry have a number of limitations. This has encouraged investigation of less conventional alternatives to conventional parametric methods, including fractal modeling and geostatistical simulation to isolate such signals or patterns (Cheng et al. 1994; Gonçalves et al. 2001; Li et al. 2003; Lima et al. 2003; Carranza 2009; Zuo 2011; Sadeghi et al. 2012, 2015).

Geochemical distributions, including those derived from regional regolith geochemical mapping studies, have been shown to display fractal behavior at various spatial scales (Bölviken et al. 1992). Such geochemical patterns develop through a variety of processes over geological time. It is the superimposition of several element enrichment or depletion episodes that results in the final element concentration distributions displaying multifractal behavior, containing a hierarchy of geochemical dispersion patterns observed at local to continental scales within most sampling materials (Zuo and Wang 2016, 2019). Differences observed in the fractal characteristics of regolith geochemical patterns relate to a number of factors. Whereas the main factor is typically a variation in

parent lithology (Cheng 2014; Nazarpour et al. 2015; Cámara et al. 2016), variation may also relate to the effects of mineralization and the task of separating element concentrations and/or samples into the categories traditionally termed “anomalous” versus “background” (Cohen and Bowell 2014).

The concept of fractal geometry was introduced to explain self-similarity or self-affinity at various scales in natural systems with models based on power-law spatial distributions in data (Mandelbrot 1983). Fractal or multifractal models are derived from power-law relationships between variables, such as the observed relationship between element concentration and the number of samples exceeding a given concentration.

A fractal model can be represented in the general form:

$$N(r) = C r^{-D} : r > 0, D \in \mathbb{R} > 0$$

where r is a characteristic linear measure, $N(r)$ is the number of objects with characteristic linear measure $\geq r$ ($\equiv N(\geq r)$), C is a constant of proportionality (a prefactor parameter), and D is the generalized fractal dimension (Shen and Zhao 1998). Some parametric distributions in geochemical data are intrinsically fractal (Shen and Cohen 2005).

Fractal relationships manifest as linear segments on (log-log) plots of the chosen fractal characteristics, in similar fashion to the identification of populations displaying particular distribution types and determining their parameters via probability or Q - Q plots (Sinclair 1991). The fractal dimensions are represented by the slopes of straight lines fitted to corresponding plots (Spalla et al. 2010; Sadeghi et al. 2012; Khalajmasoumi et al. 2017). In geochemical data, the slope for a fractal dimension can be interpreted as a measure of the element enrichment intensity (Afzal et al. 2010; Bai et al. 2010; Sadeghi et al. 2012; Zuo and Wang 2016). A change in slope or line orientation results in a new representative population for the given variables. In regional regolith geochemical data, this may be associated with a change in processes (e.g., different parent lithologies) or the existence of atypical conditions (such as the influence of mineralization). The threshold values, classifying different populations, are the breakpoints of the straight lines on such log-log plots (Carranza 2009; Afzal et al. 2011), providing a relatively simple graphical means of splitting data into populations.

A growing range of spatial and parametric fractal methods has been applied to geochemical data, to associate univariate or multivariate fractal dimensions to specific geochemical processes or domains. Whereas statistically self-similar fractals are typically isotropic, some may be anisotropic and display patterns at various scales depending on spatial orientations (Turcotte 1997), if a spatial fabric exists within the data.

In general, geochemical landscapes (Fortescue 1992) contain multiple fractal dimensions in regional geochemical data and thresholds and may be considered as spatially intertwined sets of monofractals (Feder 1988; Stanley and Meakin 1988). This may also be applicable to patterns with continuous spatial variability (Agterberg 1994, 2001). The multifractal behavior of variables within geochemical landscapes can be associated with the distributions of the probability density and spatial distributions of geochemical data (Cheng and Agterberg 1996; Carranza 2009; Zuo et al. 2009; Zuo and Wang 2016). In large, regional geochemical datasets, geochemical populations associated with mineralization are typically small and distinct as mineralization is typically of limited areal extent and is generally geochemically very distinct from other “background” or non-mineralization-related populations.

In conventional parametric methods, outlying values are generally removed prior to development of models (such as regression analysis or factor analysis) as they may significantly bias models. This partly derives from many conventional methods assuming geochemical element distributions following a normal or “normalisable” distribution (Reimann and Filzmoser 1999; Limpert et al. 2001; Sadeghi et al. 2015). Geochemical concentrations in the majority of studies do not follow a normal distribution or are the summation of a large number of component populations that either appear *en mass* to exhibit normality or log-normality (Li et al. 2003; Bai et al. 2010; He et al. 2013; Luz et al. 2014; Afzal et al. 2010). *Q-Q* plots are also based on the normality of data distribution neglecting the shape and intensity of anomalous areas. Fractal modeling does not typically require or assume normality and is largely unaffected by the presence of even extreme outliers. Therefore, fractal methods are gaining preference over conventional parametric methods (Afzal et al. 2010). Depending on the method used, fractal modeling can simultaneously incorporate population as well as spatial distribution characteristics of geochemical data (He et al. 2013; Luz et al. 2014).

Some fractal models have been developed to distinguish between different populations in geochemical datasets. This is to characterize geochemical anomalous values that could be related to the effects of mineralization. Such fractal models are number-size (N-S) (Mandelbrot 1983) and its 3D equivalent (Sadeghi et al. 2012), concentration-area (C-A) and perimeter-area (P-A) (Cheng et al. 1994, 1996; see the entry ► “Concentration-Area Plot”), spectrum-area (S-A) based on the relation between power spectrum and the occupied area of geochemical data values (Cheng et al. 1999; see the entry ► “Spectrum-Area Method”), singularity (Cheng 1999; Daya Sagar et al. 2018) based on the singularity spectrum, concentration-volume (C-V) (Afzal et al. 2011), spectrum-volume (S-V) (Afzal et al. 2012), concentration-concentration (C-C) (Sadeghi 2021), concentration-distance from centroids (C-DC) (Sadeghi and Cohen 2021a), category-based

fractal modeling (Sadeghi and Cohen 2021b) simulated size-number (SS-N) (Sadeghi et al. 2015), and global simulated size-number (GSS-N) (Madani and Sadeghi 2019).

Most such models are based on the relation between element concentrations and geometrical properties such as area and perimeter. The S-A model is used to differentiate geochemical anomalies from background by spectral analysis in the frequency domain through Fourier series analysis (Cheng 1999). The singularity model is applied to characterize the uniqueness degree of geological and geochemical properties based on its ability to delineate target areas, which are smoothed by conventional contouring models (Cheng 2007, 2008, 2012; Cheng and Agterberg 2009; Zou et al. 2013; Zuo and Wang 2016).

In exploration geochemical data, changes in fractal dimensions, indicated by the fractal classification models and ensuing clustering of samples, are used to cluster or partition samples and areas in geochemical maps. In large-scale regional geochemical mapping, the different classes represent the effects of variation in parent lithology, regolith development, and some environmental factors. Populations that are influenced by the effects of mineralization or contamination are typically represented by small numbers of samples that have conventionally been referred to as “geochemical anomalies” and the remainder as representing local or general “background” populations.

Methodology: Singularity Model

Singularity modeling is a form of fractal/multifractal modeling, which has been proposed and developed by Cheng (1999) to use in remote sensing and satellite image processing, with further work by Cheng (2007) focusing on its application in the study of stream sediment geochemical data for prediction of undiscovered mineral deposits. Some other studies on geochemical anomaly recognition have been undertaken using this method; e.g., Cheng and Zhao (2011) proposed that geochemical anomalies could be recognized for mineral deposit prediction through singularity theory, which has been developed to characterize nonlinear mineralization processes; Cheng (2012) applied the singularity theory for mapping geochemical anomalies related to buried sources and for predicting undiscovered mineral deposits in concealed areas; Wang et al. (2012) proposed a tectonic-geochemical exploration model based on the application of singularity theory to fault data; Xiao et al. (2012) combined the singularity mapping technique and spatially weighted principal component analysis (PCA) to delineate geochemical anomalies; Arias et al. (2012) demonstrated the robustness of C-A multifractal model and singularity technique for enhancing weak geochemical anomalies, in comparison with TS methods; and Zuo et al. (2013) compared the robustness of

C-A and S-A fractal models to singularity analysis in order to delineate geochemical anomalies in covered areas. Moreover, Zuo and Wang (2016) published a review paper to discuss and compare the theories, benefits, limitations, applications, and relations of C-A, S-A, concentration-distance (C-D), and C-V multifractal models in addition to the singularity model, especially for use in the recognition of weak anomalies associated with buried sources.

In general, the singularity modeling has been developed based on local singularity exponent. Such exponents are calculated by creating a number of maps (e.g., geochemical maps) with different scales. Such calculations are applicable to quantify the element concentration/depletion scaling properties. Based on Cheng (2012), singularity model is highly applicable to define weak but complex anomalous areas associated with mineral deposits, especially in areas covered or concealed by deserts, regolith, or vegetation.

The singularity phenomenon can be described in both 2D and 3D spaces based on a power-law relation between area (A) or volume (V) occupied by the target mineralization and the target metal/element total amount in that area ($\mu(A)$) or volume ($\mu(V)$). Such relationships have been formulated as follows (Cheng 2007):

$$\text{For } 3D : \mu(V_i) \propto V_i^{\frac{2}{3}}$$

where α is the singularity index. The metal/element concentration in this equation can be defined as $C(V_i) = \mu(V_i)/V_i$, and consequently:

$$C(V_i) \propto V_i^{\frac{2}{3}-1}$$

The same process for 2D models would result in:

$$\mu(A_i) \propto A_i^{\frac{1}{2}}$$

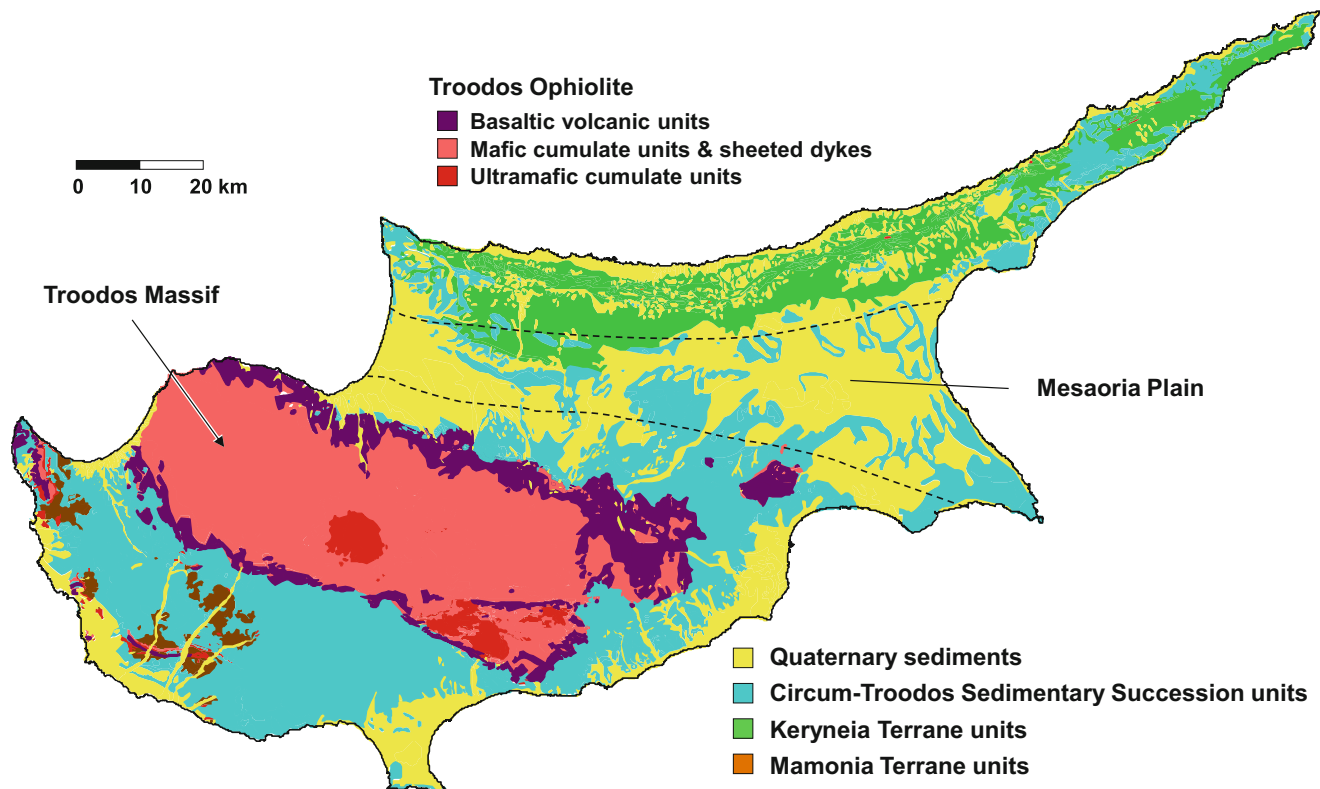
Then,

$$C(A_i) \propto A_i^{\frac{1}{2}-1}$$

The singularity index can be calculated using the following equation (Liu et al. 2019):

$$X = c \cdot e^{\alpha-E}$$

where X depicts element concentrations, c represents a constant value, ε is the normalized distance measure, α is the singularity index, and E is the Euclidian dimension (usually 2 in geochemical anomaly mapping) (Agterberg 2012; Zuo et al. 2013). In order to estimate the singularity index, we can



Singularity Analysis, Fig. 1 Simplified geology of Cyprus showing the Troodos Ophiolite (TO) flanked by the younger Circum-Troodos Sedimentary Succession (CTSS) and coastal alluvium-colluvium deposits (Sadeghi 2020; after Cohen et al. 2011)

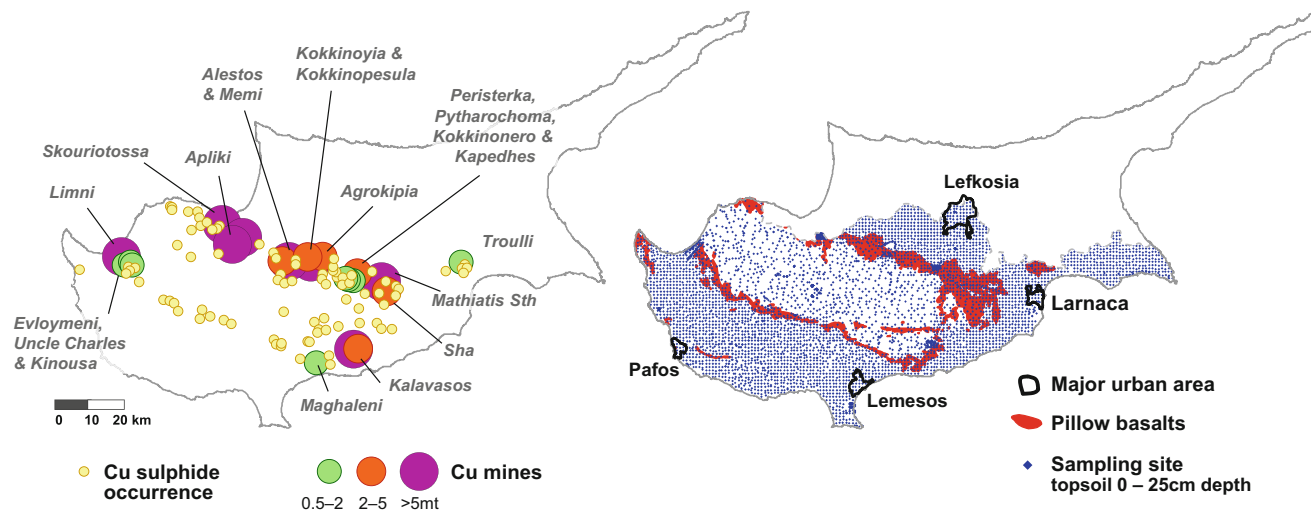
apply window-based methods. The process of estimating the singularity index has been discussed in detail in Cheng 2007, 2012 and Zuo et al. 2013.

The main point here that needs to be taken into account is that anomalous positive singularity index values (with $\alpha < 2$) and those with negative singularity index values (with $\alpha > 2$) are usually associated with enrichments and depletions of element concentrations, respectively. However, when $\alpha \approx 2$,

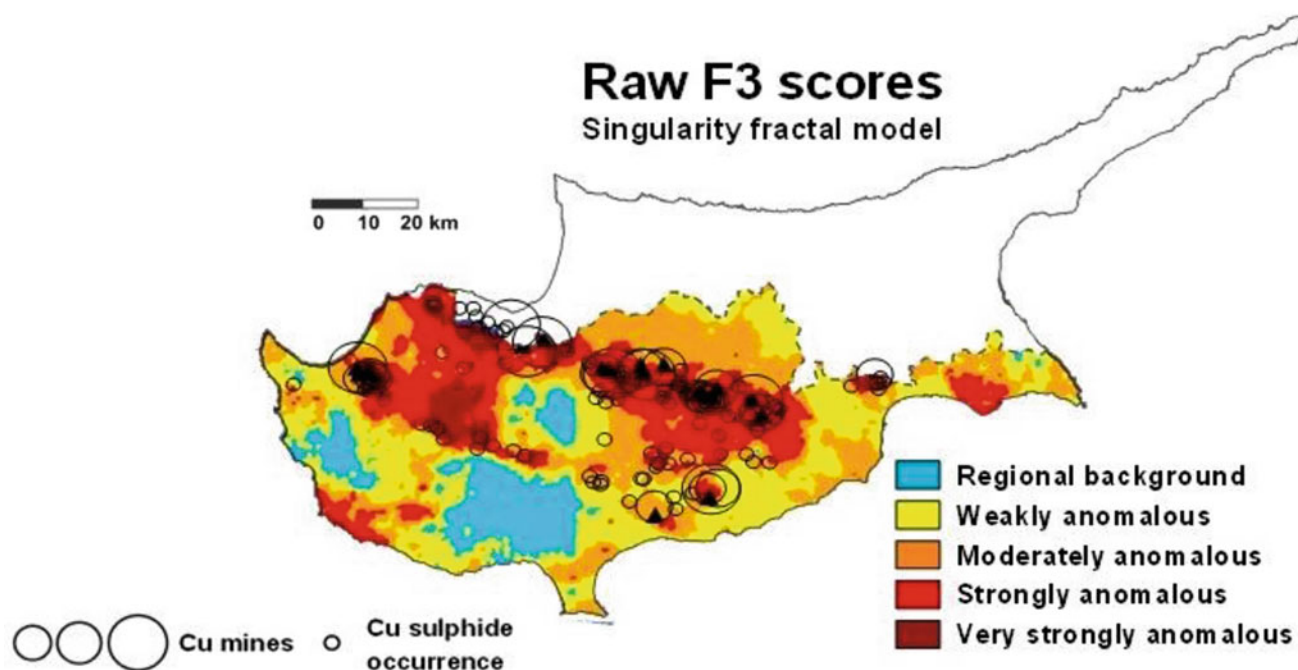
there is no positive or negative geochemical singularity (Xiao et al. 2012).

Case Study: Cyprus

Sadeghi (2020) applied the singularity model to the high-density soil geochemical atlas of Cyprus (GAC) with the



Singularity Analysis, Fig. 2 Location of Cyprus-style VMS Cu deposits and mines, the host basalt units and the 5,515 soil sampling locations (Sadeghi 2020)



Singularity Analysis, Fig. 3 Classified singularity model of the Cyprus-style VMS Cu mineralization in Cyprus (Sadeghi 2020, 2021)

objective to isolate patterns associated with the Cyprus-style Cu deposits (against a high but variable background Cu content in most of the lithologies hosting the mineral deposits) (Figs. 1 and 2).

Factor analysis was previously completed on the dataset by Cohen et al. (2011, 2012a,b). All significant factors are related to general lithologically controlled geochemical variations in the soil. At the scale of sampling undertaken, none of the factors could be related to either effects of mineralization or anthropogenic influences. So, the singularity model is applied to the third factor (Fig. 3). The model has strongly defined the highly anomalous areas including the relevant known mineral deposits and mines. Moreover, several other areas in eastern and SW Cyprus have been recognized that could be potential target areas for further exploration.

Summary and Conclusions

Singularity index, from a multifractal point of view, is applicable to quantify the amount of enrichment or depletion in mineralization. It is also applicable to generate characteristic maps of mineralization based on different classes from background to highly anomalous populations, especially in concealed, buried, and undiscovered areas. Phenomena such as mineralization have both spatial and temporal properties although, even in quite small areas or in short time, both properties can be defined by spatial and temporal singularity models.

Cross-References

- [Concentration-Area Plot](#)
- [Exploration Geochemistry](#)
- [Spectrum-Area Method](#)

Bibliography

- Afzal P, Khakzad A, Moarefvand P, Rashidnejad Omran N, Esfandiari B, Fadakar Alghalandis B (2010) Geochemical anomaly separation by multifractal modeling in Kahang (Gor Gor) porphyry system, Central Iran. *J Geochem Explor* 104:34–46
- Afzal P, Fadakar Alghalandis Y, Khakzad A, Moarefvand P, Rashidnejad Omran N (2011) Delineation of mineralization zones in porphyry Cu deposits by fractal concentration-volume modeling. *J Geochem Explor* 108:220–232
- Afzal P, Fadakar Alghalandis Y, Moarefvand P, Rashidnejad Omran N, Asadi Haroni H (2012) Application of power-spectrum-volume fractal method for detecting hypogene, supergene enrichment, leached and barren zones in Kahang Cu porphyry deposit, Central Iran. *J Geochem Explor* 112:131–138
- Agterberg FP (1994) Fractal, multifractals and change of support. In: Dimitrakopoulos R (ed) *Geostatistics for the next century*. Kluwer, Dordrecht, pp 223–234
- Agterberg FP (2001) Multifractal simulation of geochemical map patterns. In: Merriam DF, Davis JC (eds) *Geologic modeling and simulation in sedimentary systems*. Kluwer-Plenum, New York, pp 327–346
- Agterberg FP (2012) Multifractals and geostatistics. *J Geochem Explor* 122:113–122
- Albanese S, Cicchella D, Lima A, De Vivo B (2018) Geochemical mapping of urban areas. *Environ Geochem* 2nd ed, pp 133–151
- Arias M, Gumiel P, Martín-Izard A (2012) Multifractal analysis of geochemical anomalies: a tool for assessing prospectivity at the SE border of the Ossa Morena Zone, Variscan Massif (Spain). *J Geochem Explor* 122:101–112
- Bai J, Porwal A, Hart C, Ford A, Yu L (2010) Mapping geochemical singularity using multifractal analysis: application to anomaly definition on stream sediments data from Funin Sheet, Yunnan, China. *J Geochem Explor* 104:1–11
- Bölviken B, Stokke PR, Feder J, Jöessang T (1992) The fractal nature of geochemical landscapes. *J Geochem Explor* 43:91–109
- Bonham-Carter GF, Agterberg FP, Wright DF (1988) Integration of geological datasets for gold exploration in Nova Scotia. *Photogramm Eng Rem Sci* 54:1585–1592
- Bonham-Carter GF, Agterberg FP, Wright DF (1989) Weights of evidence modeling: a new approach to mapping mineral potential. In: Agterberg FP, Bonham-Carter GF (eds) *Statistical applications in the Earth sciences*. Geological survey of Canada, Ottawa, Ontario Paper, vol 89, pp 171–183
- Cámara J, Gómez-Miguel V, Ángel Martín M (2016) Identification of bedrock lithology using fractal dimensions of drainage networks extracted from medium resolution LiDAR Digital Terrain Models. *Pure Appl Geophys* 173:945–961
- Carranza EJM (2009) Geochemical anomaly and mineral prospectivity mapping in GIS. In: *Handbook of exploration and environmental geochemistry*, vol 11. Elsevier, Amsterdam. 368 p
- Cheng Q (1999) Multifractal interpolation. In: SJ Lippard, A Naess, R Sinding-Larsen (eds) *Proceedings 5th annual conferences international association mathematical geology*, Trondheim, Norway 245–250
- Cheng Q (2007) Mapping singularities with stream sediment geochemical data for prediction of undiscovered mineral deposits in Gejiu, Yunnan Province, China. *Ore Geol Rev* 32:314–324
- Cheng Q (2008) Non-linear theory and power-law models for information integration and mineral resources quantitative assessments. *Math Geol* 40:503–532
- Cheng Q (2012) Singularity theory and methods for mapping geochemical anomalies caused by buried sources and for predicting undiscovered mineral deposits in covered areas. *J Geochem Explor* 122:55–70
- Cheng Q (2014) Vertical distribution of elements in regolith over mineral deposits and implications for mapping geochemical weak anomalies in covered areas. *Geochem Explor Environ Anal* 14:277–289
- Cheng Q, Agterberg FP (1996) Multifractal modeling and spatial statistics. *Math Geol* 28:1–16
- Cheng Q, Agterberg FP (2009) Singularity analysis of ore-mineral and toxic trace elements in stream sediments. *Comput Geosci* 35: 234–244
- Cheng Q, Zhao P (2011) Singularity theories and methods for characterizing mineralization processes and mapping geo-anomalies for mineral deposit prediction. *Geosci Front* 2:67–79
- Cheng Q, Agterberg FP, Ballantyne SB (1994) The separation of geochemical anomalies from background by fractal methods. *J Geochem Explor* 51:109–130
- Cheng Q, Agterberg FP, Bonham-Carter GF (1996) A spatial analysis method for geochemical anomaly separation. *J Geochem Explor* 56: 183–195
- Cheng Q, Xu Y, Grunsky E (1999) Integrated spatial and spectral analysis for geochemical anomaly separation. In: Lippard SJ,

- Naess A, Sinding-Larsen R (eds) Proc. of the IAMG conference, vol 1. Trondheim, Norway, pp 87–92
- Chiprés J, Castro-Larragoitia J, Monroy M (2009) Exploratory and spatial data analysis (EDA-SDA) for determining regional background levels and anomalies of potentially toxic elements in soils from Catorce-Matehuala, Mexico. *Appl Geochem* 24:1579–1589
- Cohen DR, Howell RJ (2014) Chap 24: Exploration geochemistry. In: Scott SD (ed) *Treatise on geochemistry: Vol. 13, Geochemistry of mineral deposits*. Elsevier. ISBN 978 174223 306 2
- Cohen DR, Rutherford NF, Morisseau E, Zissimos AM (2011) The geochemical atlas of Cyprus. UNSW Press, Sydney
- Cohen DR, Rutherford NF, Morisseau E, Christoforou E, Zissimos AM (2012a) Anthropogenic versus lithological influences on soil geochemical patterns in Cyprus. *Geochem Explor Environ Anal* 12: 349–360
- Cohen DR, Rutherford NF, Morisseau E, Zissimos AM (2012b) Geochemical patterns in the soils of Cyprus. *Sci Total Environ* 420: 250–262
- Darnley AG (1990) International geochemical mapping: a new global project. *J Geochem Explor* 39:1–13
- Daya Sagar BS, Cheng Q, Agterberg F (2018) *Handbook of mathematical geosciences*. Springer, Cham
- Demetriades A, Birke M, Albanese S, Schoeters I, De Vivo B (2015) Continental, regional and local scale geochemical mapping. *J Geochem Explor* 154:1–5
- Feder J (1988) *Fractals*. Plenum, New York
- Fortescue JAC (1992) Landscape geochemistry: retrospect and prospect. *Appl Geochem* 7:1–53
- Gonçalves MA, Mateus A, Oliveira V (2001) Geochemical anomaly separation by multifractal modelling. *J Geochem Explor* 72(2): 91–114
- Grunsky EC, de Caritat P (2020) State-of-the-art analysis of geochemical data for mineral exploration. *Geochem Explor Environ Anal* 20(2): 217–232
- Guillén MT, Delgado J, Albanese S, Nieto JM, Lima A, De Vivo B (2011) Environmental geochemical mapping of Huelva municipality soils (SW Spain) as a tool to determine background and baseline values. *J Geochem Explor* 109:59–69
- He J, Yao S, Zhang Z, You G (2013) Complexity and productivity differentiation models of metallogenic indicator elements in rocks and supergene media around Daijiazhuang Pb-Zn deposit in Dangchang County, Gansu Province. *Nat Resour Res* 22:19–36
- Khalajmasoumi M, Sadeghi B, Carranza EJM, Sadeghi M (2017) Geochemical anomaly recognition of rare earth elements using multifractal modeling correlated with geological features, Central Iran. *J Geochem Explor* 181:318–332
- Li C, Ma T, Shi J (2003) Application of a fractal method relating concentrations and distances for separation of geochemical anomalies from background. *J Geochem Explor* 77:167–175
- Lima A, De Vivo B, Cicchella D, Cortini M, Albanese S (2003) Multifractal IDW interpolation and fractal filtering method in environmental studies: an application on regional stream sediments of Campania Region (Italy). *Appl Geochem* 18:1853–1865
- Limpert E, Stahel WA, Abbt M (2001) Log-normal distributions across the sciences: keys and clues. *Bioscience* 51:341–352
- Liu Y, Cheng Q, Carranza EJM, Zhou K (2019) Assessment of geochemical anomaly uncertainty through geostatistical simulation and singularity analysis. *Nat Resour Res* 28:199–212
- Luz F, Mateus A, Matos JX, Gonçalves MA (2014) Cu- and Zn-soil anomalies in the NE Border of the South Portuguese Zone (Iberian Variscides, Portugal) identified by multifractal and geostatistical analyses. *Nat Resour Res* 23:195–215
- Madani N, Sadeghi B (2019) Capturing hidden geochemical anomalies in scarce data by fractal analysis and stochastic modeling. *Nat Resour Res* 28:833–847
- Mandelbrot BB (1983) *The fractal geometry of nature*, 2nd edn. Freeman, New York
- Najafi A, Karimpour MH, Ghaderi M (2014) Application of fuzzy AHP method to IOCG prospectivity mapping: a case study in Taherabad prospecting area, eastern Iran. *Int J Appl Earth Obs* 33:142–154
- Nazarpour A, Omran NR, Rostami Paydar G, Sadeghi B, Matroud F, Mehrabi Nejad A (2015) Application of classical statistics, logratio transformation and multifractal approaches to delineate geochemical anomalies in the Zarshuran gold district, NW Iran. *Chem Erde-Geochem* 75:117–132
- Reimann C, Filzmoser P (1999) Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. *Environ Geol* 39: 1001–1014
- Sadeghi B (2020) Quantification of uncertainty in geochemical anomalies in mineral exploration. PhD thesis, University of New South Wales
- Sadeghi B (2021) Concentration-concentration fractal modelling: a novel insight for correlation between variables in response to changes in the underlying controlling geological-geochemical processes. *Ore Geol Rev.* <https://doi.org/10.1016/j.oregeorev.2020.103875>
- Sadeghi B, Cohen D (2021a) Concentration-distance from centroids (C-DC) multifractal modeling: A novel approach to characterizing geochemical patterns based on sample distance from mineralization. *Ore Geol Rev.* <https://doi.org/10.1016/j.oregeorev.2021.104302>
- Sadeghi B, Cohen D (2021b) Category-based fractal modelling: A novel model to integrate the geology into the data for more effective processing and interpretation. *J Geochem Explor.* <https://doi.org/10.1016/j.gexplo.2021.106783>
- Sadeghi B, Moarefvand P, Afzal P, Yasrebi AB, Daneshvar Saein L (2012) Application of fractal models to outline mineralized zones in the Zaghia iron ore deposit, Central Iran. *J Geochem Explor* 122: 9–19
- Sadeghi B, Madani N, Carranza EJM (2015) Combination of geostatistical simulation and fractal modeling for mineral resource classification. *J Geochem Explor* 149:59–73
- Shen W, Cohen DR (2005) Fractally invariant distributions and an application in geochemical exploration. *Math Geol* 37:895–909
- Shen W, Zhao PD (1998) Theory study of fractal statistical model and its application in geology. *Sci Geol Sin* 33:234–243
- Sinclair AJ (1991) A fundamental approach to threshold estimation in exploration geochemistry: probability plots revisited. *J Geochem Explor* 41:1–22
- Spalla ML, Gosso G, Moratta AM, Zucali M, Salvi F (2010) Analysis of natural tectonic systems coupled with numerical modelling of the polycyclic continental lithosphere of the Alps. *Int Geol Rev* 52: 1268–1302
- Stanley HE, Meakin P (1988) Multifractal phenomena in physics and chemistry. *Nature* 335(6189):405–409
- Thiombane M, Di Bonito M, Albanese S, Zuzolo D, Lima A, De Vivo B (2019) Geogenic versus anthropogenic behaviour and geochemical footprint of Al, Na, K and P in the Campania region (Southern Italy) soils through compositional data analysis and enrichment factor. *Geoderma* 335:12–26
- Tukey JW (1977) *Exploratory data analysis*. Addison-Wesley, Reading
- Turcotte DL (1997) *Fractals and Chaos in geology and geophysics*, 2nd edn. Cambridge University Press, Cambridge
- Wang G, Carranza EJM, Zuo R, Hao Y, Du Y, Pang Z, Sun Y, Qu J (2012) Mapping of district-scale potential targets using fractal models. *J Geochem Explor* 122:34–46
- Xiao F, Chen J, Zhang Z, Wang C, Wu G, Agterberg F (2012) Singularity mapping and spatially weighted principal component analysis to identify geochemical anomalies associated with Ag and Pb-Zn polymetallic mineralization in Northwest Zhejiang, China. *J Geochem Explor* 122:90–100
- Zuluaga MC, Norini G, Ayuso R, Nieto JM, Lima A, Albanese S, De Vivo B (2017) Geochemical mapping, environmental assessment and Pb isotopic signatures of geogenic and anthropogenic sources in three localities in SW Spain with different land use and geology. *J Geochem Explor* 181:172–190

- Zuo R (2011) Identifying geochemical anomalies associated with Cu and Pb-Zn skarn mineralization using principal component analysis and spectrum-area fractal modeling in the Gangdese Belt, Tibet (China). *J Geochem Explor* 111:13–22
- Zuo R, Wang J (2016) Fractal/multifractal modeling of geochemical data: a review. *J Geochem Explor* 164:33–41
- Zuo R, Wang J (2019) ArcFractal: an ArcGIS add-in for processing geoscience data using fractal/multifractal models. *Nat Resour Res* 29:3–12
- Zuo R, Cheng Q, Agterberg FP, Xia Q (2009) Application of singularity mapping technique to identification local anomalies using stream sediment geochemical data: a case study from Gangdese, Tibet, Western China. *J Geochem Explor* 101:225–235
- Zuo R, Xia Q, Zhang D (2013) A comparison study of the C–A and S–A models with singularity analysis to identify geochemical anomalies in covered areas. *Appl Geochem* 33:165–172

Smoothing Filter

Alejandro C. Frery

School of Mathematics and Statistics, Victoria University of Wellington, Wellington, New Zealand

Synonyms

Low-pass filter

Definition

Smoothing filters aim at improving the signal-to-noise ratio.

Consider real-valued images defined on a finite regular grid S of size $m \times n$. Denote this support $S = \{1, \dots, m\} \times \{1, \dots, n\} \subset \mathbb{N}^2$. Elements of S are denoted by $(i, j) \in S$, where $1 \leq i \leq m$ is the row, and $1 \leq j \leq n$ is the column, or by $s \in \{1, \dots, mn\}$. An image is then a function of the form $f: S \rightarrow \mathbb{R}$, i.e., $f \in S^{\mathbb{R}}$, and a *pixel* is a pair $(s, f(s))$.

A central notion in filters is that of a *neighborhood*. The neighborhood of any site $s \in S$ is any set of sites that does not include s , denoted $\partial_s \subset S \setminus \{s\}$, obeying the symmetry relationship $t \in \partial_s \Leftrightarrow s \in \partial_t$, where “ $\setminus$ ” denotes the difference between sets, i.e., $A \setminus B = A \cap B^c$, and B^c is the complement of the set B . Image filters on the data domain are typically defined with respect to a relatively small squared neighborhood of odd side ℓ of the form

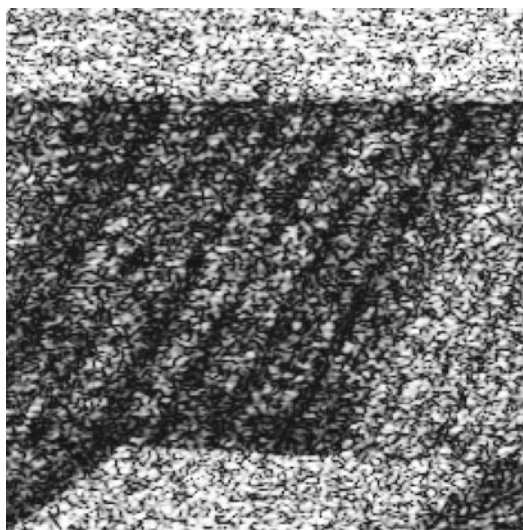
$$\partial_{(i,j)} = \left(\left(\left[i - \frac{\ell-1}{2}, i + \frac{\ell-1}{2} \right] \times \left[j - \frac{\ell-1}{2}, j + \frac{\ell-1}{2} \right] \right) \right) \cap S \setminus (i,j). \quad (1)$$

“small” in the sense that $\ell \ll m$ and $\ell \ll n$. We define the mask on the support

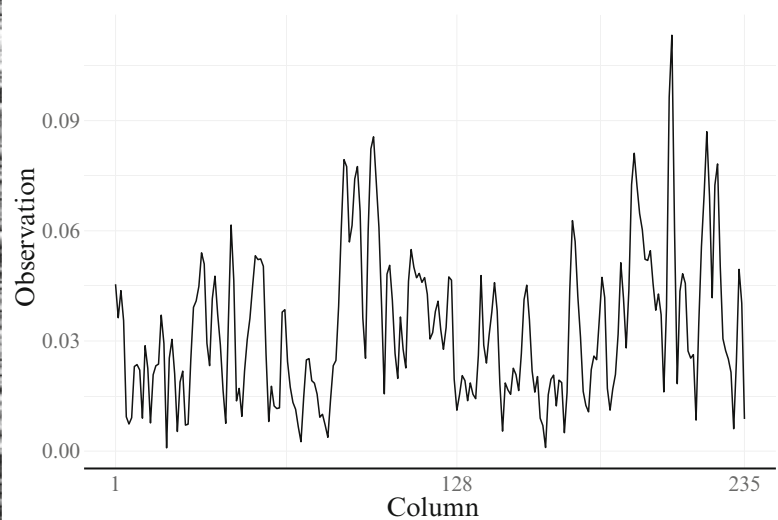
$$M = \left[-\frac{\ell-1}{2}, \frac{\ell-1}{2} \right] \times \left[-\frac{\ell-1}{2}, \frac{\ell-1}{2} \right],$$

with (odd) side ℓ , and it is denoted h_M .

Figure 1a shows the image we will use to illustrate the effect of smoothing filters. The image has 235×235 pixels, and it shows the amplitude data of a polarimetric synthetic aperture radar, horizontal-vertical channel. The effect of speckle is evident, but it is possible to see linear features.



(a) Original image



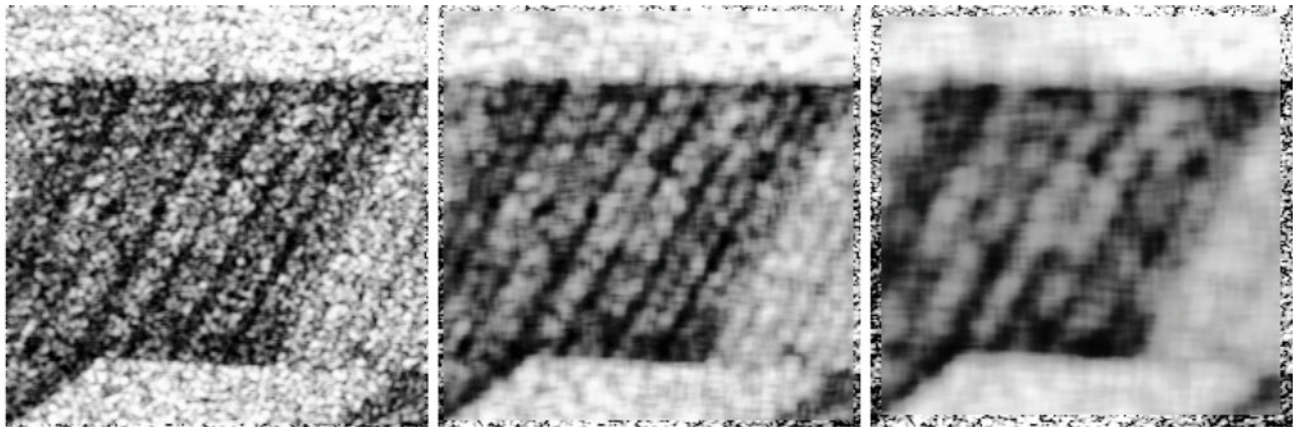
(b) Middle horizontal transect

Smoothing Filter, Fig. 1 Original image and transect

The “perfect” filter would eliminate the speckle, while retaining the details. Figure 1b shows the values of pixels in line 128, i.e., the transect at the middle of the image. Although the image shows dark and bright regions in this transect, the individual observations hinder such information due to the presence of speckle.

Linear Filters

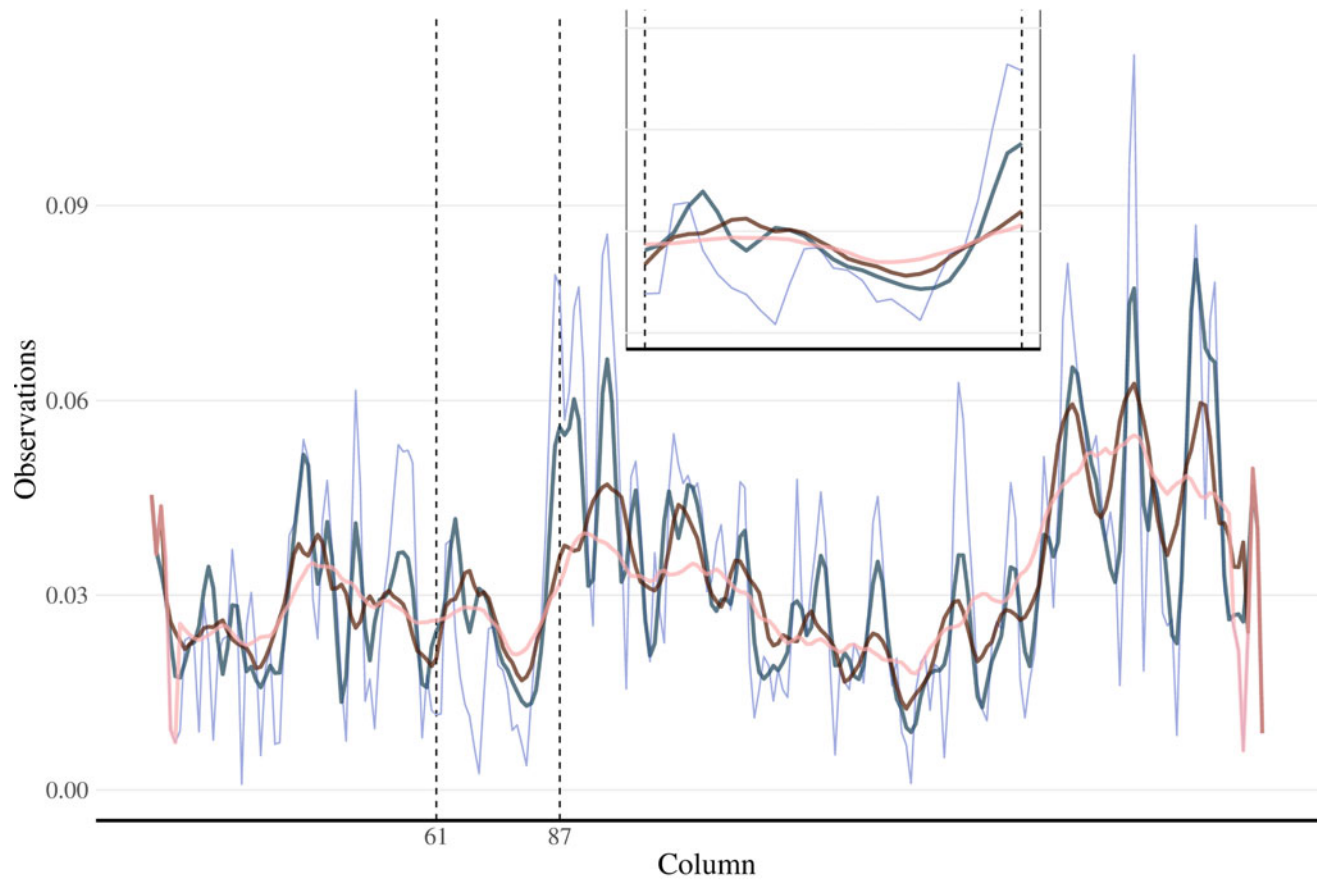
Consider $\Psi : S^{\mathbb{R}} \rightarrow S^{\mathbb{R}}$ an image transformation. It is said to be linear if for any pair of real numbers α, β and any pair of images $f_1, f_2 \in S^{\mathbb{R}}$, holds that $\Psi(\alpha f_1 + \beta f_2) = \alpha \Psi(f_1) + \beta \Psi(f_2)$. Linear filters belong to the class of linear transformations.



(a) Mean over 3×3

(b) Mean over 7×7

(c) Mean over 13×13



(d) Transects and zoom

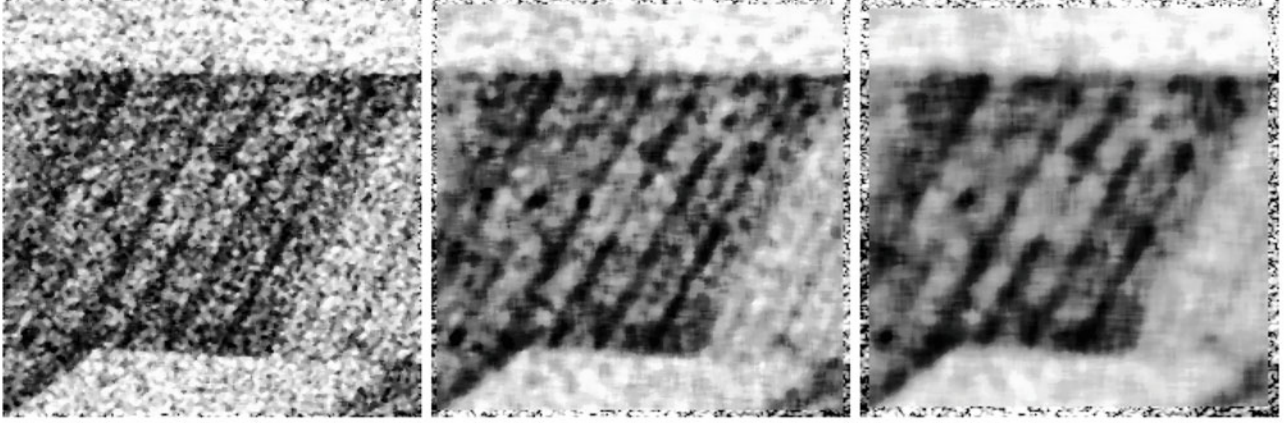
Smoothing Filter, Fig. 2 Mean filtered images and transects: original data in lilac, 3×3 in teal, 7×7 in brown, and 13×13 in pink

They are defined by means of a mask, and they are also called *convolutional filters*.

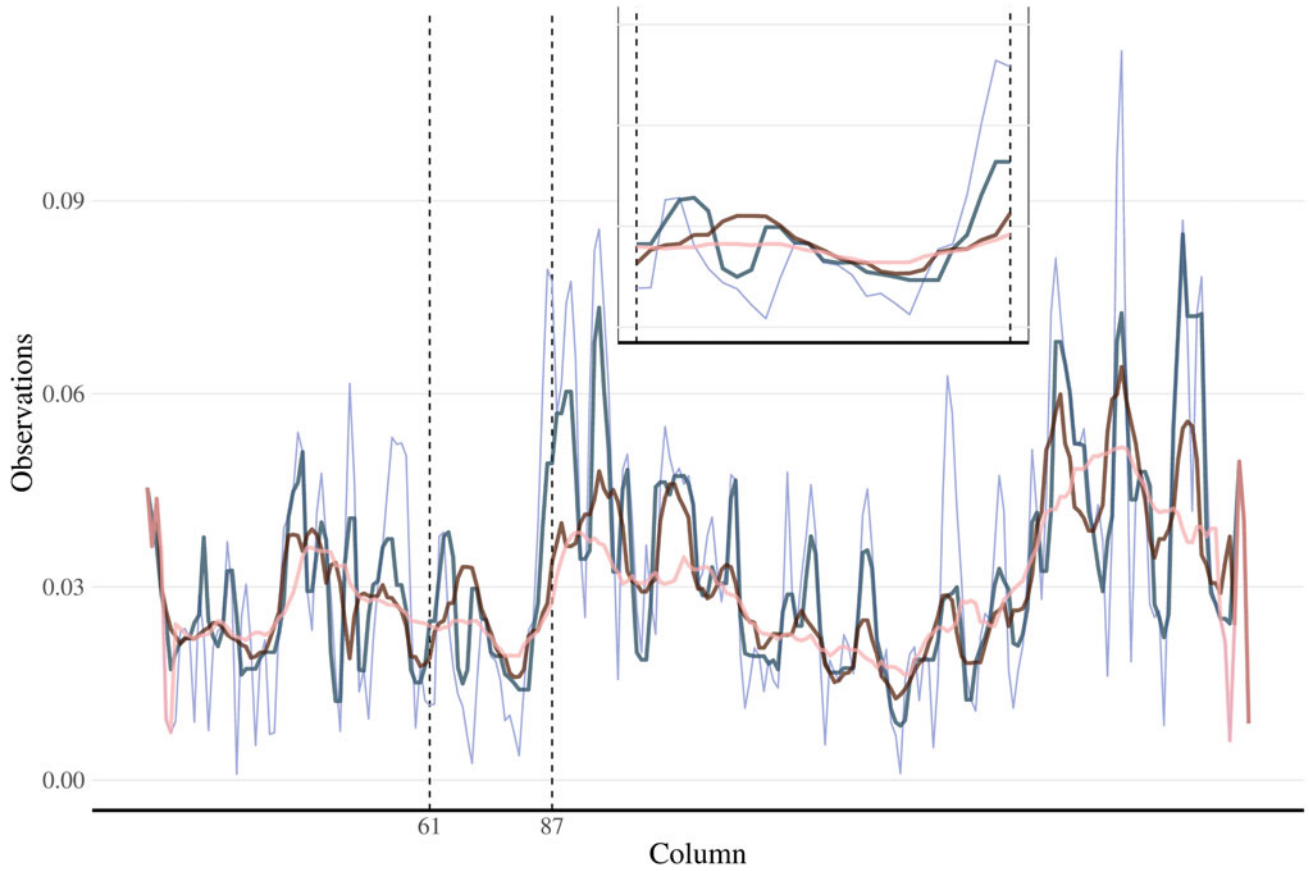
The convolution of the image $f \in S^{\mathbb{R}}$ by the mask $h_M \in M^{\mathbb{R}}$ is a new image $g \in S^{\mathbb{R}}$ given by

$$g(i, j) = \sum_{-\frac{\ell-1}{2} \leq i', j' \leq \frac{\ell-1}{2}} f(i-i', j-j') h_M(i', j'), \quad (2)$$

in every $(i, j) \in S$, and it is denoted $g = f * h_M$.



(a) Median over 3×3 windows (b) Median over 7×7 windows (c) Median over 13×13 windows



(d) Transects and zoom

Smoothing Filter, Fig. 3 Median filtered images and transects: original data in lilac, 3×3 in teal, 5×5 in brown, and 7×7 in pink

If all the elements of the mask are nonnegative, then it is customary to say that the resulting filter is a “low pass filter.” This is related to the fact that, in the frequency domain representation, such filters tend to reduce the high-frequency content, which is mainly related to noise and to small details. In the following we will illustrate some of these filters, and their effect on the data presented in Fig. 1a.

The class of *mean filters* is defined by masks h_M where all or some values take the same (positive) value, being the rest zero. In order to preserve the mean value of the image f , the usual choice for this value is the reciprocal of the sum of entries of h_M , as presented in the following:

$$h_M^5 = \begin{pmatrix} 0 & 1/5 & 0 \\ 1/5 & 1/5 & 1/5 \\ 0 & 1/5 & 0 \end{pmatrix}, h_M^9 = \begin{pmatrix} 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \end{pmatrix},$$

and $h_M^{169} = (h(i, j) = 1/169)$ in every $-6 \leq i, j \leq 6$.

Figure 2a, b, and c are the result of applying convolutions with masks h_M^9 , h_M^{49} , and h_M^{169} to the image shown in Fig. 1a. These filtered images show the typical effect of low-pass filters: The noise is reduced, but small details are also affected. The bigger the support of the mean filter is, the more intense both effects are.

The transect shown in Fig. 2d shows the effect of the mean filters. The detail presented in the inset plot illustrates how the noise reduction goes along with the blurring of sharp edges.

Statistical Filters

This kind of filters employs statistical ideas. Many of them also belong to the class of linear filters, but many others do not.

The median filter returns in each coordinate s the median value of the observations belonging to $\overline{\partial}_s$. Its effect is similar to that of the mean filter, but it tends to preserve better edges at the expense of not reducing the additive noise as effectively as the latter does. On the other hand, it is excellent for reducing impulsive noise. Figure 3a, b, and c show the effect of the median filter and how it varies according to the window size. Figure 3d shows the original and median-filtered values along the transect.

More Details

A more model-based approach can be found in the book by Velho et al. (2008). Other important references are the works by Barrett and Myers (2004), Jain (1989), Lim (1989), Gonzalez and Woods (1992), Myler and Weeks (1993), and Russ (1998), among many others. Frery and Perciano (2013) discuss these and other filters, and how to implement them, while Frery (2012) discusses their properties and applications to digital document processing.

Choosing or designing a smoothing filter requires a good knowledge of both the kind of data and noise, and of the desired effect.

Summary

Smoothing filters aim at reducing noise while improving the signal-to-noise ratio. They often achieve such a task at the expense of introducing blurring and loosing details.

Bibliography

- Barrett HH, Myers KJ (2004) Foundations of image science. Pure and applied optics. Wiley-Interscience, New York
- Frery AC (2012) Image filtering. In: Mello CAB, Oliveira ALI, Santos WP (eds) Digital document analysis and processing. Nova Science Publishers, pp 53–69. https://www.novapublishers.com/catalog/product_info.php?products_id=28733
- Frery AC, Perciano T (2013) Introduction to image processing with R: learning by examples. Springer. <http://www.springer.com/computer/image+processing/book/978-1-4471-4949-1>
- Gonzalez RC, Woods RE (1992) Digital image processing, 3rd edn. Addison-Wesley, Reading
- Jain AK (1989) Fundamentals of digital image processing. Prentice-Hall International Editions, Englewood Cliffs
- Lim JS (1989) Two-dimensional signal and image processing. Prentice hall signal processing series. Prentice Hall, Englewood Cliffs
- Myler HR, Weeks AR (1993) The pocket handbook of image processing algorithms in C. Prentice Hall, Englewood Cliffs, New Jersey
- Russ JC (1998) The image processing handbook, 3rd edn. CRC Press, Boca Raton
- Velho L, Frery AC, Miranda J (2008) Image processing for computer graphics and vision, 2nd edn. Springer, London. <https://doi.org/10.1007/978-1-84800-193-0>

Spatial Analysis

Qiyu Chen², Gang Liu^{1,2}, Xiaogang Ma³ and Xiang Que⁴

¹State Key Laboratory of Biogeology and Environmental Geology, Wuhan, Hubei, China

²School of Computer Science, China University of Geosciences, Wuhan, Hubei, China

³Department of Computer Science, University of Idaho, Moscow, ID, USA

⁴Computer and Information College, Fujian Agriculture and Forestry University, Fuzhou, China

Definition

Spatial analysis is a quantitative study of spatial phenomena in geography and even earth science. Its main ability is to

manipulate spatial data into different forms and extract and mine potential information and knowledge. Spatial analysis derives from geographic information science (GIScience) and has been expanded and applied to a wide range of fields, including geoscience, architecture, environmental science, etc. Spatial analysis has been a core of Geographic Information System (GIS). Spatial analysis capability, especially to the extraction and transmission of spatial hidden information, is the main sign of GISs that differ from other information systems.

Situation and Development of Spatial Analysis Technologies

Since the appearance of maps, people have consciously or unconsciously carried out various types of spatial analysis. For example, measuring the distance and area between geographic elements on the map, and using the map for tactical research and strategic decision-making. With the application of modern computer technology into cartography and geography, geographic information system (GIS) began to gestate and develop. Digital maps stored in the computer show a broader application field. Using computers to analyze maps, extract information, discover knowledge, and support spatial decision-making has become an important research content of GIS. “Spatial Analysis” has also become a specialized term in this field. Spatial analysis is the core and soul of GIS, and it is one of the main signs of GISs differing from other information systems (Cressie and Wike 2015). Spatial analysis, combined with the attribute information of spatial data, can provide powerful and abundant analysis and mining functions for spatial data. Therefore, the position of spatial analysis in GIS is self-evident. So far, GIS and spatial analysis technologies have been applied in a wide range of fields including earth science, architecture, environmental science, social science, etc. (Fotheringham and Rogerson 2013).

Spatial analysis aims to obtain derivative information and new knowledge from the relationship between spatial object. It is a complex process of extracting information from one or more spatial data sources (Burrough 2001). Spatial analysis uses geographic computing and spatial characterization to mine potential spatial patterns and knowledge. Its essence includes 1) detecting patterns in spatial data, 2) studying the relationship between multivariate data and establishing spatial data models, 3) making spatial data more intuitive to express its potential meaning, and 4) improving the prediction and control capabilities of geospatial events. Obviously, spatial analysis mainly uses the joint analysis of spatial data and spatial models to mine the potential information among spatial objects. The basic information of these spatial objects consists of their spatial location, distribution, form, distance,

orientation, topological relationship, etc. Among them, distance, orientation, and topological relationship constitute the spatial relationship of spatial objects. It is the spatial characteristics between geographic entities and can be used as the basis for data organization, query, analysis, and reasoning. Combining the spatial data and attribute data of the spatial objects can perform spatial calculation and analysis of many specialized tasks (De Smith et al. 2018). Many spatial analysis methods have been implemented in GIS software. For instance, a large number of spatial analysis tools have been integrated in ArcToolBox, such as spatial information classification, overlay, network analysis, neighborhood analysis, geostatistical analysis, etc. (Fischer and Getis 2009).

Research Objects

Spatial analysis is a general term for related methods of analyzing spatial data, and it is a sign of the advancement of GISs. Early GISs emphasized simple spatial query, and the analysis functions were really weak. With the development of GIS, users need more and more complex spatial analysis functions. These requirements promote the development of spatial analysis technology and also make a variety of spatial analysis functions. According to the different features of spatial data, it can be divided into: ① analysis operation based on spatial graphic data; ② data operation based on non-spatial attributes; ③ joint operation of spatial and non-spatial data. The basis of spatial analysis is geospatial data. Its main goal is to use a variety of geometric logic operations, mathematical statistical analysis, algebraic operations, and other mathematical methods to solve the actual problems of geographic space.

Main Contents

Spatial analysis involves a variety of contents, which can be summarized into the following types.

1. Spatial location: the location information of a spatial object is transferred by means of the spatial coordinate system. It is the foundation of spatial object characterization.
2. Spatial distribution: group positioning information of similar spatial objects, including distribution, trend, comparison, etc.
3. Spatial morphology: geometric form of spatial objects.
4. Spatial distance: the proximity of space objects.
5. Spatial relationship: related relationship of spatial objects, including topology, orientation, similarity, correlation, etc.

Basic Methods

Spatial Query and Measurement

Spatial query is to obtain the location and attribute information of the spatial objects according to certain conditions to form a new data set. Spatial query methods can be divided into the following types: a) Location query, b) Hierarchical query, c) Regional query, d) Conditional query, and e) Spatial relationship query. There are a variety of spatial relationships between spatial entities, including topological, sequence, distance, orientation, and so on. Querying and locating spatial entities through spatial relationships is one of the features of GISs that are different from other information systems. Generally, GIS software has the function of measuring points, lines, and areas. Spatial measurement has different meanings for different points, lines, and areas. a) Point features (0D): coordinates; b) Linear features (1D): length, direction, curvature; c) Area features (2D): area, perimeter, shape, etc.; d) Body features (3D): volume, surface area, etc.

Buffer Analysis

Buffer analysis is one of the spatial analysis tools to quantify the proximity problem of geospatial objects. It aims to automatically build a set of polygons around points, lines, areas, and other geographic entities. Proximity describes how close two geospatial objects are. In reality, buffer zones reflect a range of influence or service area of a geospatial object. According to the features of geospatial objects (point, line, area, and body), the corresponding buffer analysis methods have been developed.

Overlay Analysis

The overlay analysis of GIS is the operation of superimposing the data composed of related thematic layers to generate a new data layer. The result is a synthesis of the attributes of the original two or more layers. Overlay analysis includes not only the comparison of spatial relationships, but also the comparison of attribute relationships. Overlay analysis can be divided into the following categories: visual information overlay, points and polygons overlay, lines and polygons overlay, polygons overlay, and raster overlay. Logical operations are commonly used in overlay analysis, and logical expressions are used to analyze and process the logical relationships between geospatial objects from different layers.

Network Analysis

Network analysis is a basic model in operations. Its fundamental purpose is to study and plan how to arrange a network and make it run best (Okabe and Sugihara 2012). The analysis and modeling of geospatial networks (such as transportation

networks) and urban infrastructure networks (such as various network cables, power lines, telephone lines, water supply and drainage pipelines, etc.) are the main objects of GIS network analysis. The basic idea is that human activities always tend to choose the best spatial location according to a certain goal. Such problems are numerous in social and economic activities, so it is of great significance to study network problems in GIS. Network analysis includes: path analysis (seeking the best path), address matching (essentially a query of geospatial locations) and resource allocation.

Spatial Statistical Analysis

Due to the interaction between spatial phenomena in different directions and different distances, traditional mathematical statistics cannot well solve the problems of spatial sampling, spatial interpolation, and the relationship between two or more spatial datasets. Therefore, the methods of spatial statistical analysis emerged. In the 1960s, Matheron formed a new branch of statistics through a lot of theoretical research, namely, spatial statistics, also called geostatistics (Matheron 1963).

Spatial statistics is based on the theory of regionalized variables. It uses the variograms as the main tool to study the spatial interaction and change laws of geospatial objects or phenomena (Unwin 1996). The spatial statistical methods assume that all the values in the study area are non-independent and correlated with each other. In the context of space or time, it is called autocorrelation. By detecting whether the variation in a position depends on the variation of other positions in its neighborhood, it can be used to judge whether the variation bears spatial autocorrelation. The important task of spatial statistical analysis is to reveal the correlation rules of spatial data and use these rules to predict unknown points. Spatial statistics contains two significant tasks: 1) analyzing the variograms or covariograms of spatial variability and structures and 2) Kriging interpolation for spatial local estimation.

Spatial interpolation is to generate a continuous surface using samples collected at different locations. Kriging is the most common and widely used spatial interpolation method. Kriging and its variants are based on the theoretical analysis of variogram or covariogram. They are methods for unbiased and optimal estimation of regionalization variables in a limited area.

Conclusions and Outlook

Spatial analysis is a spatial data analysis technique based on the location and morphology of geospatial objects. Its

purpose is to discover, extract, mine, and transmit potential spatial information and knowledge. Spatial analysis technology has been applied in many fields including geography, geology, surveying and mapping, architecture, environmental science, and social science. Different application fields present different meanings to spatial analysis. Their emphasis is different, but they all explain the connotation of spatial analysis from different aspects: either focusing on geometric analysis, or focusing on geostatistics and modeling. In summary, GIS spatial analysis is a technology that uses geometric analysis, statistical analysis, mathematical modeling, geographic calculation, and other methods to describe, analyze, and model the spatial relationship of in geospatial entities, and further provide services for spatial decision.

Especially, spatial statistics or geostatistics has become the basis of quantitative research in geosciences. Geostatistics is widely used in the characterization of complex heterogeneous geological phenomena and structures, mapping the prospect of mineral resources, reserve calculation of mineral resources, reservoir prediction and simulation, etc. With the further development of artificial intelligence and big data technologies, using artificial neural networks, machine learning, and deep learning technologies to mine and analyze the deeper correlations and knowledge hidden behind the massive data has become the hotspots and key issues in the joint field of mathematical geoscience and GIS.

Cross-References

- [Geostatistics](#)
- [Interpolation](#)
- [Spatial Statistics](#)

Bibliography

- Burrough PA (2001) GIS and geostatistics: essential partners for spatial analysis. *Environ Ecol Stat* 8(4):361–377
- Cressie N, Wikle C (2015) *Statistics for spatio-temporal data*. Wiley, Hoboken
- De Smith M, Longley P, Goodchild M (2018) *Geospatial analysis – a comprehensive guide*, 6th edn. The Winchelsea Press
- Fischer MM, Getis A (eds) (2009) *Handbook of applied spatial analysis: software tools, methods and applications*. Springer Science & Business Media
- Fotheringham S, Rogerson P (eds) (2013) *Spatial analysis and GIS*. CRC Press
- Matheron G (1963) Principles of geostatistics. *Econ Geol* 58(8):1246–1266
- Okabe A, Sugihara K (2012) *Spatial analysis along networks: statistical and computational methods*. Wiley, Hoboken
- Unwin DJ (1996) GIS, spatial analysis and spatial statistics. *Prog Hum Geogr* 20(4):540–551

Spatial Autocorrelation

Donato Posa and Sandra De Iaco

Department of Economics, Section of Mathematics and Statistics, University of Salento, Lecce, Italy

Synonyms

Covariance function; Spatial correlation; Variogram

Definition

Spatial autocorrelation is an assessment of the correlation between two random variables which describe the same aspect of the phenomenon under study, referred to two locations of the domain. The suffix “auto” is justified since in some sense the spatial autocorrelation quantifies the correlation of a variable with itself over space. However, the expressions “spatial autocorrelation” and “spatial correlation” are used interchangeably in literature. Most of the theoretical results of classical statistics consider the observed values, as independent realizations of a random variable: this assumption makes the statistical theory much easier. Unfortunately, this last hypothesis cannot be valid if the observations are measured in space (or in time). Historically, the concept of spatial autocorrelation has naturally been considered an extension of temporal autocorrelation: however, time is one-dimensional, and only goes in one direction, ever forward, on the other hand, the physical space has (at least) two dimensions.

Any subject which regards data collected at spatial locations, such as soil science, ecology, atmospheric science, geology, image processing, epidemiology, forestry, astronomy, needs suitable tools able to analyze dependence between observations at different locations.

A Brief Overview

In the last years, spatial data analysis has developed peculiar methods to provide a quantitative description of several spatial variables in various fields of applications, from environmental science to mining engineering, from medicine to biology, from economics to finance and insurance.

The present contribution will be devoted to spatial autocorrelation for *geostatistical data*: indeed, besides *geostatistics*, distributions on *lattices* and *spatial point processes* are considered further branches of spatial statistics. In

particular, geostatistical data refer to observations related to an underlying spatially continuous phenomenon, i.e., to measurements which can be taken at any point of a given spatial domain and utilize the formalism of the random functions (RFs) (Christakos 2005; Journel and Huijbregts 1981; Matheron 1971). Most of the applications use the variogram or the covariance function to characterize spatial correlation. A further and more general development to geostatistical methods has been provided by the theory of the *intrinsic random functions of order k* (Matheron 1973) and by *multiple point geostatistics* (Krishnan and Journel 2003).

Although real-valued covariance functions have been mostly utilized in several applications, it should be properly underlined that a covariance, according to Bochner's theorem, is a complex valued function: this relevant aspect is often neglected. For this purpose, in recent years, a significant and more general contribution to the theory of spatial autocorrelation has been given thanks to the formalism of *complex RFs*, which generalize the theory of real RFs (Posa 2020); indeed, the theory of complex valued RFs provides an important contribution in the study of various phenomena in oceanography, signal analysis, meteorology, and geophysics. In particular, in the last part of this entry, it will be underlined which tools of spatial autocorrelation are more appropriate and suitable, according to the various and different case studies.

Real Valued Random Functions

Geostatistical techniques rely on the theory of RFs: in particular, in the whole first part of the entry the discussion will be devoted to real valued RFs, whereas in the last part of the entry, a brief outline on complex valued RFs will also be given.

Consider the value $z(\mathbf{s})$, $\mathbf{s} \in D$, of a spatial variable z (which belongs to the set of the real numbers), where $\mathbf{s}$ is the vector of the spatial location and D is the spatial domain. Any sampled measure $z(\mathbf{s})$ with $\mathbf{s} \in D$ is considered as a realization of the random variable $Z(\mathbf{s})$. The set of all random variables over the domain D is called RF. Hence, the sampled values of a spatial phenomenon can be interpreted as a finite realization of a spatial RF (SRF), i.e.,

$$\{Z(\mathbf{s}); \mathbf{s} \in D\}, D \subseteq \mathbb{R}^n, \quad n \in \mathbb{N}_+. \quad (1)$$

By assuming the existence of the first and second order moments for Z , the SRF can be decomposed as follows:

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + Y(\mathbf{s}), \quad (2)$$

where $\mu(\mathbf{s}) = E[Z(\mathbf{s})]$ is the expected value of Z , usually called drift, and Y is a zero mean SRF, which describes the random

fluctuations (or the micro scale variability) of the phenomenon under study around the expected value.

In applied sciences usually the distribution function which characterizes the SRF is unknown and cannot be derived empirically due to the small number of realizations (in most cases only one sequence of measurements is available). Thus, the characterization of the SRF is limited to its statistical moments of order up to two. The part of the general theory that studies the properties of a SRF utilizing only the first and second order moments is called *correlation theory*. In particular, the expected value μ (moment of first order), the covariance C , and the semivariogram γ (moments of second order) of an SRF Z are defined, respectively, as follows:

$$\mu(\mathbf{s}) = E[Z(\mathbf{s})], \quad \mathbf{s} \in D, \quad (3)$$

$$C(\mathbf{s}_i, \mathbf{s}_j) = E[(Z(\mathbf{s}_i) - \mu(\mathbf{s}_i))(Z(\mathbf{s}_j) - \mu(\mathbf{s}_j))], \quad \mathbf{s}_i \in D, \quad \mathbf{s}_j \in D, \quad (4)$$

$$2\gamma(\mathbf{s}_i, \mathbf{s}_j) = \text{Var}[(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))], \quad \mathbf{s}_i \in D, \quad \mathbf{s}_j \in D. \quad (5)$$

Note that 2γ is called *variogram* and differs from the semivariogram by the constant.

Stationarity

Stationarity is a prior decision which refers to the SRF, not to the sampled values; this hypothesis makes inference possible from the observed values over the spatial domain D . The SRF Z is said to be stationary of order two, if its expected value exists and is independent of the location, its covariance C exists and depends only on the separation vector $\mathbf{h}$ between two spatial locations, i.e.,

$$E[Z(\mathbf{s})] = \mu, \quad \mathbf{s} \in D, \quad (6)$$

$$C(\mathbf{h}) = E[(Z(\mathbf{s} + \mathbf{h}) - \mu)(Z(\mathbf{s}) - \mu)], \quad \mathbf{s} \in D, \mathbf{s} + \mathbf{h} \in D. \quad (7)$$

As a consequence, even the semivariogram γ exists and depends only on the separation vector $\mathbf{h}$ between two spatial locations, i.e.,

$$2\gamma(\mathbf{h}) = E[(Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s}))^2], \quad \mathbf{s} \in D, \mathbf{s} + \mathbf{h} \in D. \quad (8)$$

In case of a stationary random field, the covariance function, which is represented by a one-parameter function, is called *covariogram*. The covariance function usually decreases when the distance between two spatial locations increases; on the other hand, the semivariogram, by definition, is a variance, hence it usually increases when the distance between two spatial locations increases. It is well known that the covariance C and the semivariogram γ are connected through the following relationship:

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}). \quad (9)$$

A detailed and interesting overview and comprehensive discussion about stationarity is given in Myers (1989). Note also that the variogram and the covariance functions are the main tools utilized to describe the spatial dependence; moreover, the corresponding models are used to make predictions.

Intrinsic Hypotheses

A weaker stationarity hypothesis, often recalled in the literature, assumes that, for every vector $\mathbf{h}$, the increment $(Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s}))$ is second order stationary: in this case, Z is called an intrinsic RF (IRF), i.e.,

$$E[(Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s}))] = 0, \quad 2\gamma(\mathbf{h}) = E[(Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s}))^2].$$

Hence, the processes that satisfy the hypotheses of second-order stationarity are a subset of the processes that satisfy the intrinsic hypotheses. This aspect justifies why the semivariogram function γ is more utilized than the covariance function in several applications.

Intrinsic Random Functions of Order k

In geostatistics, it is well known that if there is a trend in the data, then a bias affects the variogram of the residuals in universal kriging. Historically, this has been the starting point to construct models in a such a way to reach stationarity by considering increments of higher order than the ones defined through the intrinsic hypotheses: indeed, these last are defined as IRF of order zero. Through the IRF of order k (IRF k) theory (Matheron 1973) a wider class of covariance functions (the generalized covariance functions) has been obtained and applied in order to overcome the bias, above mentioned, and to reach stationarity through a suitable order of increments. The theory of IRF k generalizes the approach utilized in time series (ARIMA models): indeed, Z is an IRF k , with $k \in \mathbb{N}$, if:

$$W(\mathbf{s}) = \sum_{i=1}^m \lambda_i Z(\mathbf{s}_i + \mathbf{s}), \quad \mathbf{s} \in D,$$

is second order stationary, where $\mathbf{s}_i \in D$, $i = 1, \dots, m$, and $A = (\lambda_1, \dots, \lambda_m)^T$ is a vector of real numbers such that $(f_j(\mathbf{s}_1), \dots, f_j(\mathbf{s}_m))A = 0$, where f_j , $j = 1, \dots, k + 1$, are mixed monomials.

The generalized covariances, which characterize IRF k , are even functions and are unique up to an even polynomial of degree $2k$. In summary, the RFs which are second order stationary are a subset of the RFs which are intrinsically stationary, i.e., IRF of order 0 (IRF0), and these last are a subset of IRF1 and so forth. A detailed discussion on IRF k , on the properties and construction of generalized covariances

can be found in Chiles and Delfiner (1999), in Christakos (1984) and in Cressie (1993).

Properties of Covariance and Semivariogram Functions

The family of the real covariance functions satisfies the following properties:

$$|C(\mathbf{h})| \leq C(\mathbf{0}), \quad C(\mathbf{h}) = C(-\mathbf{h}), \quad (10)$$

and then $C(\mathbf{0}) \geq 0$.

If C_i , $i = 1, \dots, n$ are covariance functions defined in $\mathbb{R}^n$ and a_i , $i = 1, \dots, n$ are non-negative coefficients, then C_P and C_S , defined as follows:

$$C_P(\mathbf{h}) = \prod_{i=1}^n C_i(\mathbf{h}), \quad C_S(\mathbf{h}) = \sum_{i=1}^n a_i C_i(\mathbf{h})$$

are covariance functions.

If $C_n(\mathbf{h})$, $n \in \mathbb{N}$ are covariance functions in $\mathbb{R}^n$, then $C(\mathbf{h}) = \lim_{n \rightarrow \infty} C_n(\mathbf{h})$ is a covariance function, if the limit exists.

If $C(\mathbf{h}; a)$ is a covariance in $\mathbb{R}^n$ for all values of $a \in A \subseteq \mathbb{R}$ and if $\mu(da)$ is a positive measure on A , then $\int_A C(\mathbf{h}; a) \mu(da)$

is a covariance function, if the integral exists for all $\mathbf{h} \in D$.

Moreover, the condition for a function to be a covariance is that it be positive definite, namely:

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j C(\mathbf{s}_i - \mathbf{s}_j) \geq 0 \quad (11)$$

for any $\mathbf{s}_i \in D$, any $a_i \in \mathbb{R}$, $i = 1, \dots, n$, and any positive integer n . The function C is strictly positive definite if it excludes the chance that the quadratic form in (11) is equal to zero $\forall n \in \mathbb{N}_+$, any choice of distinct points $(\mathbf{s}_i)$, $\forall a_i \in \mathbb{R}$ (with at least one different from zero). Note that strict positive definiteness is desirable, since it ensures the invertibility of the kriging coefficient matrix; a thorough discussion on strict positive definiteness for covariance functions is given in De Iaco and Posa (2018). A semivariogram function satisfies the following properties:

$$\gamma(\mathbf{h}) = \gamma(-\mathbf{h}), \quad \gamma(\mathbf{0}) = 0, \quad \lim_{\|\mathbf{h}\| \rightarrow \infty} \frac{\gamma(\mathbf{h})}{\|\mathbf{h}\|^2} = 0;$$

if γ_i , $i = 1, \dots, n$ are semivariogram functions defined in $\mathbb{R}^n$ and a_i , $i = 1, \dots, n$ are non-negative coefficients, then $\gamma_S(\mathbf{h}) = \sum_{i=1}^n a_i \gamma_i(\mathbf{h})$ is a semivariogram function. Moreover, a semivariogram is a conditional negative definite function, namely,

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j \gamma(\mathbf{s}_i - \mathbf{s}_j) \leq 0, \quad \sum_{i=1}^n a_i = 0 \quad (12)$$

for any $\mathbf{s}_i \in D$, $a_i \in \mathbb{R}$, $i = 1, \dots, n$, and any positive integer n .

Although a number of studies provide theoretical results in terms of the covariance, the spatial correlation is usually analyzed through the variogram, which is preferred to the covariance for different reasons (Cressie 1993).

Peculiar Features of Spatial Autocorrelation

Some relevant peculiar characteristics, such as isotropy, separability, and symmetry, concerning spatial autocorrelation will be described hereafter and will be given in terms of the covariance function. Note that the definition of isotropy can also be given in terms of the variogram; however, the definition of separability can only be described in terms of the covariance function. These special features, which are described in detail in (De Iaco et al. 2019), play an important role for the choice of a suitable correlation model for the spatial variable.

Isotropy/Anisotropies

Geostatistical applications usually utilize, although very restrictive, the hypothesis of isotropy. Let Z be a SRF which is second order stationary and let C be its covariance function. Then Z is *isotropic* on $\mathbb{R}^n$ ($n \in \mathbb{N}_+$), if there exists a function C_0 such that:

$$C(\mathbf{s}_1 - \mathbf{s}_2) = C_0(\|\mathbf{h}\|), \quad \mathbf{s}_1, \mathbf{s}_2 \in D, \mathbf{s}_1 - \mathbf{s}_2 = \mathbf{h}, \quad (13)$$

where $\|\cdot\|$ indicates the norm in the Euclidean space $\mathbb{R}^n$. Definition (13) can also be given in terms of the variogram.

If C is an isotropic covariance function in $\mathbb{R}^n$, then C is also an isotropic covariance function in $\mathbb{R}^m$, with $m < n$; the same property holds for a variogram function γ . However the converse could not be true.

As a consequence, the definition of *anisotropy* stems from the rejection of (13). In Yaglom (1987), some relevant properties concerning the Eq. (13) are described in detail. In classical geostatistics, the class of anisotropic SRF has been often classified in two main subsets, according to two different kinds of anisotropy: geometric anisotropy and zonal anisotropy, which can be equally defined in terms of the covariance or variogram. The first subset includes SRF characterized by a variogram function which presents the same sill value in all directions, but the range changes with the spatial direction. Hence, given an isotropic covariance function C_0 , it can be transformed to a geometric anisotropic covariance function C through a linear transformation of its lag vector

$\mathbf{h} \in \mathbb{R}^n$, i.e., $C(\mathbf{h}) = C_0(\|\mathbf{A}\mathbf{h}\|)$, where the matrix $\mathbf{A}$ is obtained as the product of a rotation matrix with a scaling parameters diagonal matrix.

On the other hand, according to Journel and Huijbregts (1981), by setting equal to zero some values of the previous diagonal matrix, a *zonal anisotropy* is obtained; hence, zonal anisotropy can be defined as a special case of geometric anisotropy.

A covariance function, characterized by zonal anisotropy, can be obtained through the sum of covariance functions, where each covariance is defined on a proper subspace of the spatial domain. As shown hereafter, the following covariance function in $\mathbb{R}^n$ is given as the sum of the models C_1 and C_2 , i.e.,

$$C(\mathbf{h}) = C_1(\mathbf{h}_1) + C_2(\mathbf{h}_2), \quad \mathbf{h} = (\mathbf{h}_1, \mathbf{h}_2),$$

where $\mathbf{h}_1 \in \mathbb{R}^{n_1}$ and $\mathbf{h}_2 \in \mathbb{R}^{n_2}$, with $\mathbb{R}^n = \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$.

Zimmerman (1993) introduced a peculiar classification of anisotropy. In particular, he dropped out the definition of “zonal anisotropy” for the more explanatory terms such as “range anisotropy” (also slope anisotropy), “sill anisotropy,” and “nugget anisotropy,” and to save the classical term “geometric anisotropy,” which is interpreted as a special case of range anisotropy.

Separability

Let C be a covariance function on $\mathbb{R}^n$;

- if $C(\mathbf{h}) = C_1(h_1)C_2(h_2) \cdots C_n(h_n)$, and each $C_i: \mathbb{R} \rightarrow \mathbb{R}$ is a covariance function, $i = 1, \dots, n$, then C is *fully separable*;
- if $C(\mathbf{h}) = C_1(\mathbf{h}_1)C_2(\mathbf{h}_2) \cdots C_k(\mathbf{h}_k)$, $k < n$, and each C_i is a covariance function on $\mathbb{R}^{n_i}$, $i = 1, \dots, k$, with $\mathbb{R}^n = \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_k}$, then C is *partially separable*.

If a covariance function is fully separable, then it is partially separable, whereas the converse is not true.

Symmetry

Symmetry is a further characteristic of a SRF which is second order stationary; in particular:

- A SRF is *axially symmetric*, if $C(h_1, \dots, h_j, \dots, h_n) = C(-h_1, \dots, h_j, \dots, h_n) = C(h_1, \dots, -h_j, \dots, h_n)$ for a given j . This last property is named *axial or reflection symmetry* in $\mathbb{R}^2$;
- A SRF is *quadrant symmetric*, if $C(h_1, \dots, h_j, \dots, h_n) = C(h_1, \dots, -h_j, \dots, h_n)$ for all $j = 1, \dots, n$.

In particular, in $\mathbb{R}^2$, a SRF is

- *Laterally or diagonally symmetric*, if $C(h_1, h_2) = C(h_2, h_1)$, $(h_1, h_2) \in \mathbb{R}^2$.
- *Complete symmetric*, if both diagonal and reflection symmetric properties are satisfied, i.e., $C(h_1, h_2) = C(-h_1, h_2) = C(h_2, h_1) = C(-h_2, h_1)$, $(h_1, h_2) \in \mathbb{R}^2$.

Note that the set of isotropic covariance functions is a subset of the class of symmetric covariance functions.

Remarks

Symmetry, strict positive definiteness, and separability do not require any form of stationarity for the SRF. On the other hand, the property of isotropy is appropriate only for second order stationary SRFs. A covariance function can be separable and isotropic if and only if all the factors of the product model are Gaussian covariance functions (De Iaco et al. 2020).

Covariance and Variogram Models

Given the difficulties to verify conditions (11) and (12), it is advisable for users to look for the best model of spatial correlation, for a given case study, among the wide parametric families whose members are known to respect the above conditions. Specific details about the construction and characteristics of covariance and variogram models are given in Chiles and Delfiner (1999), Cressie (1993), and Journel and Huijbregts (1981).

Several parametric families of covariance and variogram models can be found in the previous mentioned geostatistical textbooks. Among the various variogram models, the most utilized, such as the *exponential*, the *gaussian*, the *rational quadratic*, and the *power* model, are valid in $\mathbb{R}^n$, $n \geq 1$, whereas the *spherical* model is valid in $\mathbb{R}^n$, with $n \leq 3$. With the exception of the power model which is unbounded, all the previous models are bounded. Spatial correlation with an alternance of positive and negative values, caused by a periodicity of the spatial phenomenon, is often modeled by the *hole effect* or *wave* variogram.

Some isotropic semivariogram models proposed in the literature for a random function that satisfies at least the intrinsic hypotheses are reported below. For simplicity of notation, the semivariogram is assumed to be dependent on the modulus of $\mathbf{h}$, that is $\gamma(\mathbf{h}) = \gamma(h)$, where $h = \|\mathbf{h}\|$.

• Spherical model.

$$\gamma(h) = \begin{cases} C \left[1, 5 \frac{h}{a} - 0, 5 \left(\frac{h}{a} \right)^3 \right] & 0 \leq h \leq a \\ C & h > a \end{cases} \quad (14)$$

where $a \in \mathbb{R}_+$ is the *range* and $C \in \mathbb{R}_+$ is the *sill* value.

• Exponential model.

$$\gamma(h) = C[1 - \exp(-h/a)], \quad (15)$$

where $a \in \mathbb{R}_+$ is the *range* and $C \in \mathbb{R}_+$ is the *sill* value. This model reaches the sill value asymptotically, then the value a' for which $\gamma(a') = 0, 95C$ is such that $a' = 3a$ and it is called *effective range*.

• Gaussian model.

$$\gamma(h) = C[1 - \exp(-h^2/a^2)], \quad (16)$$

where $a \in \mathbb{R}_+$ is the *range* and $C \in \mathbb{R}_+$ is the *sill* value. Similarly to the previous model, it reaches the sill value asymptotically and the value a' for which $\gamma(a') = 0, 95C$ is such that $a' = \sqrt{3}a$.

• Power model.

$$\gamma(h) = \omega h^\alpha, \quad (17)$$

with $\alpha \in [0, 2]$.

This function is concave for $0 < \alpha \leq 1$, while it is convex for $1 < \alpha < 2$.

- **Hole effect model.** The *hole effect* models are used to describe the behavior of a semivariogram which does not grow monotonically, as h increases.

– *Hole Effect 1.*

$$\gamma(h) = C[1 - \cos(h/a)], \quad (18)$$

with $a \in \mathbb{R}_+$ and $C \in \mathbb{R}_+$. This model, valid in $\mathbb{R}$, presents a hole effect, which is described by a periodic component, without damping in the oscillations.

– *Hole Effect 2.*

$$\gamma(h) = C[1 - \exp(-h/a_1) \cos(h/a_2)], \quad (19)$$

with $a_1, a_2 \in \mathbb{R}_+$ and $C \in \mathbb{R}_+$. This model is valid in $\mathbb{R}$; moreover, it is valid in $\mathbb{R}^2$ if and only if $a_2 \geq a_1$ and in $\mathbb{R}^3$ if and only if $a_2 \geq a_1\sqrt{3}$ (Yaglom 1987).

– *Hole Effect 3.*

$$\gamma(h) = C[1 - (a/h) \sin(h/a)], \quad (20)$$

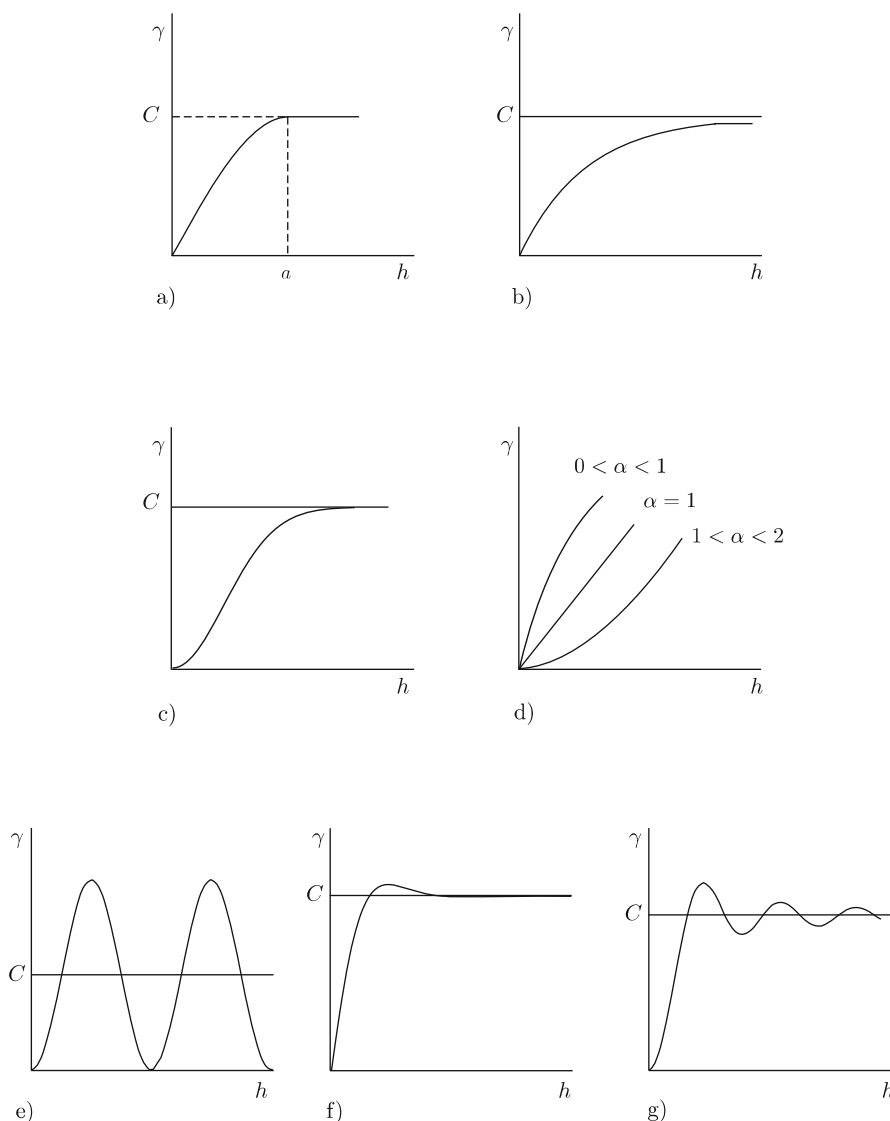
with $a \in \mathbb{R}_+$ and $C \in \mathbb{R}_+$. This model is valid in $\mathbb{R}^3$ and similarly to the previous model presents a damping effect in the oscillations.

Figure 1 shows the semivariogram models analytically described above.

Further classes and specific properties of correlation models, which are constructed through the modified Bessel functions, can be found in Matern (1980).

Spatial Autocorrelation,

Fig. 1 (a) Spherical model; (b) Exponential model; (c) Gaussian model; (d) Power model; (e) hole effect 1 model; (f) hole effect 2 model; (g) hole effect 3 model

**Characteristics of Spatial Autocorrelation**

In the present section some characteristics of the variogram functions are briefly discussed, essentially because they are useful to choose an appropriate model of spatial correlation. At this purpose, the behavior near the origin and the behavior for large distances are often relevant issues.

Sill and Range

For what concerns the behavior of a semivariogram for large distances, two typical situations are usually encountered: the semivariogram systematically increases at large distances, otherwise it stabilizes around a particular value, which is called sill, while the corresponding distance at which the sill is reached is called range. In this last case, for distances greater than the range, there is absence of spatial correlation.

Nested Structures

The variogram can reveal nested structures, that is, hierarchical structures, each characterized by its own range (Journel and Huijbregts 1981).

Behavior Near the Origin

The importance of the behavior of a variogram close to the origin is related to the regularity of the spatial variable. In particular, a *parabolic behavior* is typical of a spatial variable which is very regular and the corresponding RF is differentiable in the mean square sense. A *linear behavior* is typical of a spatial variable which is not so regular as in the previous case, while the corresponding RF is continuous but not differentiable in the mean square sense. If the semivariogram presents a discontinuity at the origin, called *nugget effect*, then the spatial variable is very irregular. The denomination

nugget effect originates from the strong variability presented in gold deposits. However, the discontinuity of a variogram at the origin can be caused by several reasons, well described in the classical geostatistical references (Journel and Huijbregts 1981). At last, if the variogram function is a flat curve, then this behavior is typical of absence of spatial correlation.

Hole Effect

If the variogram function presents some bumps for some distances, which correspond to negative values (holes) of the covariance function, then the physical interpretation of this behavior derives from the presence of couples of locations, at the same distances, where relatively low values of the spatial variable are combined with relatively high values and vice versa.

Periodicities

As well known, temporal observations frequently present a periodic behavior, because of the natural influence of cycles which characterize human activities and natural phenomena: this periodic behavior is captured by a correlation measure. On the other hand, spatial variables are rarely characterized by periodicities, except in some peculiar cases, such as the behavior of sedimentary rocks.

Presence of a Drift

If the sample variogram increases faster than a parabola for large distances, then this behavior reflects the presence of a trend in the data, hence the expected value of the SRF cannot be assumed constant over the spatial domain.

Structural Analysis

The whole geostatistical analysis can be summarized through the following steps:

1. Look for a model of the SRF from which data could reasonably be derived.
2. Estimate the variogram as a measure of the spatial correlation exhibited by the data.
3. Fit a suitable model to the estimated variogram.
4. Use this last model in the kriging system for spatial prediction.

In particular, structural analysis is executed on the realizations of a second order stationary SRF. These realizations correspond directly to the observed values (if the macro scale component can be reasonably supposed constant over the domain) or to the residuals otherwise.

Structural analysis begins with the estimation of the semi-variogram or covariogram of the residual SRF Y defined in (2). Given the set $A = \{\mathbf{s}_i, i = 1, 2, \dots, n\}$ of data locations, the semi-variogram can be estimated through the sample semi-variogram $\hat{\gamma}$ as follows:

$$\hat{\gamma}(\mathbf{r}_s) = \frac{1}{2|L(\mathbf{r}_s)|} \sum_{L(\mathbf{r}_s)} [Y(\mathbf{s} + \mathbf{h}) - Y(\mathbf{s})]^2 \quad (21)$$

where $|L(\mathbf{r}_s)|$ is the cardinality of the set $L(\mathbf{r}_s) = \{(\mathbf{s} + \mathbf{h}) \in A, (\mathbf{s}) \in A : \mathbf{h} \in \text{Tot}(\mathbf{r}_s)\}$, and $\text{Tot}(\mathbf{r}_s)$ is a specified tolerance region around $\mathbf{r}_s$. Similarly, the sample covariogram $\hat{C}$ is defined below:

$$\hat{C}(\mathbf{r}_s) = \frac{1}{|L(\mathbf{r}_s)|} \sum_{L(\mathbf{r}_s)} [Y(\mathbf{s} + \mathbf{h}) - \bar{Y}][Y(\mathbf{s}, t) - \bar{Y}], \quad (22)$$

where $\bar{Y}$ is the sample mean.

The second step of structural analysis consists in fitting a theoretical model to the sample spatial semi-variogram or covariogram. As already underlined in one of the previous sections, various classes of covariance functions or semi-variograms are available. Variogram or covariogram modeling has been a research focus in this field for a long-time. Maximum likelihood and least squares are the two common ways to achieve this model fitting goal. Some details on the fitting aspects can be found in Cressie (1993).

After modeling the spatial empirical variogram/covariogram surface, the subsequent step is to evaluate the reliability of the fitted model through the application of cross-validation and jackknife techniques. Then, if the model performance is satisfactory, the same model can be used to predict the variable under study over the spatial domain by using kriging.

Through the cross-validation technique all the observed values are removed, one at a time, and at each location where the observed value has been temporarily removed, the spatial variable is estimated by utilizing all the remaining observed values and the selected variogram model. At last, the estimated values are compared with the observed ones through the correlation coefficient. On the other hand, the technique of jackknife operates in a different way; in particular, two different data sets of observed values are considered. The first data set is the one used in structural analysis to estimate and model the correlation function (variogram); this last model is then used to predict the spatial variable at the locations of the second data set. Hence, these last estimated values are compared with the true values (which are known) of the second data set, called for this reason, control data set.

From a computational point of view, the literature offers various packages which perform structural analysis in two and three dimensions, such as the Geostatistical Software Library GSLIB and its version for Windows WinGSLIB (Deutsch and Journel 1998), SGeMS (Remy et al. 2009), as well as the commercial software ISATIS. Moreover, in the R environment, there are various packages devoted to geostatistical modeling, such as gstat and RGeostats (Pebesma and Wesseling 1998; Renard et al. 2014).

Multiple Point Statistics

As previously pointed out, the analysis of spatial correlation in classical geostatistics is based on the variogram or covariance function. However, these tools, which are known as two point statistics, could not be appropriate for modeling phenomena which present curvilinear structures or, more generally, complex spatial patterns. At this purpose, multiple point statistics (Krishnan and Journel 2003) represent a solution to overcome these difficulties since the relations among the variables at more than two points at a time are considered. Unfortunately, spatial data are usually sparse, hence they are not very informative for the application of multiple point statistic techniques. These last problems can be overcome by utilizing training images from which multiple point statistics can be provided.

Complex Random Functions

Complex RFs, as a generalization of real RFs, can be useful to model vectorial data in $\mathbb{R}^2$: for example, a wind field can be viewed as a complex variable by considering the intensity of the wind speed and its direction at each spatial location (De Iaco and Posa 2016). Let $\mathbb{R}^n$ and $\mathbb{C}$ be the n -dimensional Euclidean space and the set of complex numbers, respectively. Let Z_1 and Z_2 be the real components of the following complex RF Z :

$$Z(\mathbf{s}) = Z_1(\mathbf{s}) + iZ_2(\mathbf{s}), \mathbf{s} \in \mathbb{R}^n,$$

where $i^2 = -1$, with i the imaginary unit.

If Z is a second order stationary RF, its covariance function is defined as follows:

$$C(\mathbf{h}) = E[(Z(\mathbf{s}) - m)\overline{(Z(\mathbf{s} + \mathbf{h}) - m)}]$$

and it can be expressed in the following form: $C(\mathbf{h}) = C_{re}(-\mathbf{h}) + iC_{im}(\mathbf{h})$, where

$$C_{re}(\mathbf{h}) = C_{Z_1}(\mathbf{h}) + C_{Z_2}(\mathbf{h}), C_{im}(\mathbf{h}) = C_{Z_2Z_1}(\mathbf{h}) - C_{Z_1Z_2}(\mathbf{h})$$

C_{re} is an even function and it is a real covariance function, whereas C_{im} is an odd function and it is not a covariance function.

Note that the following properties generalize the previous ones given in (10) for a real valued covariance function: $C(-\mathbf{h}) = \overline{C(\mathbf{h})}$; $C(\mathbf{0}) \geq 0$; $|C(\mathbf{h})| \leq C(\mathbf{0})$.

The theorem of Bochner (1959) specifies that any continuous covariance function can be represented as follows:

$$C(\mathbf{h}) = \int_{\mathbb{R}^n} \exp(i\boldsymbol{\omega}^T \mathbf{h}) dF(\boldsymbol{\omega}), \quad (23)$$

where F is a non-negative and finite measure on $\mathbb{R}^n$. Hence, according to the previous theorem, a covariance is a complex valued function.

A detailed discussion on the construction of complex and, in particular, of real covariance functions has been recently given in Posa (2020). In particular, the symmetry of the spectral distribution function or the property of the spectral density function to be an even function, guarantee that the covariance is a real function. Otherwise the covariance is a complex function.

Summary and New Results

In the present section, a summary on spatial autocorrelation, as well as some recent and significant results will be provided.

- One of the main issues, which often requires a specific answer, concerns the measure of spatial autocorrelation to be utilized for the case study at hand. Indeed, the answer to this kind of question is not so easy, as could appear or as often underlined in the literature. In fact, in most of the geostatistical applications the variogram is preferred to the covariance function, essentially because the first function satisfies the intrinsic hypotheses, which describe a wider class of spatial RFs respect to the processes which satisfy the hypotheses of second order stationarity. Several papers provide a comparison between these two correlation functions and the above choice is often justified by further reasons (Cressie and Grondona 1992).
- On the other hand, the variogram, by definition, is a variance, hence it is a real valued function; according to Bochner's theorem, the covariance is a complex valued function and it is suitable to analyze vectorial data in two dimensions, as shown in some interesting applications concerning phenomena which often happen in oceanography, environmental and ecological sciences, geophysics, in meteorology, and signal analysis.
- If the spatial variable is a scalar quantity, the variogram or, more generally, the formalism of the IRFk could be a suitable choice for the measure of spatial autocorrelation.
- In presence of curvilinear or complex spatial patterns, if a training image is available, then the techniques of multiple point statistics are recommended, as shown in several applications.
- Wide classes of parametric families of complex covariance functions have been recently constructed (Posa 2020). These complex covariance models could be applied in a spatial context, in several dimensional spaces, as well as in

a spatiotemporal domain to model the correlation structure of complex valued random fields.

- Although it is well known that the difference of two covariance functions could not be a covariance function, several classes of models for the difference of two covariance functions have been recently constructed in the complex domain and in the subset of the real domain (Posa 2021). All the relevant issues and consequences which stem from this last result are also discussed in the same paper.

Cross-References

- [Multiple Point Statistics](#)
- [Markov Random Fields](#)
- [Stationarity](#)
- [Variance](#)

References

- Bochner S (1959) Lectures on Fourier integrals. Princeton University Press, Princeton, 338 p
- Chiles J, Delfiner P (1999) Geostatistics. Wiley, New York, 687 p
- Christakos G (1984) On the problem of permissible covariance and variogram models. *Water Resour Res* 20(2):251–265
- Christakos G (2005) Random field models in earth sciences. Dover, Mineola, 512 p
- Cressie N (1993) Statistics for spatial data. Wiley, New York, 900 p
- Cressie N, Grondona M (1992) A comparison of variogram estimation with covariogram estimation. In: Mardia KV (ed) The art of statistical science: a tribute to G.S. Watson. Wiley, Chichester, pp 191–208
- De Iaco S, Posa D (2016) Wind velocity prediction through complex kriging: formalism and computational aspects. *Environ Ecol Stat* 23(1):115–139
- De Iaco S, Posa D (2018) Strict positive definiteness in geostatistics. *Stoch Env Res Risk A* 32:577–590
- De Iaco S, Posa D, Cappello C, Maggio S (2019) Isotropy, symmetry, separability and strict positive definiteness for covariance functions: a critical review. *Spat Stat* 29:89–108
- De Iaco S, Posa D, Cappello C, Maggio S (2020) On some characteristics of Gaussian covariance functions. *Int Stat Rev* 89(1):36–53
- Deutsch CV, Journel AG (1998) GSLib: geostatistical software library and user's guide, Applied Geostatistics Series, 2nd edn. Oxford University Press, New York
- Journel AG, Huijbregts CJ (1981) Mining geostatistics. Academic Press, London, 610 p
- Krishnan S, Journel AG (2003) Spatial connectivity: from variograms to multiple-point measures. *Math Geol* 35(8):915–925
- Matern B (1980) Spatial variation, 2, lecture notes in statistics. Springer, New York
- Matheron G (1971) The theory of regionalized variables and its applications. Les Cahiers du Centre de Morphologie Mathématique in Fontainebleau, Fontainebleau, 211 p
- Matheron G (1973) The intrinsic random functions and their applications. *Adv Appl Probab* 5:439–468
- Myers DE (1989) To be or not to be... Stationary? That is the question. *Math Geol* 21:347–362
- Pebesma EJ, Wesseling CG (1998) Gstat: a program for geostatistical modelling, prediction and simulation. *Comput Geosci* 24(1):17–31. [https://doi.org/10.1016/S0098-3004\(97\)00082-4](https://doi.org/10.1016/S0098-3004(97)00082-4)
- Posa D (2020) Parametric families for complex valued covariance functions: some results, an overview and critical aspects. *Spat Stat* 39(2):1–20. <https://doi.org/10.1016/j.spasta.2020.100473>
- Posa D (2021) Models for the difference of continuous covariance functions. *Stoch Env Res Risk A* 35:1369–1386
- Remy N, Boucher A, Wu J (2009) Applied geostatistics with SGeMS: a user's guide. Cambridge University Press, New York
- Renard D, Desassis N, Beucher H, Ors F, Laporte F (2014) RGeostats: the geostatistical package. Mines Paris Tech
- Yaglom AM (1987) Correlation theory of stationary and related random functions, vol I. Springer, Berlin, 526 p
- Zimmerman DL (1993) Another look at anisotropy in geostatistics. *Math Geol* 25(4):453–470

Spatial Data

Sabrina Maggio and Claudia Cappello
Dipartimento di Scienze dell'Economia, Università del Salento, Lecce, Italy

Synonyms

Geodatabase; Geographical data; Georeferenced data; Territorial data

Definition

Data which present a spatial component are called spatial data (Fischer and Wang 2011). In a nutshell, they are sample data of a random field referred to a spatial domain.

Overview

Spatial data are a collection of observations measured in different locations on a spatial domain. They are interpreted as a finite realization of a random field, whose distribution law (often unknown) provides a likelihood measure of the spatial evolution of the phenomena under study.

Spatial data have two characteristics, that is:

- They are non-repetitive, since only one observation is available in every single location.
- They are spatially correlated, which means that the phenomenon under study varies in space, but the variations are correlated over some distances.

For this last reason, the usual assumption of independence, made in classical statistical inference, is not reasonable in

Spatial Data, Table 1 Classification and characteristics of spatial data, together with the R packages available for their analysis

Types of data ^a	Variable type	Spatial index	Examples	R packages ^b
Geostatistical	Random variables discrete or continuous	Variable defined everywhere in the domain	Soil pH Air temperature Soil clay content	gstat geoR geospt RandomFields RGeostats (MINES 2020)
Lattice	Random variables discrete or continuous	Point/area objects fixed in the domain	Disease rate Crime rate Land use Total fertility rate	spdep spgwr spatialreg DCluster
Point patterns	Random variables discrete or continuous (if measured)	Randomly located points	Location of trees, of restaurants, of bicycle theft crimes	spatstat splancs spatial SpatialKernel

^aClassification suggested by Cressie (1993)

^bAn exhaustive list is available on the CRAN Task View: Analysis of Spatial Data (Bivand 2021)

spatial statistics. Indeed spatial data usually show dependencies along any direction in the domain, and their intensity is generally closely connected with the distance between locations and weakens as the distance arises (Cressie 1993). This feature of the spatial data recalls Tobler's first law of geography: "everything is related to everything else, but near things are more related than distant things" (Tobler 1970).

Spatial data are used in various scientific fields, such as agriculture, geology, environmental sciences, hydrology, epidemiology, and economics (Müller 2007). In the geographical information system, the term spatial data is used to refer to features (points, lines, and polygons) linked to a geographic location on the Earth; in this sense, the features are georeferenced. Note that each specific application often requires an appropriate data resolution, since some features of interest are scale-dependent (Flury et al. 2021).

Types of Spatial Data

Over the past 40 years, many studies have been carried out on spatial data analysis considered from different points of view (Ripley 1981; Isaaks and Srivastava 1989; Goovaerts 1997; Chilès and Delfiner 2012; Schabenberger and Gotway 2005 and others).

Given a spatial random field in a d -dimensional space $\{Z(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$, where $Z(\mathbf{s})$ denotes a spatial random variable and $\mathbf{s}$ a location of the spatial domain D , the spatial data can be classified with respect to the nature of the domain D , as follows:

- Geostatistical data
- Lattice data
- Point patterns

A summary of these spatial data classification is provided in Table 1. For further details on the spatial data classification, the readers may refer to Cressie (1993).

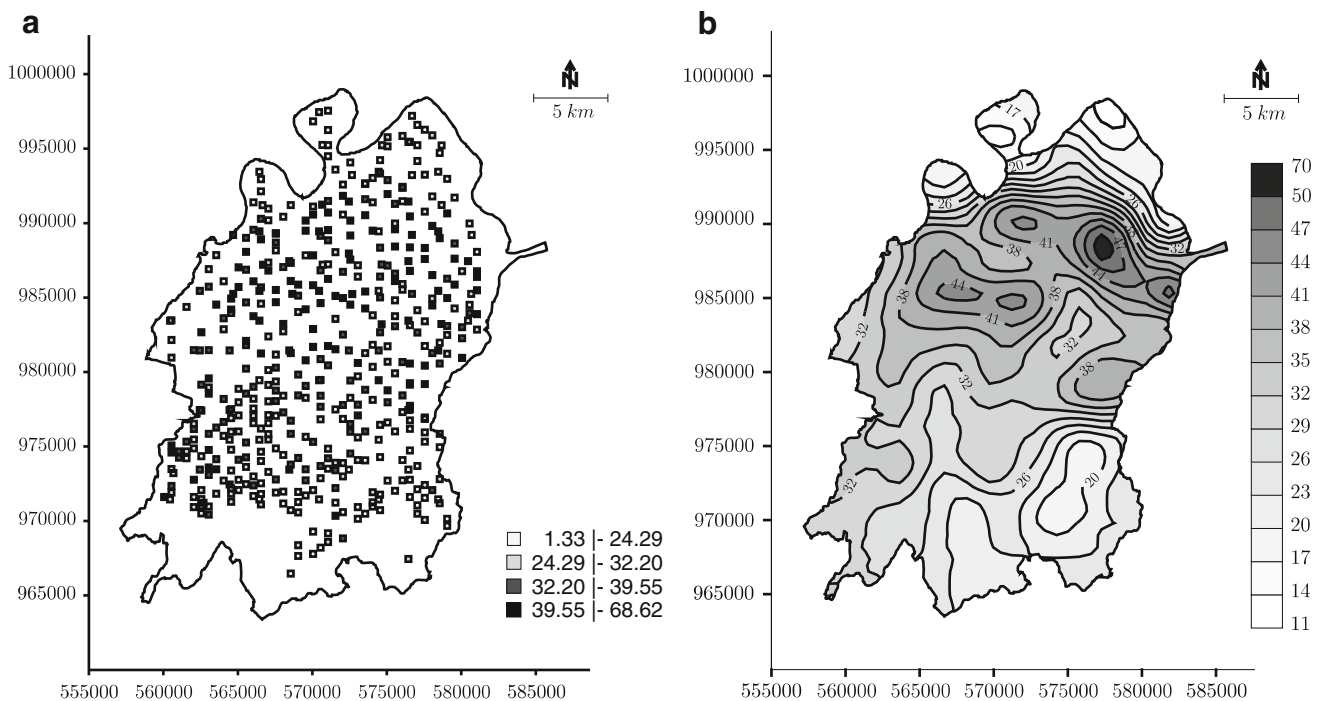
Geostatistical Data

In the case of geostatistical data, the sample data are considered as a finite realization of the random field $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$. The values that can be observed at the i -th spatial point is interpreted as a possible realization of a random variable, which is denoted as $Z(\mathbf{s}_i)$, where $\mathbf{s}_i$ represents the spatial location (Cressie 1993). Moreover, the peculiar feature of geostatistical data is that the domain $D \subset \mathbb{R}^d$ is continuous and fixed. If $d = 2$, the spatial location $\mathbf{s}$ is characterized by the Cartesian coordinates (x, y) . In this field of spatial data, the locations are treated as explanatory variables and the values recorded at these locations as response variables.

The spatial domain is continuous since the random variable $Z(\mathbf{s})$ can be observed at any point of D ; hence between two spatial locations $\mathbf{s}_i$ and $\mathbf{s}_j$, an infinite/non-countable number of other spatial locations can be fixed. It is worth noting that the adjective continuous refers to the spatial domain and not to the type of the random variable measured on it, which could be continuous or discrete.

Furthermore, the spatial domain is fixed in the sense that the spatial points $\mathbf{s}_i$, $i = 1, \dots, n$, are non-random (Schabenberger and Gotway 2005; Cressie 1993).

It is important to highlight that since the spatial domain is continuous, it cannot be sampled totally; consequently, the sample data is sampled at a finite number of locations, chosen at random or by a deterministic method. The sampling design takes into account sampling costs, benefits, and experimental constraints (i.e., locations of measuring stations).



Spatial Data, Fig. 1 (a) Gray scale map and (b) contour map of the soil clay percentage in an Italian district (Source data: soil map of Italy, year 2000, <https://esdac.jrc.ec.europa.eu/>)

Since there are only few measurements of the phenomena under study over the spatial domain, in geostatistics the main objective is the estimation of the random variable $Z(\mathbf{s})$ at unsampled spatial location, starting from observations available at a finite set of locations. The kriging is the most and widely used method for predicting the random variable in an unsampled location, by means of the observed data. This optimal prediction method requires the knowledge of a variogram/covariogram model, for this reason a relevant effort is devoted to estimate the spatial correlation function (through the variogram or the covariogram) as well as to fit a model to it.

In the following, an example regarding geostatistical data is proposed.

Example 1 Soil clay content is a key parameter in soil science, since it influences magnitudes and rates of several chemical, physical, and hydrological phenomena in soils. Given a spatial sampling, the observations of this feature taken at the sampled spatial locations can be considered as geostatistical data. Indeed, the soil clay content can be evaluated at any location of a given domain D , but only few measurements at some spatial locations are available. One of the main goals might be to define a continuous surface across the domain in order to estimate the spatial variability of the soil clay content over the area under study.

In Fig. 1, a location map of 485 sample points regarding the soil clay percentage in an Italian district and the corresponding kriging estimates, over the entire domain, are shown (Posa and De Iaco 2009). The estimates are obtained by using the ordinary kriging method and a variogram model.

Lattice Data

Lattice data are spatial data where the domain D is a countable collection of spatial regions at which the random variable is observed. The spatial regions are such that none of them can intersect each other, and the regions are characterized by a neighborhood structure. For example, in a 2D space, the spatial domain can be expressed as the set of the areas A_i , $i = 1, \dots, n$, which cover the territory under study D :

$$D = \{A_1, A_2, \dots, A_n\} \quad A_i \subset D \subset \mathbb{R}^2 \quad A_1 \cup A_2 \cup \dots \cup A_n = D$$

with $A_i \cap A_j = \emptyset$, $i, j \in \mathbb{N}_+$, $i \neq j$.

Hence, in this context, the spatial domain is fixed and discrete. In particular, the spatial domain is

- Fixed, in the sense that the spatial regions A_i , $i = 1, \dots, n$, in D are not stochastic (Schabenberger and Gotway 2005).
- Discrete, since the number of spatial regions could be infinite, although countable.

The spatial data could refer to a regular lattice (i.e., the cells/regions show the same form and size) or irregular lattice (i.e., the spatial domain is divided into cells/regions according to the natural or policy barriers). Moreover, the data do not correspond to a specific point in space, but to the entire region. Thus, it is usually assigned to each area a representative location in space, which is usually the centroid of the region, and then recorded for each centroid the corresponding spatial coordinates (i.e., longitude and latitude) in a 2D domain. Therefore, the spatial domain can be also formalized as follows:

$$D = \{(i; x_i, y_i) : i = 1, \dots, n\}$$

where i denotes the i -th region under study, x_i and y_i are the longitude and latitude of the i -th centroid, respectively.

Similarly to the geostatistical context, the values that can be observed at the i -th region are interpreted as a possible realization of a random variable, which is denoted as $Z(\mathbf{s}_i)$, where $\mathbf{s}_i$ represents the centroid of the corresponding region and refers to the representative location (Cressie 1993). In other contributions (see, for example, Schabenberger and Gotway 2005), the notation $Z(A_i)$ is used, in order to underline the areal nature of lattice data, where A_i is the i -th sampled areal unit.

It is important to point out that a specific feature of the lattice data is that they are exhaustive over the spatial domain. Moreover, they usually concern aggregated data over the sample region. However, the aggregation of the data causes sometimes some problems; in particular if the level of aggregation is chosen arbitrarily, the *modifiable areal unit problem* can arise (Openshaw and Taylor 1979), since the size of the areas on which aggregation is applied influences the correlation between variables. This problem occurs frequently in ecology, where most of ecological variables depend on other covariates, acting at different spatial scales (Saveliev et al. 2007).

Note that, in the analysis of the lattice data, the estimation of the random variable at unsampled points is not relevant, since the phenomenon under study is known over the entire domain, but the main goal is to model the spatial process using the spatial regression method. In the model, the variable of interest is predicted by using other variables measured on the same areas as well as by including the spatial dependence of neighboring regions, since if regions are nearby, the random variables collected at these areas could be spatially correlated.

In order to include the spatial component in the regression model, the covariance structure must be identified. This last requires to define previously a neighborhood structure by means of the $(n \times n)$ spatial weights matrix $\mathbf{W}$, whose elements w_{ij} are defined as follows:

$$w_{ij} = \begin{cases} 1 & \text{if } A_i \text{ and } A_j \text{ share a common border,} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Note that symmetry of the weights is not a requirement (Schabenberger and Gotway 2005). Regions which are neighboring and with similar values identify a spatial pattern, which is representative of a cluster, and the selection of correlated variables with the one of interest helps in the detection of explicative factors for the analyzed phenomenon.

For lattice data modeling, the wider applied approaches are based on the linear regression models. However, if the residuals highlight the presence of spatial correlation, alternative models which take into account the spatial dependence could be used. Among these, it is worth mentioning the conditional spatial autoregression (CAR), simultaneous spatial autoregression (SAR), and moving average (MA) models.

In the following, an example regarding lattice data is proposed.

Example 2 The total fertility rate (TFR) is the average number of children that a woman would potentially have in her child-bearing years (i.e., for ages 15–49).

In Fig. 2, the map of the centroids of the regions and the color map of this demographic index, available for each region in France, are provided (De Iaco et al. 2015). This is a typical example of lattice data, where the domain D is a set of 21 counties; thus it is fixed and discrete. Moreover, the TFR is an aggregate statistic over each region.

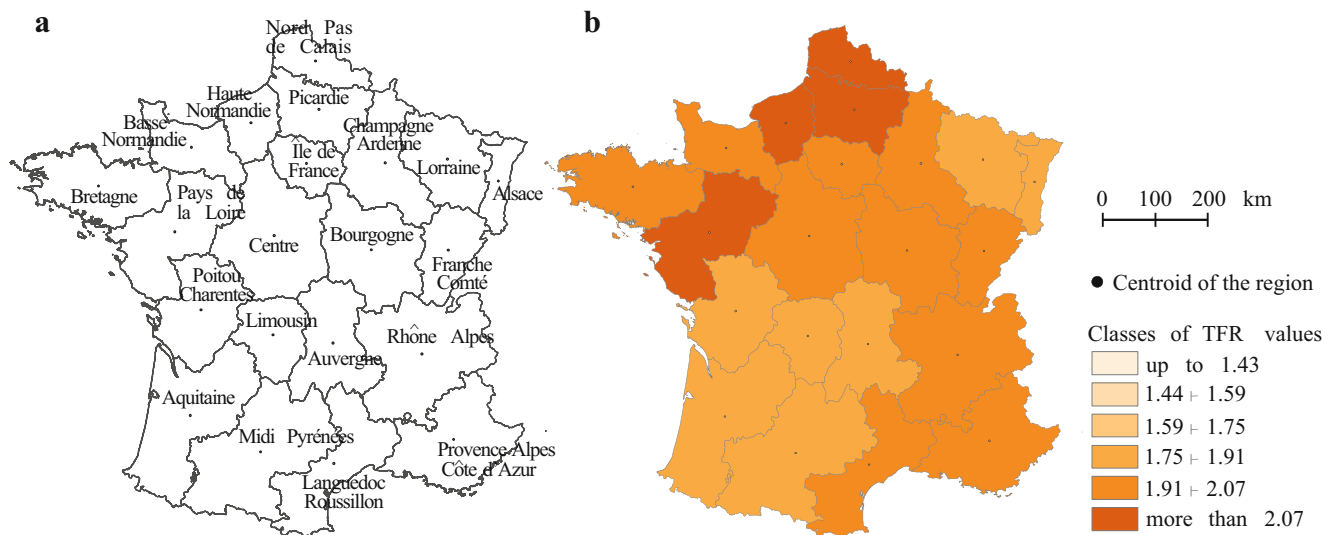
Point Patterns

Geostatistical and lattice data are measured over a fixed, non-stochastic domain, where the term “fixed” is used to highlight that the domain D does not change if different realizations of the spatial process are considered. Moreover, for these types of data, the analyst is interested in the measurement of the attribute Z over the spatial domain D . On the other hand, if the analyst intends to focus only on the location at which an event occurs, such a dataset is known as spatial point pattern.

Diggle (2003) defined the point process as a “stochastic mechanism which generates a countable set of events,” where the events are the points of the spatial pattern in a bounded region. The point patterns are said to be *sampled* or *mapped*, if the events are partially observed or if all the events are recorded, respectively.

For a spatial point process, the spatial domain D is a collection of points on the hyperplane

$$D = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n\}$$



Spatial Data, Fig. 2 (a) Location map of the regional centroids and (b) color map of the yearly TFR registered in 2011 for the 21 regions of mainland France (Source data: Eurostat, 2013, <http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics>)

In general, no random variable $Z(\mathbf{s})$ is observed at these points. Otherwise, if an attribute is measured in each point, the realization of a point process $\{Z(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$ represents a pattern of points over the domain D . In order to distinguish these two cases, in the literature the terms *unmarked* and *marked* point pattern are often used. In particular, the former is given to the locations where the event occurs, while the latter is used when an attribute $Z(\cdot)$ is also observed at these points. For instance, the locations at which trees emerge in a forest, the spatial distribution of schools in a city, or the locations of an earthquake are examples of unmarked point patterns. Nevertheless, the set of locations regarding the height of trees in a forest, the size of schools in a city, or the magnitude of earthquakes are examples of marked point patterns.

It is important to highlight that, in contrast with geostatistical and lattice data, which can be measured at any point of the spatial domain D , a point process takes nonzero values only in specific points that are the ones in which the underlying process occurs (Hristopulos 2020). Moreover, in spatial point pattern statistics, the spatial locations as well as the values measured in each point are treated as response variables.

Given a random set D , for each region $A_i \in D$, $N(A_i)$ denotes the number of points in the region A_i and $\lambda(A_i)$ is the average number of events in the same area. The point pattern is named *completely random process* or, alternatively, *Poisson point process* if the following properties are satisfied:

- The intensity function $\lambda(\cdot)$, which is the average number of points in each area, is constant.
- The numbers of events in two areas A_i and A_j , such as $A_i \cap A_j = \emptyset$, $i, j \in \mathbb{N}_+$, $i \neq j$, in the domain D are

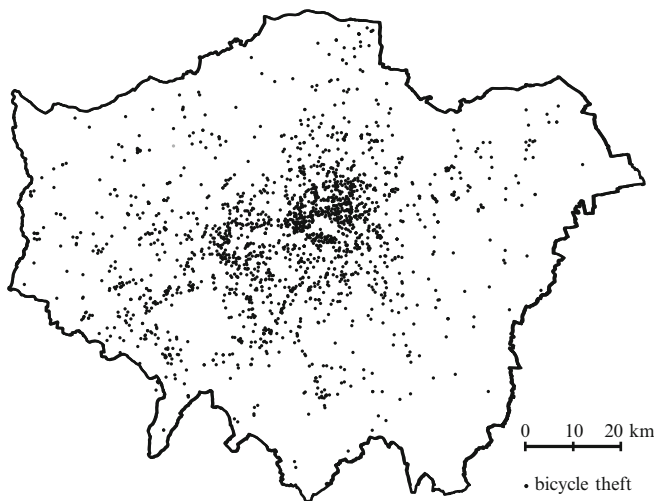
independent; hence the probability that an event occurs is equal in each spatial location.

Note that this model is widely used in spatial point pattern analysis, to compare the locations of the events with this reference distribution. However, the Poisson point process does not always adequately describe phenomena with spatial clustering of points and spatial regularity. In the literature, a variety of other models were proposed, among these the Cox process, the inhibition process, or the Markov point process (Cressie 1993; Schabenberger and Gotway 2005).

The objectives of point pattern analysis vary according to the scientist's purposes. In general, it is relevant to focus on the spatial distribution of the observed events and make inference about the underlying process that generates them. In particular, it could be interesting to analyze the distribution of the points in space, in terms of intensity of the point pattern, as well as the existence of possible interactions between events, which may reveal the tendency of events to appear clustered, independently, or regularly spaced (Bivand et al. 2013). For example in ecology, a target could be determining the spatial distribution of a tree species over a study area or assessing whether two or more species are equally distributed. Furthermore, in spatial epidemiology, it could be challenging to check whether or not the cases of a certain disease are clustered. This approach can be also used in human sciences, to identify, for example, areas where crimes occur, the activity spaces of individuals, the catchment areas of services, such as banks and restaurants.

In the following, an example regarding point patterns is proposed.

Example 3 In Fig. 3, a map of bicycle theft crimes occurring in London (September 2017) is shown. This represents an



Spatial Data, Fig. 3 Location map of bicycle thefts in the city of London in September 2017 (Source data: UK Police Department 2017, <https://data.police.uk/data/archive/>)

example of unmarked spatial point pattern, which consists of 2218 locations of bicycle thefts recorded within the city of London.

Summary

The spatial data include information on the location and the measured attributes. They have two fundamental characteristics: (a) dependency between specific locations along any direction, where the intensity becomes weaker as the distance arises, and (b) non-repetitive, since only one observation is available in every single location.

As reported in Cressie (1993), the spatial data types have been classified with respect to the characteristics of the domain D , as geostatistical data, lattice data, and point patterns.

Cross-References

- [Geostatistics](#)
- [Multiple Point Statistics](#)
- [Random Variable](#)
- [Spatial Analysis](#)
- [Spatial Statistics](#)

Bibliography

Bivand RS (2021) CRAN task view: analysis of spatial Data. Version 2021-03-01, URL: <https://CRAN.R-project.org/view=Spatial>

- Bivand RS, Pebesma E, Gomez-Rubio V (2013) Applied spatial data analysis with R, 2nd edn. Springer, New York, 405 pp
- Chilès JP, Delfiner P (2012) Geostatistics: modeling spatial uncertainty, 2nd edn. Wiley, Hoboken, p 734
- Cressie NAC (1993) Statistics for spatial data, Revised edn. Wiley, New York, p 416
- De Iaco S, Palma M, Posa D (2015) Spatio-temporal geostatistical modeling for French fertility predictions. *Spat Stat* 14:546–562
- Diggle PJ (2003) Statistical analysis of spatial point patterns, 2nd edn. Arnold, London, p 267
- Fischer MM, Wang J (2011) Spatial data analysis. Models, methods and techniques. Springer, Heidelberg/Dordrecht/London/New York, p 91
- Flury R, Gerber F, Schmid B, Furrer R (2021) Identification of dominant features in spatial data. *Spat Stat* 41:25. <https://doi.org/10.1016/j.spasta.2020.100483>
- Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York, p 483
- Hristopulos D (2020) Random fields for spatial data modeling: a primer for scientists and engineers. Springer, Netherlands, p 867
- Isaaks EH, Srivastava RM (1989) An introduction to applied geostatistics. Oxford University Press, New York, p 561
- MINES ParisTech/ARMINES (2020) RGeostats: the geostatistical R Package. Free download from <http://cg.enscm.fr/rgeostats>
- Müller WG (2007) Collecting spatial data. Optimum design of experiments for random fields. Springer, Berlin/Heidelberg, 250 pp
- Openshaw S, Taylor PJ (1979) A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In: Wrigley N (ed) Statistical applications in the spatial sciences. Pion, London, pp 127–144
- Posa D, De Iaco S (2009) Geostatistica: Teoria e Applicazioni. Giappichelli, Torino, p 264
- Ripley BD (1981) Spatial statistics. John Wiley & Sons, New York, p 252
- Saveliev AA, Mukharamova SS, Zuur AF (2007) Analysis and modelling of lattice data. In: Analysing ecological data. Statistics for biology and health. Springer, New York, pp 321–339
- Schabenberger O, Gotway CA (2005) Statistical methods for spatial data analysis: texts in statistical science, 1st edn. CRC Press, Boca Raton, p 512
- Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. *J Econ Geogr* 46:234–240

Spatial Data Infrastructure and Generalization

Jagadish Boodala, Onkar Dikshit and
Nagarajan Balasubramanian
Department of Civil Engineering, Indian Institute of
Technology Kanpur, Kanpur, UP, India

Definition

The Spatial Data Infrastructure (SDI) is a framework that facilitates the reuse and integration of spatial data and services available from different sources. SDI framework is a collection of policies, spatial data, technologies, humans, and institutional arrangements that ease the sharing and reuse of spatial data. The main component of SDI is spatial data. The

generalization plays an essential role in situations where the spatial data of different levels of detail or scales are to be combined and made homogeneous. The transformation of spatial data from more detailed to the abstract level of detail is called generalization.

Introduction

Traditional maps were the outcome of spatial observations for many years to get the spatial context (Tóth et al. 2012). With technological advancements, conventional paper maps were replaced by digital spatial data. Initially, the public sector was responsible for capturing and producing a significant portion of the spatial data, such as National Mapping and Cadastral Agencies (Toomanian 2012). The private sector has shown interest in the spatial data when its applications have increased in areas like location-based services, agriculture, intelligent navigation, etc. The licensing issues of authoritative spatial data from public and private sectors lead to the

crowdsourced data called volunteered geographic information (Toomanian 2012).

The availability of spatial data from the public, private, and crowdsourced have resulted in the data explosion. Furthermore, the number of spatial data processing tools has also increased. The cost of data capturing is high; hence there is a necessity to reuse the existing spatial data. There is a need for a framework, called SDI, to make better sense and use of the current spatial data. One such legal framework to establish European SDI is INSPIRE. INSPIRE stands for the Infrastructure for Spatial Information in the European Community. INSPIRE's main objective is to assist in policy-making related to environmental issues by improving the accessibility and interoperability of spatial data among European Union member states (Duchêne et al. 2014; INSPIRE 2007a).

As the spatial data are coming from various sources, there is a possibility that the same real-world phenomena are modeled and represented by different viewpoints and also at different levels of detail or scales. The multiple representations are quite common, which can be derived and handled by a generalization process (Tóth 2007). The following sections



Spatial Data Infrastructure and Generalization, Fig. 1 Features of different themes at the same level of detail. (Contains OS data © Crown copyright and database right (2017))

highlight the requirements set out by the INSPIRE Directive (INSPIRE 2007a) to achieve data interoperability and explain how generalization can help in achieving it.

INSPIRE: Multiple Representations and Data Consistency

The possibility of combining spatial data sets, without repetitive human intervention, into coherent data sets whose value is enhanced is called data interoperability (INSPIRE 2007a). Recital 6 of the INSPIRE Directive provides five common principles of INSPIRE (INSPIRE 2007a, b). The first three principles form the basis for defining data interoperability (INSPIRE 2014).

Article 8(3) and Article 10(2) of the INSPIRE Directive provide the requirements to handle multiple representations and data consistency (INSPIRE 2007a). Hence, the elements “(Q) Consistency between data,” and “(R) Multiple representations” are included as the data interoperability components (INSPIRE 2014).

The component “(Q) Consistency between data” provides the rules to maintain consistency between the data corresponding to the same location or between the same data collected at different levels of detail or scales (INSPIRE 2007a). However, the component “(R) Multiple representations” explains how generalization can be used to derive multiple representations (INSPIRE 2014). Another data interoperability component that specifies the target scale range is the element “(S) Data capturing.” This component defines the data capturing rules or selection criteria according to the target scale (INSPIRE 2014).

The requirements of INSPIRE discussed above helps to pinpoint the situations where generalization finds its worth.

Generalization in the Context of INSPIRE

The first principle of INSPIRE (INSPIRE 2007b), “data should be collected only once,” motivates to apply an automatic generalization to derive spatial data of abstract levels of



Spatial Data Infrastructure and Generalization, Fig. 2 Building features at two different levels of detail. (Contains OS data © Crown copyright and database right (2017))

detail. These different representations of the same real-world phenomena are called multiple representations.

The generalization is a process of reducing the level of detail or scale of the spatial data by applying a broad set of operators called generalization operators. The high-level classification of generalization operators is either based on automation (Foerster et al. 2007) or the cartographer's (Roth et al. 2011) point of view. However, the essential list of generalization operators includes selection, simplification, classification, collapse, merge, typification, displacement, enhancement, and elimination. The data capturing rules, for example, minimum area or length, defined in INSPIRE, are useful while performing the selection operation.

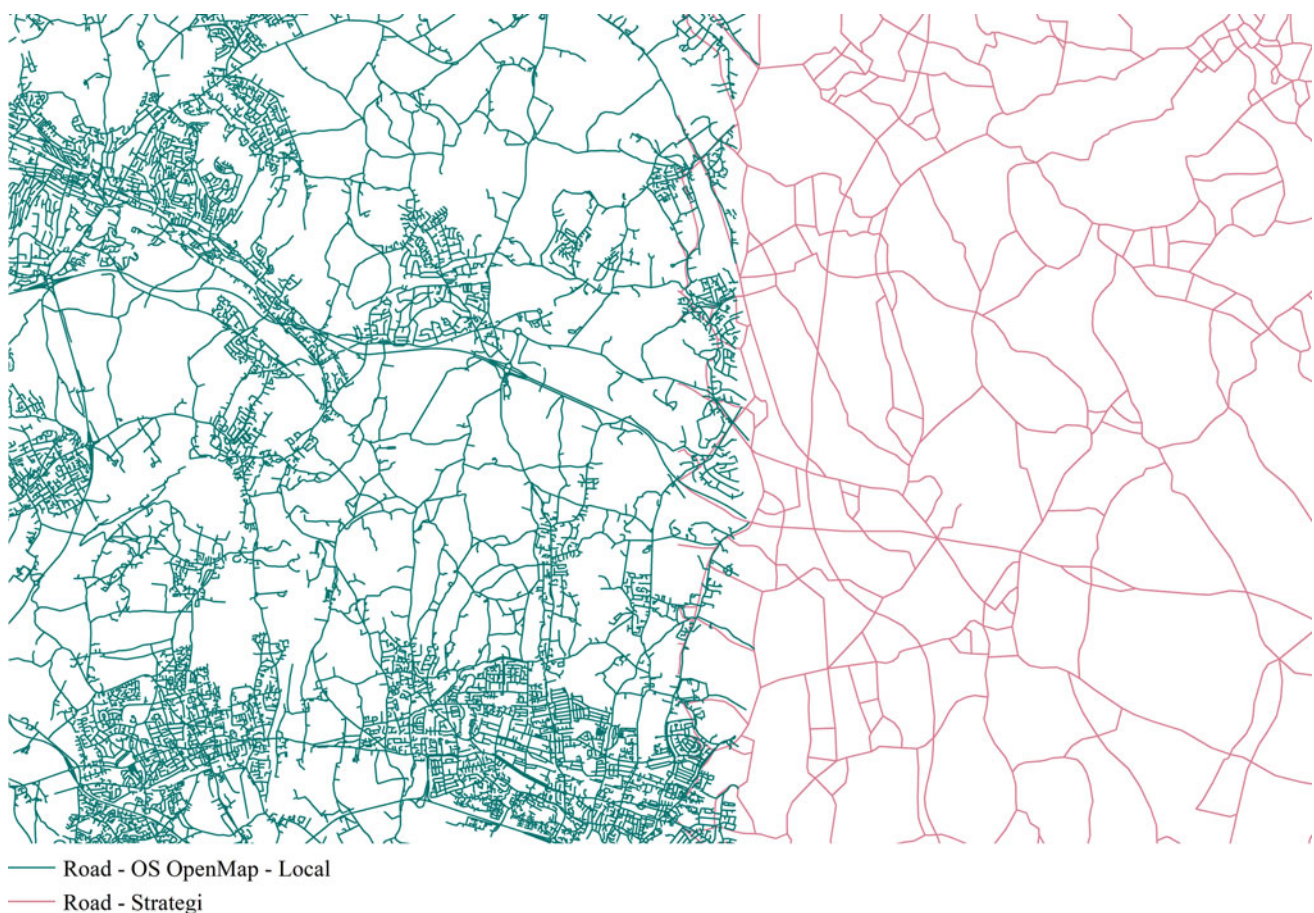
The generalization is a complex process. It is challenging to achieve unique results because of the subjectivity involved in the process itself (Stoter 2005). The issues with generalization are its subjectivity, complexity, and immaturity. Hence, generalization service is not part of the INSPIRE data transformation services; however, multiple representations are allowed in INSPIRE (INSPIRE 2008).

The modeling of multiple representations in INSPIRE paves the way for easy integration of spatial data sets from

different sources. While dealing with multiple representations and data sets from different sources, it is necessary to maintain consistency between the datasets. INSPIRE identifies different types of consistency checks: within a data set, between features of different themes at the same level of detail, and between the same feature at different levels of detail (INSPIRE 2014).

Figure 1 shows the scenario where the features of different themes are represented together in a product of scale 1:10000. The generalization process should consider the consistency rules defined between all the features. These rules are developed based on logical, functional, and geometrical relationships between features. For example, Building features must not overlap between themselves and with the Woodland features. The consistency rules are expressed as topological constraints to avoid inconsistencies in the data set and guide the generalization process. The "OS OpenMap – Local" product of Ordnance Survey, UK, is used for the illustration in Fig. 1.

The consistency between multiple representations of the same feature at different levels of detail should be maintained. Fig. 2 explains the importance of this consistency



Spatial Data Infrastructure and Generalization, Fig. 3 Road features at two different levels of detail. (Contains OS data © Crown copyright and database right (2017 & 2015))

requirement. It shows the two representations of the Building features in products of 1:10000 (OS OpenMap – Local) and 1:25000 (OS VectorMap District) scales. The advantages of automatic generalization are twofold. One is in deriving the abstract representations, and the other is in deriving the consistent data sets. Maintaining links between multiple representations from existing data sets is quite challenging than maintaining it between the representations derived via automatic generalization. The established links help in maintaining data consistency (INSPIRE 2008).

Figure 3 illustrates the situation frequently encountered in SDI. It considers the Road features from products of 1:10000 (OS OpenMap – Local) and 1:250000 (Strategi) scales. In this case, the spatial data user has two different data sets of different levels of detail or scales separately covering the region of his/her interest. It is clear from Fig. 3 that the spatial data sets covering the region of the user's interest are heterogeneous. In such cases, generalization comes into the picture to homogenize the level of the detail or scale of the combined spatial data set.

Summary

An increase in spatial data applications in various fields attracted investments in spatial data capturing from both the public and private sectors. Nevertheless, the data capturing cost remains high. Some policies or decisions, for example, related to the environment or disaster management, have to be taken at the national or international level. In such cases, spatial data from different sources are required to support the policy or decision-making. The spatial data from different sources cannot be integrated and used instantaneously because of different standards followed in capturing it. Also, the accessibility of the spatial data sets is not easy and quick. All these aspects lead to the SDI, and the INSPIRE Directive aims to establish a European Union SDI. The spatial data is an essential component of INSPIRE, and for that matter, of any SDI. Hence, INSPIRE defines a total of 20 elements as the data interoperability components. Among those 20 elements, generalization finds its application in only 3 components to achieve data interoperability. They are “(Q) Consistency between data,” “(R) Multiple representations,” and “(S) Data capturing.” In the future, generalization can be used in INSPIRE view services to provide high-quality cartographic data in an INSPIRE geportal.

Cross-References

- [Data Visualization](#)
- [Geographical Information Science](#)
- [Metadata](#)
- [Spatial Data](#)

Bibliography

- Duchêne C, Baella B, Brewer CA, Burghardt D, Battenfield BP, Gaffuri J, Käuferle D, Lecordix F, Maugeais E, Nijhuis R, Pla M, Post M, Regnaud N, Stanislawski LV, Stoter J, Tóth K, Urbanke S, van Altena V, Wiedemann A (2014) Generalisation in practice within National Mapping Agencies. In: Burghardt D, Duchêne C, Mackaness W (eds) Abstracting geographic information in a data rich world. Springer International Publishing, Cham, pp 329–391
- Foerster T, Stoter J, Köbben B (2007) Towards a formal classification of generalization operators. In: Proceedings of the 23rd international cartographic conference – cartography for everyone and for you. International Cartographic Association, Moscow
- INSPIRE (2007a) Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). <https://eur-lex.europa.eu/eli/dir/2007/2/oj>. Accessed 29 Oct 2020
- INSPIRE (2007b) INSPIRE Principles. <https://inspire.ec.europa.eu/inspire-principles/9>. Accessed 29 Oct 2020
- INSPIRE (2008) D2.6: Methodology for the development of data specifications, v3.0. <https://inspire.ec.europa.eu/documents/methodology-development-data-specifications-baseline-version-d-26-version-30>. Accessed 29 Oct 2020
- INSPIRE (2014) D2.5: Generic conceptual model, Version 3.4. <https://inspire.ec.europa.eu/documents/inspire-generic-conceptual-model>. Accessed 29 Oct 2020
- Roth RE, Brewer CA, Stryker MS (2011) A typology of operators for maintaining legible map designs at multiple scales. *Cartograph Perspect* 68:29–64. <https://doi.org/10.14714/CP68.7>
- Stoter JE (2005) Generalisation within NMA's in the 21st century. In: Proceedings of the 22nd international cartographic conference: mapping approaches into a changing world. International Cartographic Association, A Coruña
- Toomanian A (2012) Methods to improve and evaluate spatial data infrastructures. Department of Physical Geography and Ecosystem Science. Lund University, Lund
- Tóth K (2007) Data consistency and multiple-representation in the European spatial data infrastructure. In: 10th ICA workshop on generalisation and multiple representation, Moscow
- Tóth K, Portele C, Illert A, Lutz M, Nunes de Lima M (2012) A conceptual model for developing interoperability specifications in spatial data infrastructures. EUR, Scientific and technical research series, vol 25280. Publications Office of the European Union, Luxembourg

Spatial Statistics

Noel Cressie and Matthew T. Moores
School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW, Australia

Definition

Spatial statistics is an area of study devoted to the statistical analysis of data that have a spatial label associated with them. Geographers often refer to the “location information” associated with the “attribute information,” whose study defines a research area called “spatial analysis.” Many of the ways to manipulate spatial data are driven by algorithms with no

uncertainty quantification associated with them. When a spatial analysis is statistical, that is, when it incorporates uncertainty quantification, it falls in the research area called spatial statistics. The primary feature of spatial statistical models is that nearby attribute values are more statistically dependent than distant attribute values; this is a paraphrasing of what is sometimes called the First Law of Geography (Tobler 1970).

Introduction

Spatial statistics provides a probabilistic framework for giving answers to those scientific questions where spatial-location information is present in the data, and that information is relevant to the questions being asked. The role of probability theory in (spatial) statistics is to model the uncertainty, both in the scientific theory behind the question, and in the (spatial) data coming from measurements of the (spatial) process that is a representation of the scientific theory.

In spatial statistics, uncertainty in the scientific theory is expressed probabilistically through a spatial stochastic process, which can be written most generally as:

$$\{Y(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}, \quad (1)$$

where $Y(\mathbf{s})$ is the random attribute value at location $\mathbf{s}$, and $\mathcal{D}$ is a subset of a d -dimensional space, for illustration Euclidean space $\mathbb{R}^d$, that indexes all possible spatial locations of interest. Contained within $\mathcal{D}$ is a (possibly random) set D that indexes those parts of $\mathcal{D}$ relevant to the scientific study. We shall see below that D can have different set properties, depending upon whether the spatial process is a geostatistical process, a lattice process, or a point process.

It is convenient to express the joint probability model defined by random $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$ and random D in the following shorthand, $[Y, D]$, which we refer to as the spatial process model. Now,

$$[Y, D] = [Y|D][D], \quad (2)$$

where for generic random quantities A and B , their joint probability measure is denoted by $[A, B]$; the conditional probability measure of A given B is denoted by $[A|B]$; and the marginal probability measure of B is denoted by $[B]$. In this review of spatial statistics, expression (2) formalizes the general definition of a spatial statistical model given in Cressie (1993, Section 1.1).

The model (2) covers the three principal spatial statistical areas according to three different assumptions about $[D]$, which leads to three different types of spatial stochastic process, $[Y|D]$; these are described further in the next section “Spatial Process Models”. Spatial statistics has, in the past, classified its methodology according to the types of spatial

data, denoted here as $\mathbf{Z}$ (e.g., Ripley 1981; Upton and Fingleton 1985; and Cressie 1993), rather than the types of spatial processes Y that underly the spatial data.

In this review, we classify our spatial-statistical modeling choices according to the *process model* (2). Then the *data model*, namely, the distribution of the data $\mathbf{Z}$ given both Y and D in (2), is the straightforward conditional-probability measure,

$$[\mathbf{Z}|Y, D]. \quad (3)$$

For example, the spatial data $\mathbf{Z}$ could be the vector $(Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))'$, of imperfect measurements of Y taken at given spatial locations $\{\mathbf{s}_1, \dots, \mathbf{s}_n\} \subset D$, where the data are assumed to be *conditionally independent*. That is, the data model is

$$[\mathbf{Z}|Y, D] = \prod_{i=1}^n [Z(\mathbf{s}_i)|Y, D]. \quad (4)$$

Notice that while (4) is based on conditional independence, the *marginal* distribution, $[\mathbf{Z}|D]$, does *not* exhibit independence: The spatial-statistical dependence in $\mathbf{Z}$, articulated in the First Law of Geography that was discussed in section “Definition”, is inherited from $[Y|D]$ and (4) as follows:

$$[\mathbf{Z}|D] = \int [\mathbf{Z}|Y, D][Y|D]dY.$$

Another example is where the randomness is in D but not in Y . If D is a spatial point process (a special case of a random set), then the data $\mathbf{Z} = \{N, \mathbf{s}_1, \dots, \mathbf{s}_N\}$, where N is the random number of points in the now-bounded region $\mathcal{D}$, and $D = \{\mathbf{s}_1, \dots, \mathbf{s}_N\}$ are the random locations of the points. If there are measurements (sometimes called “marks”) $\{Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_N)\}$ associated with the random points in D , these should be included within $\mathbf{Z}$. That is,

$$\mathbf{Z} = \{N, (\mathbf{s}_1, Z(\mathbf{s}_1)), \dots, (\mathbf{s}_N, Z(\mathbf{s}_N))\}. \quad (5)$$

This description of spatial statistics given by (2) and (3) captures the (known) uncertainty in the scientific problem being addressed, namely, *scientific uncertainty* through the spatial process model (2) and *measurement uncertainty* through the data model (3). Together, (2) and (3) define a *hierarchical statistical model*, here for spatial data, although this hierarchical formulation through the conditional probability distributions, $[\mathbf{Z}|Y, D]$, $[Y|D]$, and $[D]$ for general Y and D , is appropriate throughout all of applied statistics.

It is implicit in (2) and (3) that any parameters θ associated with the process model and the data model are known. We now discuss how to handle parameter uncertainty in the hierarchical statistical model. A Bayesian would put a

probability distribution on θ : Let $[\theta]$ denote the parameter model (or prior) that captures parameter uncertainty. Then, using obvious notation, all the uncertainty in the problem is expressed through the joint probability measure,

$$[\mathbf{Z}, Y, D, \theta] = [\mathbf{Z}, Y, D | \theta][\theta] \quad (6)$$

$$= [\mathbf{Z} | Y, D, \theta][Y | D, \theta][D | \theta][\theta]. \quad (7)$$

A *Bayesian hierarchical model* uses the decomposition (7), but there is also an *empirical hierarchical model* that substitutes a point estimate $\hat{\theta}$ of θ into the first factor on the right-hand side of (6), resulting in its being written as,

$$[\mathbf{Z}, Y, D | \hat{\theta}] = [\mathbf{Z} | Y, D, \hat{\theta}][Y | D, \hat{\theta}][D | \hat{\theta}]. \quad (8)$$

Finding efficient estimators of θ from the spatial data $\mathbf{Z}$ is an important problem in spatial statistics, but in this review we emphasize the problem of spatial prediction of Y . In what follows, we shall assume that the parameters are either known or have been estimated. Hence, for convenience, we can drop $\hat{\theta}$ in (8) and express the uncertainty in the problem through the joint probability measure,

$$[\mathbf{Z}, Y, D] = [\mathbf{Z} | Y, D][Y | D][D], \quad (9)$$

and Bayes' Rule can be used to infer the unknowns Y and D through the *predictive distribution*:

$$[Y, D | \mathbf{Z}] = \frac{[\mathbf{Z} | Y, D][Y | D][D]}{[\mathbf{Z}]}; \quad (10)$$

here $[\mathbf{Z}]$ is the normalization constant that ensures that the right-hand side of (10) integrates or sums to 1. If the spatial index set D is *fixed and known* then we can drop D from (10), and Bayes' Rule simplifies to:

$$[Y | \mathbf{Z}] = \frac{[\mathbf{Z} | Y][Y]}{[\mathbf{Z}]}, \quad (11)$$

which is the *predictive distribution* of Y (when D is fixed and known). It is this expression that is often used in spatial statistics for prediction. For example, the well-known *simple kriging predictor* can easily be identified as the predictive mean of (11) under Gaussian distributional assumptions for both (2) and (3) (Cressie and Wikle 2011, pp. 139–141).

Our review of spatial statistics starts with a presentation in the next section “[Spatial Process Models](#)”, of a number of commonly used spatial process models, which includes multivariate models. Following that, the section “[Spatial Discretization](#)”, turns attention to discretization of $\mathcal{D} \subset \mathbb{R}^d$, which is an extremely important consideration when actually computing the predictive distribution (10) or (11). The

extension of spatial process models to spatiotemporal process models is discussed in the section “[Spatiotemporal Processes](#)”. Finally, in the section “[Conclusion](#)”, we briefly discuss important recent research topics in spatial statistics, but due to a lack of space we are unable to present them in full. It will be interesting to see 10 years from now, how these topics will have evolved.

Spatial Process Models

In this section, we set out various ways that the probability distributions $[Y | D]$, $[Y]$, and $[D]$, given in Bayes' Rule (10), can be represented in the spatial context. These are not to be confused with $[\mathbf{Z} | D]$ and $[\mathbf{Z}]$, the probability distributions of the spatial data. In many parts of the spatial-statistics literature, this confusion is noticeable when researchers build models directly for $[\mathbf{Z} | D]$. Taking a hierarchical approach, we capture knowledge of the scientific process starting with the statistical models, $[Y | D]$ and $[D]$, and then we model the measurement errors and missing data through $[\mathbf{Z} | Y, D]$. Finally, Bayes' Rule (10) allows inference on the unknowns Y and D through the predictive distribution, $[Y, D | \mathbf{Z}]$.

We present three types of spatial process models, where their distinction is made according to the index set D of all spatial locations at which the process Y is defined. For a *geostatistical process*, $D = D^G$, which is a known set over which the locations vary continuously and whose area (or volume) is >0 . For a *lattice process*, $D = D^L$, which is a known set whose locations vary discretely and the number of locations is countable; note that the area of D^L is equal to zero. For a *point process*, $D = D^P$, which is a random set made up of random points in $\mathcal{D}$.

Geostatistical Processes

In this section, we assume that the spatial locations D are given by D^G , where D^G is known. Hence D can be dropped from any of the probability distributions in (10), resulting in (11). This allows us to concentrate on Y and, to feature the spatial index, we write Y equivalently as $\{Y(\mathbf{s}) : \mathbf{s} \in D^G\}$. A property of geostatistical processes is that D^G has positive area and hence is uncountable.

Traditionally, a geostatistical process has been specified up to second moments. Starting with the most general specification, we have

$$\mu_Y(\mathbf{s}) \equiv E(Y(\mathbf{s})); \quad \mathbf{s} \in D^G \quad (12)$$

$$C_Y(\mathbf{s}, \mathbf{u}) \equiv \text{cov}(Y(\mathbf{s}), Y(\mathbf{u})); \quad \mathbf{s}, \mathbf{u} \in D^G. \quad (13)$$

From (12) and (13), an optimal spatial linear predictor $\hat{Y}(\mathbf{s}_0)$ of $Y(\mathbf{s}_0)$ can be obtained that depends on spatial data

$\mathbf{Z} \equiv (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))'$. This is an n -dimensional vector indexed by the data's n known spatial locations, $D^{G*} \equiv \{\mathbf{s}_1, \dots, \mathbf{s}_n\} \subset D^G$. In practice, estimation of the parameters θ that specify completely (12) and (13) can be problematic due to the lack of replicated data, so Matheron (1963) made stationarity assumptions that together are now known as *intrinsic stationarity*. That is, for all $\mathbf{s}, \mathbf{u} \in D^G$, assume

$$E(Y(\mathbf{s})) = \mu_Y^0 \quad (14)$$

$$\text{var}(Y(\mathbf{s}) - Y(\mathbf{u})) = 2\gamma_Y^0(\mathbf{s} - \mathbf{u}), \quad (15)$$

where (15) is equal to $C_Y(\mathbf{s}, \mathbf{s}) + C_Y(\mathbf{u}, \mathbf{u}) - 2C_Y(\mathbf{s}, \mathbf{u})$. The quantity $2\gamma_Y^0(\cdot)$ is called the *variogram*, and $\gamma_Y^0(\cdot)$ is called the *semivariogram* (or occasionally the *semivariance*).

If the assumption in (15) were replaced by

$$\text{cov}(Y(\mathbf{s}), Y(\mathbf{u})) = C_Y^0(\mathbf{s} - \mathbf{u}), \text{ for all } \mathbf{s}, \mathbf{u} \in D^G, \quad (16)$$

then (16) and (14) together are known as *second-order stationarity*. Matheron chose (15) because he could derive optimal-spatial-linear-prediction (i.e., kriging) equations of $Y(\mathbf{s}_0)$ without having to know or estimate μ_Y^0 . Here, “optimal” is in reference to a spatial linear predictor $\hat{Y}(\mathbf{s}_0)$ that minimizes the mean-squared prediction error (MSPE),

$$E \left[\left(\hat{Y}(\mathbf{s}_0) - Y(\mathbf{s}_0) \right)^2 \right], \text{ for any } \mathbf{s}_0 \in D^G, \quad (17)$$

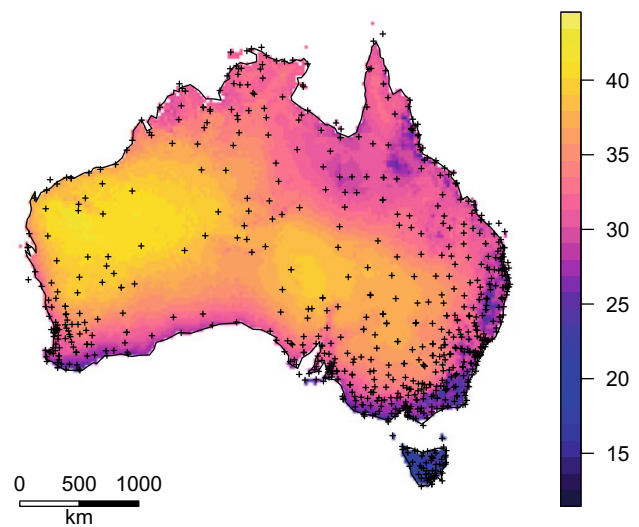
where $\hat{Y}(\mathbf{s}_0) \equiv \sum_{i=1}^n \lambda_i Z(\mathbf{s}_i)$. The minimization in (17) is with respect to the coefficients $\{\lambda_i : i = 1, \dots, n\}$ subject to the unbiasedness constraint, $E(\hat{Y}(\mathbf{s}_0)) = E(Y(\mathbf{s}_0))$, or equivalently subject to the constraint $\sum_{i=1}^n \lambda_i = 1$ on $\{\lambda_i\}$. With optimally chosen $\{\lambda_i\}$, $\hat{Y}(\mathbf{s}_0)$ is known as the *kriging predictor*. Matheron called this approach to spatial prediction *ordinary kriging*, although it is known in other fields as BLUP (Best Linear Unbiased Prediction); Cressie (1990) gave the history of kriging and showed that it could also be referred to descriptively as *spatial BLUP*.

The constant-mean assumption (14) can be generalized to $E(Y(\mathbf{s})) \equiv \mathbf{x}(\mathbf{s})'\boldsymbol{\beta}$, for $\mathbf{s} \in D^G$, which is a linear regression where the regression coefficients $\boldsymbol{\beta}$ are unknown and the covariate vector $\mathbf{x}(\mathbf{s})$ includes the entry 1. Under this assumption on $E(Y(\mathbf{s}))$, ordinary kriging is generalized to *universal kriging*, also notated as $\hat{Y}(\mathbf{s}_0)$. Figure 1 shows the universal-kriging predictor of Australian temperature in the month of January 2009, mapped over the whole continent D^G , where the spatial locations $D^{G*} = \{\mathbf{s}_1, \dots, \mathbf{s}_n\}$ of weather stations that supplied the data $\mathbf{Z}$ are superimposed. Formulas for $\hat{Y}(\mathbf{s}_0)$ can be found in, for example, Chilès and Delfiner (2012, Section 3.4).

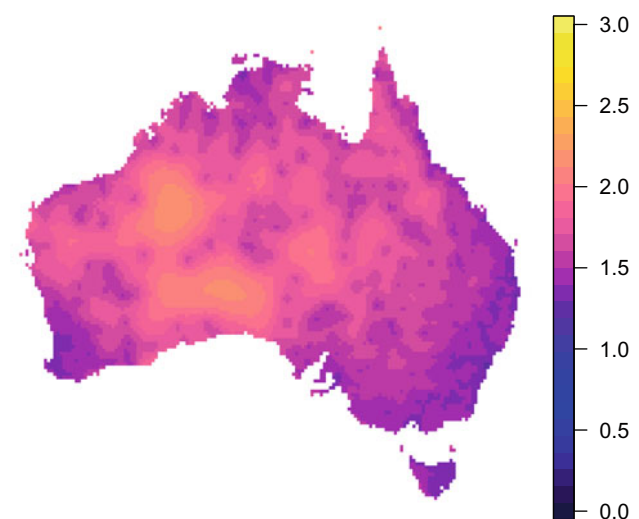
The optimized MSPE (17) is called the *kriging variance*, and its square root is called the *kriging standard error*:

$$\sigma_k(\mathbf{s}_0) \equiv \left\{ E \left(\left(\hat{Y}(\mathbf{s}_0) - Y(\mathbf{s}_0) \right)^2 \right) \right\}^{1/2}, \text{ for any } \mathbf{s}_0 \in D^G. \quad (18)$$

Figure 2 shows a map over D^G of the kriging standard error associated with the kriging predictor mapped in Fig. 1. It can be shown that a smaller $\sigma_k(\mathbf{s}_0)$ corresponds to a higher density of weather stations near $\mathbf{s}_0$. While ordinary and universal kriging produce an optimal linear predictor, there is an even better predictor, the *best optimal predictor (BOP)*, which is



Spatial Statistics, Fig. 1 Map of a kriging predictor of Australian temperature in January 2009, superimposed on spatial locations of data



Spatial Statistics, Fig. 2 Map of the kriging standard error (18) for the kriging predictor shown in Fig. 1

the best of all the best predictors obtained under a variety of constraints (e.g., linearity). From Bayes' Rule (10), the predictor that minimizes the MSPE (16) without any constraints is $Y^*(\mathbf{s}_0) \equiv E(Y(\mathbf{s}_0) | \mathbf{Z})$, which is the mean of the predictive distribution. Notice that the BOP, $Y^*(\mathbf{s}_0)$, is unbiased, namely, $E(Y^*(\mathbf{s}_0)) = E(Y(\mathbf{s}_0))$, without having to constrain it to be so.

Lattice Processes

In this section, we assume that the spatial locations D are given by D^L , a known countable subset of $\mathbb{R}^d$. This usually represents a collection of grid nodes, pixels, or small areas and the spatial locations associated with them; we write the countable set of all such locations as $D^L \equiv \{\mathbf{s}_1, \mathbf{s}_2, \dots\}$. Each $\mathbf{s}_i$ has a set of neighbors, $\mathcal{N}(\mathbf{s}_i) \subset D^L \setminus \mathbf{s}_i$, associated with it, and whose locations are spatially proximate (and note that a location is not considered to be its own neighbor). Spatial-statistical dependence between locations in lattice processes is defined in terms of these neighborhood relations.

Typically, the neighbors are represented by a spatial-dependence matrix $\mathbf{W}$ with entries w_{ij} nonzero if $\mathbf{s}_j \in \mathcal{N}(\mathbf{s}_i)$, and hence the diagonal entries of $\mathbf{W}$ are all zero. The non-diagonal entries of $\mathbf{W}$ might be, for example, inversely proportional to the distance, $\|\mathbf{s}_i - \mathbf{s}_j\|$, or they might involve some other way of moderating dependence based on spatial proximity. For example, they might be assigned the value 1 if a neighborhood relation exists and 0 otherwise. In this case, $\mathbf{W}$ is called an adjacency matrix, and it is symmetric if $\mathbf{s}_j \in \mathcal{N}(\mathbf{s}_i)$ whenever $\mathbf{s}_i \in \mathcal{N}(\mathbf{s}_j)$ and vice versa.

Consider a lattice process in $\mathbb{R}^2$ defined on the finite grid $D^L = \{(x, y) : x, y = 1, \dots, 5\}$. The first-order neighbors of the grid node (x, y) in the interior of the lattice are the four adjacent nodes, $\mathcal{N}(x, y) = \{(x-1, y), (x, y-1), (x+1, y), (x, y+1)\}$, shown as:

$$\begin{array}{ccccc} \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & \bullet & \circ & \circ \\ \circ & \bullet & \times & \bullet & \circ \\ \circ & \circ & \bullet & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ \end{array}$$

where grid node $\mathbf{s}_i$ is represented by $\times$, and its first-order neighbors are represented by $\bullet$. Nodes situated on the boundary of the grid will have less than four neighbors.

The most common type of lattice process is the *Markov random field* (MRF), which has a conditional-probability property in the spatial domain $\mathbb{R}^d$ that is a generalization of the temporal Markov property found in section “Spatiotemporal Processes”. A lattice process $\{Y(\mathbf{s}) : \mathbf{s} \in D^L\}$ is a MRF if, for all $\mathbf{s}_i \in D^L$, its conditional probabilities satisfy

$$[Y(\mathbf{s}_i) | \mathbf{Y}(D^L \setminus \mathbf{s}_i)] = [Y(\mathbf{s}_i) | \mathbf{Y}(\mathcal{N}(\mathbf{s}_i))], \quad (19)$$

where $\mathbf{Y}(A) \equiv \{Y(\mathbf{s}_j) : \mathbf{s}_j \in A\}$. The MRF is defined in terms of these conditional probabilities (19), which represent statistical dependencies between neighboring nodes that are captured differently from those given by the variogram or the covariance function. Specifically,

$$[Y(\mathbf{s}_i) | \mathbf{Y}(D^L \setminus \mathbf{s}_i)] = \frac{\exp\{-f(Y(\mathbf{s}_i), \mathbf{Y}(\mathcal{N}(\mathbf{s}_i)))\}}{C}, \quad (20)$$

where C is a normalizing constant that ensures the right-hand side of (20) integrates (or sums) to 1. Equation (20) is also known as a Gibbs random field in statistical mechanics since, under regularity conditions, the Hammersley-Clifford Theorem relates the joint probability distribution to the Gibbs measure (Besag 1974). The function $f(Y(\mathbf{s}_i), \mathbf{Y}(\mathcal{N}(\mathbf{s}_i)))$ is referred to as the potential energy, since it quantifies the strength of interactions between neighbors. A wide variety of MRF models can be defined by choosing different forms of the potential-energy function (Winkler 2003, Section 3.2). Note that care needs to be taken to ensure that specification of the model through all the conditional probability distributions, $\{[Y(\mathbf{s}_i) | \mathbf{Y}(\mathcal{N}(\mathbf{s}_i))] : \mathbf{s}_i \in D^L\}$, results in a valid joint probability distribution, $\{[Y(\mathbf{s}_i) : \mathbf{s}_i \in D^L]\}$ (Kaiser and Cressie 2000).

Revisiting the previous simple example of a first-order neighborhood structure on a regular lattice in $\mathbb{R}^2$, notice that grid nodes situated diagonally across from each other are conditionally independent. Hence, D^L can be partitioned into two sub-lattices D_1^L and D_2^L , such that the values at the nodes in D_1^L are independent given the values at the nodes in D_2^L and vice versa (Besag 1974; Winkler 2003, Section 8.1):

$$\begin{array}{ccccc} \bullet & \circ & \bullet & \circ & \bullet \\ \circ & \bullet & \circ & \bullet & \circ \\ \bullet & \circ & \bullet & \circ & \bullet \\ \circ & \bullet & \circ & \bullet & \circ \\ \bullet & \circ & \bullet & \circ & \bullet \end{array}$$

This forms a checkerboard pattern where $\{Y(\mathbf{s}) : \mathbf{s} \in D_1^L\}$ at nodes D_1^L represented by $\bullet$ are mutually independent, given the values $\{Y(\mathbf{u}) : \mathbf{u} \in D_2^L\}$ at nodes D_2^L represented by $\circ$.

Besag (1974) introduced the conditional autoregressive (CAR) model, which is a Gaussian MRF that is defined in terms of its conditional means and variances. We refer the reader to LeSage and Pace (2009) for discussion of a different lattice-process model, known as the simultaneous autoregressive (SAR) model, and a comparison of it with the CAR model. We define the CAR model as follows: For $\mathbf{s}_i \in D^L$, $Y(\mathbf{s}_i)$ is conditionally Gaussian defined by its first and second moments,

$$E(Y(\mathbf{s}_i) | \mathbf{Y}(\mathcal{N}(\mathbf{s}_i))) = \sum_{\mathbf{s}_j \in \mathcal{N}(\mathbf{s}_i)} c_{ij} Y(\mathbf{s}_j) \quad (21)$$

$$\text{var}(Y(\mathbf{s}_i) | \mathbf{Y}(\mathcal{N}(\mathbf{s}_i))) = \tau_i^2, \quad (22)$$

where c_{ij} are spatial autoregressive coefficients such that the diagonal elements $c_{1,1} = \dots = c_{n,n} = 0$, and $\{\tau_i^2\}$ are the variability parameters for the locations $\{\mathbf{s}_i\}$, respectively. Under an important regularity condition (see below), this specification results in a joint probability distribution that is multivariate Gaussian. That is,

$$\mathbf{Y} \sim \text{Gau}(\mathbf{0}, (\mathbf{I} - \mathbf{C})^{-1} \mathbf{M}), \quad (23)$$

where $\text{Gau}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$; the matrix $\mathbf{M} \equiv \text{diag}(\tau_1^2, \dots, \tau_n^2)$ is diagonal; and the regularity condition referred to above is that the coefficients $\mathbf{C} \equiv \{c_{ij}\}$ in (21) have to result in $\mathbf{M}^{-1}(\mathbf{I} - \mathbf{C})$ being a symmetric and positive-definite matrix. With a first-order neighborhood structure, such as shown in the simple example above in $\mathbb{R}^2$, the precision matrix is block-diagonal, which makes it possible to sample efficiently from this Gaussian MRF using sparse-matrix methods (Rue and Held 2005, Section 2.4).

The data vector for lattice processes is $\mathbf{Z} \equiv (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))'$, where $D^{L*} \equiv \{\mathbf{s}_1, \dots, \mathbf{s}_n\} \subset D^L$. As for the previous subsection, the data model is $[\mathbf{Z} | \{Y(\mathbf{s}) : \mathbf{s} \in D^L\}]$ which, to emphasize dependence on parameters θ , we reactivate earlier notation and write as $[\mathbf{Z} | Y, \theta]$. Now, if we write the lattice-process model as $[\{Y(\mathbf{s}) : \mathbf{s} \in D^L\} | \theta] \equiv [Y | \theta]$, then estimation of θ follows by maximizing the likelihood, $\mathcal{L}(\theta) \equiv \int [\mathbf{Z} | Y, \theta] [Y | \theta] dY$.

Regarding spatial prediction, $Y^*(\mathbf{s}_0) \equiv E(Y(\mathbf{s}_0) | \mathbf{Z}, \theta)$ is the best optimal predictor of $Y(\mathbf{s}_0)$, for $\mathbf{s}_0 \in D^L$ and known θ (e.g., Besag et al. 1991). Note that $\mathbf{s}_0$ may not belong to D^{L*} , and hence $Y^*(\mathbf{s}_0)$ is a predictor of $Y(\mathbf{s}_0)$ even when there is no datum observed at the node $\mathbf{s}_0$. Inference on unobserved parts of the process Y is just as important for lattice processes as it is for geostatistical processes.

Spatial Point Processes and Random Sets

A spatial point process is a countable collection of *random locations* $D \equiv D^P \subset \mathcal{D}$. Closely related to this random set of points is the counting process that we shall call $\{N(A) : A \subset \mathcal{D}\}$, where recall that $\mathcal{D}$ indexes all possible locations of interest, and now we assume it is bounded. For example, if A is a given subset of $\mathcal{D}$, and two of the random points $\{\mathbf{s}_i\}$ are contained in A , then $N(A) = 2$. Since $D^P = \{\mathbf{s}_i\}$ is

random and A is fixed, $N(A)$ is a random variable defined on the non-negative integers.

Clearly, the joint distributions $[N(A_1), \dots, N(A_m)]$, for any subsets $\{A_j : j = 1, \dots, m\}$ contained in $\mathcal{D}$ (possibly overlapping) and for any $m = 0, 1, 2, \dots$, are well defined. Spatial dependence can be seen through the spatial proximity between the $\{A_j\}$. Consider just two fixed subsets, A_1 and A_2 (i.e., $m = 2$) and, to avoid ambiguity caused by potentially sharing points, let $A_1 \cap A_2$ be empty. Then no spatial dependence is exhibited if, for any disjoint A_1 and A_2 , there is statistical independence; that is,

$$[N(A_1), N(A_2)] = [N(A_1)][N(A_2)]. \quad (24)$$

The basic spatial point process known as the *Poisson point process* has the independence property (24), and its associated counting process satisfies

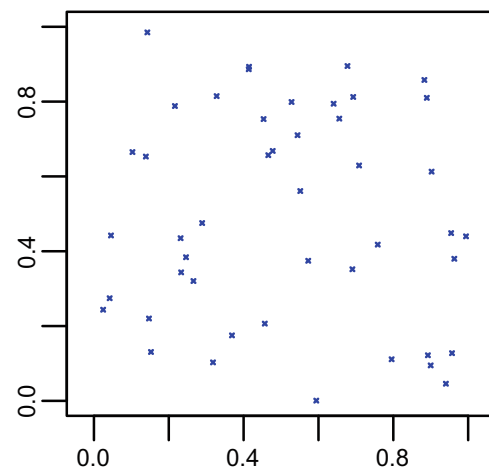
$$[N(A)] = \exp\{-\lambda(A)\} \frac{\lambda(A)^{N(A)}}{N(A)!}; \quad A \subset \mathcal{D}, \quad (25)$$

where $\lambda(A) \equiv \int_A \lambda(\mathbf{s}) d\mathbf{s}$. In (25), $\lambda(\cdot)$ is a given intensity function defined according to:

$$\lambda(\mathbf{s}) \equiv \lim_{|\delta\mathbf{s}| \rightarrow 0} \frac{E(Y(\delta\mathbf{s}))}{|\delta\mathbf{s}|}, \quad (26)$$

where $\delta\mathbf{s}$ is a small set centered at $\mathbf{s} \in \mathcal{D}$, and whose volume $|\delta\mathbf{s}|$ tends to 0.

In (25), the special case of $\lambda(\mathbf{s}) \equiv \lambda$, for all $\mathbf{s} \in \mathcal{D}$, results in a *homogeneous Poisson point process*, and a simulation of it is shown in Fig. 3. The simulation was obtained using an equivalent probabilistic representation for which the count random variable $N(\mathcal{D})$, for $\mathcal{D} = [0, 1] \times [0, 1]$, was simulated



Spatial Statistics, Fig. 3 A realization on the unit square $\mathcal{D} = [0, 1] \times [0, 1]$, of the homogeneous Poisson point process (25) with parameter $\lambda = 50$; for this realization, $N(\mathcal{D}) = 46$

according to (25). Then, conditional on $N(\mathcal{D})$, $\{\mathbf{s}_1, \dots, \mathbf{s}_{N(\mathcal{D})}\}$ was simulated independently and identically according to the uniform distribution,

$$[\mathbf{u}] = \begin{cases} \frac{1}{\lambda(\mathcal{D})}; & \mathbf{u} \in \mathcal{D} \\ 0; & \text{elsewhere.} \end{cases} \quad (27)$$

This representation explains why the homogenous Poisson point process is commonly referred to as a *Completely Spatially Random* (CSR) process, and why it is used as a baseline for testing for the absence of spatial dependence in a point process. That is, before a spatial model is fitted to a point pattern, a test of the null hypothesis that the point pattern originates from a CSR process, is often carried out. Rejection of CSR then justifies the fitting of spatially dependent point processes to the data (e.g., Ripley 1981; Diggle 2013).

Much of the early research in point processes was devoted to establishing test statistics that were sensitive to various types of departures from the CSR process (e.g., Cressie 1993, Section 8.2). This was followed by researchers' defining and then estimating spatial-dependence measures such as the second-order-intensity function and the K-function (e.g., Ripley 1981, Chapter 8), where inference was often in terms of method-of-moments estimation of these functions. More efficient likelihood-based inference came later; Baddeley et al. (2015) give a comprehensive review of these methodologies for point processes. From a modeling perspective, particular attention has been paid to the log-Gaussian Cox point processes; here, $\lambda(\cdot)$ in (25) is random, such that $\{\log(\lambda(\mathbf{s})) : \mathbf{s} \in \mathcal{D}\}$ is a Gaussian process (e.g., Møller and Waagepetersen 2003). This model leads naturally to hierarchical Bayesian inference for $\lambda(\cdot)$ and its parameters (e.g., Gelfand and Schliep 2018).

If an attribute process, $\{Y(\mathbf{s}_i) : \mathbf{s}_i \in D^P\}$, is included with the spatial point process D^P , one obtains a so-called *marked point process* (e.g., Cressie 1993, Section 8.7). For example, the study of a natural forest where both the locations $\{\mathbf{s}_i\}$ and the sizes of the trees $\{Y(\mathbf{s}_i)\}$ are modeled together probabilistically, results in a marked point process where the "mark" process is a spatial process $\{Y(\mathbf{s}_i) : \mathbf{s}_i \in D^P\}$ of tree size. Now, Bayes' Rule given by (10), where both Y and $D (= D^P)$ are random, should be used to make inference on Y and D^P through the predictive distribution $[Y, D^P | \mathbf{Z}]$. Here, $\mathbf{Z}$ consists of the number of trees, the trees' locations, and the trees' size measurements, as denoted in (5). After marginalization, we can obtain $[D^P | \mathbf{Z}]$, the predictive distribution of the spatial point process D^P .

A spatial point process is a special case of a random set, which is a random quantity in Euclidean space that was defined rigorously by Matheron (1975). Some geological

processes are more naturally modeled as set-valued phenomena (e.g., the facies of a mineralization); however, inference for random-set processes has lagged behind those for spatial point processes. It is difficult to define a likelihood based on set-valued data, which has held back statistically efficient inferences; nevertheless, basic method-of-moment estimators are often available. The most well-known random set that allows statistical inference from set-valued data is the Boolean Model (e.g., Cressie and Wikle 2011, Section 4.4).

Multivariate Spatial Processes

The previous subsections have presented single spatial statistical processes but, as models become more realistic representations of a complex world, there is a need to express interactions between multiple processes. This is most directly seen by modeling vector-valued geostatistical processes $\{\mathbf{Y}(\mathbf{s}) : \mathbf{s} \in D^G\}$, and vector-valued lattice processes $\{\mathbf{Y}(\mathbf{s}_i) : \mathbf{s}_i \in D^L\}$, where the k -dimensional vector $\mathbf{Y}(\mathbf{s}) \equiv (Y_1(\mathbf{s}), \dots, Y_k(\mathbf{s}))'$ represents the multiple processes at the generic location $\mathbf{s} \in \mathcal{D}$. Vector-valued spatial point processes can be represented as a set of k point processes, $\{\{\mathbf{s}_{1,i}\}, \dots, \{\mathbf{s}_{k,i}\}\}$, and these are presented in Baddeley et al. (2015, Chapter 14). If we adopt a hierarchical-statistical-modeling approach, it is possible to construct multivariate spatial processes whose component univariate processes could come from any of the three types of spatial processes presented in the previous three subsections. This is because, at a deeper level of the hierarchy, a core *multivariate geostatistical process* can control the spatial dependence for processes of any type, which allows the possibility of hybrid multivariate spatial statistical processes.

In what follows, we describe briefly two approaches to constructing multivariate geostatistical processes, one based on a joint approach and the other based on a conditional approach. We consider the case $k = 2$, namely, the bivariate spatial process $\{(Y_1(\mathbf{s}), Y_2(\mathbf{s}))' : \mathbf{s} \in D^G\}$, for illustration. The joint approach involves directly constructing a valid spatial statistical model from $\boldsymbol{\mu}(\mathbf{s}) \equiv (\mu_1(\mathbf{s}), \mu_2(\mathbf{s}))' \equiv (E(Y_1(\mathbf{s})), E(Y_2(\mathbf{s})))'$, for $\mathbf{s} \in D^G$, and from

$$\text{cov}(Y_l(\mathbf{s}), Y_m(\mathbf{u})) \equiv C_{lm}(\mathbf{s}, \mathbf{u}); l, m = 1, 2, \quad (28)$$

for $\mathbf{s}, \mathbf{u} \in D^G$. The bivariate-process mean $\boldsymbol{\mu}(\cdot)$ is typically modeled as a vector linear regression; hence it is straightforward to model the bivariate mean once the appropriate regressors have been chosen.

Analogous to the univariate case, the set of covariance and cross-covariance functions, $\{C_{11}(\cdot, \cdot), C_{22}(\cdot, \cdot), C_{12}(\cdot, \cdot), C_{21}(\cdot, \cdot)\}$, have to satisfy positive-definiteness conditions for the bivariate geostatistical model to be valid, and it is important to note that, in general, $C_{12}(\mathbf{s}, \mathbf{u}) \neq C_{21}(\mathbf{s}, \mathbf{u})$. There are classes of valid models that exhibit symmetric cross-

dependence, namely, $C_{12}(\mathbf{s}, \mathbf{u}) = C_{21}(\mathbf{s}, \mathbf{u})$, such as the linear model of co-regionalization (Gelfand et al. 2004). These are not reasonable models for ore-reserve estimation when there has been preferential mineralization in the ore body.

The joint approach can be contrasted with a conditional approach (Cressie and Zammit-Mangion 2016), where each of the k processes is a node of a directed acyclic graph that guides the conditional dependence of any process, given the remaining processes. Again consider the bivariate case (i.e., $k = 2$), where there are only two nodes such that $Y_1(\cdot)$ is at node 1, $Y_2(\cdot)$ is at node 2, and a directed edge is declared from node 1 to node 2. Then the appropriate way to model the joint distribution is through

$$[Y_1(\cdot), Y_2(\cdot)] = [Y_2(\cdot) | Y_1(\cdot)] [Y_1(\cdot)], \quad (29)$$

where $[Y_2(\cdot) | Y_1(\cdot)]$ is shorthand for $[Y_2(\cdot) | \{Y_1(\mathbf{s}) : \mathbf{s} \in D^G\}]$.

The geostatistical model for $[Y_1(\cdot)]$ is simply a univariate spatial model based on a mean function $\mu_1(\cdot)$ and a valid covariance function $C_{11}(\cdot, \cdot)$, which was discussed in section “Geostatistical Processes”. Now assume that $Y_2(\cdot)$ depends on $Y_1(\cdot)$ as follows: For $\mathbf{s}, \mathbf{u} \in D^G$,

$$\begin{aligned} E[Y_2(\mathbf{s}) | Y_1(\cdot)] &\equiv \mu_2(\mathbf{s}) + \int_{D^G} b(\mathbf{s}, \mathbf{v}) \\ &\quad \times (Y_1(\mathbf{v}) - \mu_1(\mathbf{v})) \, d\mathbf{v}, \end{aligned} \quad (30)$$

$$\text{cov}(Y_2(\mathbf{s}), Y_2(\mathbf{u}) | Y_1(\cdot)) \equiv C_{2|1}(\mathbf{s}, \mathbf{u}), \quad (31)$$

where $C_{2|1}(\cdot, \cdot)$ is a valid univariate covariance function and $b(\cdot, \cdot)$ is an integrable interaction function. The conditional-moment assumptions given by (30) and (31) follow if one assumes that $(Y_1(\cdot), Y_2(\cdot))'$ is a bivariate Gaussian process.

Cressie and Zammit-Mangion (2016) show that, from (30) and (31),

$$C_{12}(\mathbf{s}, \mathbf{u}) = \int_{D^G} C_{11}(\mathbf{s}, \mathbf{v}) b(\mathbf{u}, \mathbf{v}) \, d\mathbf{v} \quad (32)$$

$$C_{21}(\mathbf{s}, \mathbf{u}) = \int_{D^G} C_{11}(\mathbf{v}, \mathbf{u}) b(\mathbf{v}, \mathbf{s}) \, d\mathbf{v} \quad (33)$$

$$\begin{aligned} C_{22}(\mathbf{s}, \mathbf{u}) &= C_{2|1}(\mathbf{s}, \mathbf{u}) \\ &\quad + \int_{D^G} \int_{D^G} b(\mathbf{s}, \mathbf{v}) C_{11}(\mathbf{v}, \mathbf{w}) b(\mathbf{u}, \mathbf{w}) \, d\mathbf{v} \, d\mathbf{w}, \end{aligned} \quad (34)$$

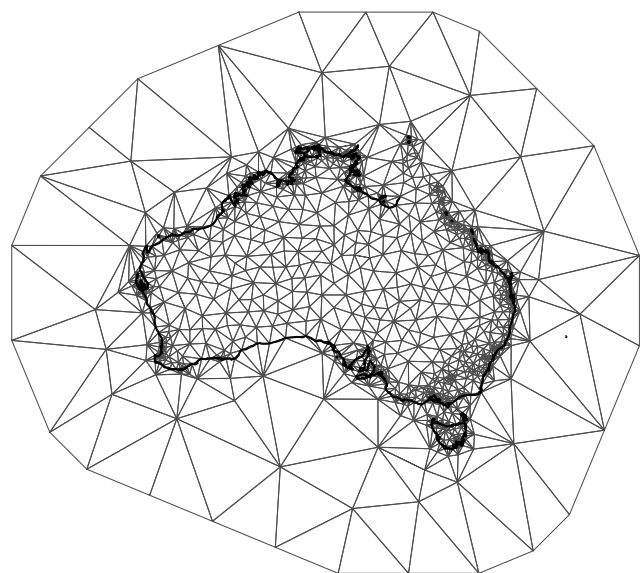
for $\mathbf{s}, \mathbf{u} \in D^G$. Along with $\mu_1(\cdot)$, $\mu_2(\cdot)$, and $C_{11}(\cdot, \cdot)$, these functions (32)–(34) define a valid bivariate geostatistical process $[Y_1(\cdot), Y_2(\cdot)]$. A notable property of the conditional approach is that asymmetric cross-dependence (i.e., $C_{12}(\mathbf{s}, \mathbf{u}) \neq C_{21}(\mathbf{s}, \mathbf{u})$) occurs if $b(\mathbf{s}, \mathbf{u}) \neq b(\mathbf{u}, \mathbf{s})$.

In summary, the conditional approach allows multivariate modeling to be carried out validly by simply specifying $\boldsymbol{\mu}(\cdot) = (\mu_1(\cdot), \mu_2(\cdot))'$ and two valid univariate covariance functions, $C_1(\cdot, \cdot)$ and $C_{2|1}(\cdot, \cdot)$. The strengths of the conditional approach are that only univariate covariance functions need to be specified (for which there is a very large body of research; e.g., Cressie and Wikle 2011, Chapter 4), and that only integrability of $b(\cdot, \cdot)$, the interaction function, needs to be assumed (Cressie and Zammit-Mangion 2016).

Spatial Discretization

Although geostatistical processes are defined on a continuous spatial domain D^G , this can limit the practical feasibility of statistical inferences due to computational and mathematical considerations. For example, kriging from an n -dimensional vector of data involves the inversion of an $n \times n$ covariance matrix, which requires order n^3 floating-point operations and order n^2 storage in available memory. These costs can be prohibitive for large spatial datasets; hence, spatial discretization to achieve scalable computation for spatial models is an active area of research.

In practical applications, spatial statistical inference is required up to a finite spatial resolution. Many approaches take advantage of this by dividing the spatial domain $\mathcal{D}$ into a lattice of discrete points in $\mathcal{D}$, as shown in Fig. 4. As a consequence of this discretization, a geostatistical process can be approximated by a lattice process, such as a Gaussian MRF (e.g., Rue and Held 2005, Section 5.1); however, sometimes this can result in undesirable discretization errors and artifacts. More sophisticated approaches have been developed



Spatial Statistics, Fig. 4 Discretization of the spatial domain $\mathcal{D}$, a convex region around Australia, into a triangular lattice

to obtain highly accurate approximations of a geostatistical (i.e., continuously indexed) spatial process evaluated over an irregular lattice, as we now discuss.

Let the original domain $\mathcal{D}$ be bounded and suppose it is tessellated into the areas $\{A_j \subset \mathcal{D} : j = 1, \dots, m\}$ that are small, non-overlapping *basic areal units* (BAUs; Nguyen et al. 2012), so that $\mathcal{D} = \bigcup_{j=1}^m A_j$, and $A_j \cap A_k$ is the empty set for any $j \neq k \in \{1, \dots, m\}$; Fig. 4 gives an example of triangular BAUs. Spatial basis functions $\{\phi_\ell(\cdot) : \ell = 1, \dots, r\}$, can then be defined on the BAUs. For example, Lindgren et al. (2011) used triangular basis functions where $r > m$, while fixed rank kriging (FRK; Cressie and Johannesson 2008; Zammit-Mangion and Cressie 2021) can employ a variety of different basis functions for $r < m$, including multi-resolution wavelets and bisquares.

Vecchia approximations (e.g., Datta et al. 2016; Katzfuss et al. 2020) are also defined using a lattice of discrete points $D^L \subset D^G \subset \mathcal{D}$, that include the coordinates of the observed data $D^{G*} = \{\mathbf{s}_1, \dots, \mathbf{s}_n\}$ and the prediction locations $\{\mathbf{s}_{n+1}, \dots, \mathbf{s}_{n+p}\}$. Let $[\mathbf{X}] \equiv [\mathbf{Z}, \mathbf{Y}]$, where data $\mathbf{Z}$ and spatial process $\mathbf{Y}$ are associated with the lattice $D^L \equiv \{\mathbf{s}_1, \dots, \mathbf{s}_n, \mathbf{s}_{n+1}, \dots, \mathbf{s}_{n+p}\}$. It is a property of joint and conditional distributions that this can be factorized into a product:

$$[\mathbf{X}] = [X(\mathbf{s}_1)] \prod_{i=2}^{n+p} [X(\mathbf{s}_i) | X(\mathbf{s}_1), \dots, X(\mathbf{s}_{i-1})]. \quad (35)$$

In the previous section, the set of spatial coordinates D^L had no fixed ordering. However, a Vecchia approximation requires that an artificial ordering is imposed on $\{\mathbf{s}_1, \dots, \mathbf{s}_{n+p}\}$. Let the ordering be denoted by $\{\mathbf{s}_{(1)}, \dots, \mathbf{s}_{(n+p)}\}$, and define the set of neighbors $\mathcal{N}(\mathbf{s}_{(i)}) \subset \{\mathbf{s}_{(1)}, \dots, \mathbf{s}_{(i-1)}\}$, similarly to “Lattice Processes,” except that these neighborhood relations are not reciprocal: If for $j < i$, $\mathbf{s}_{(j)}$ belongs to $\mathcal{N}(\mathbf{s}_{(i)})$, then $\mathbf{s}_{(i)}$ cannot belong to $\mathcal{N}(\mathbf{s}_{(j)})$. As part of the Vecchia approximation, a fixed upper bound $q \ll n$ on the number of neighbors is chosen. That is, $|\mathcal{N}(\mathbf{s}_{(i)})| \leq q$, so that the lattice formed by $\{\mathcal{N}(\mathbf{s}_{(i)}) : i = 1, \dots, n+p\}$ is a directed acyclic graph, which results in a partial order in $\mathcal{D}$.

The joint distribution $[\mathbf{X}]$ given by (35) is then approximated by:

$$[X(\mathbf{s}_{(1)})] \prod_{i=2}^{n+p} [X(\mathbf{s}_{(i)}) | \mathbf{X}(\mathcal{N}(\mathbf{s}_{(i)}))] \equiv [\tilde{\mathbf{X}}], \quad (36)$$

which is a Partially Ordered Markov Model (POMM; Cressie and Davidson 1998). This Vecchia approximation, $[\tilde{\mathbf{X}}]$, is a distribution coming from a valid spatial process on the original, uncountable, unordered index set D^G (Datta et al. 2016), which means that it can be used as a geostatistical process model with considerable computational advantages. For

example, it can be used as a random log-intensity function, $\log(\lambda(\mathbf{s}))$, in a hierarchical point-process model, or it can be combined with other processes to define models described in section “Multivariate Spatial Processes”. However, in all of these contexts it should be remembered that the resulting predictive process, $[\tilde{\mathbf{X}} | \mathbf{Z}]$, is an approximation to the true predictive process, $[\mathbf{X} | \mathbf{Z}]$.

Spatiotemporal Processes

The section “Multivariate Spatial Processes” introduced processes that were written in vector form as,

$$\mathbf{Y}(\mathbf{s}) \equiv (Y_1(\mathbf{s}), \dots, Y_k(\mathbf{s}))'; \mathbf{s} \in D^G. \quad (37)$$

In that section, we distinguished between the joint approach and the conditional approach to multivariate-spatial-statistical modeling and, under the conditional approach, we used a directed acyclic graph to give a blueprint for modeling the multivariate spatial dependence.

Now, consider a spatiotemporal process,

$$\{Y(\mathbf{s}; t) : \mathbf{s} \in D^G; t \in \mathcal{T}\}, \quad (38)$$

where $\mathcal{T}$ is a temporal index set. Clearly, if $\mathcal{T} = \{1, 2, \dots\}$ then (38) becomes a spatial process of time series, $\{Y(\mathbf{s}; 1), Y(\mathbf{s}; 2), \dots : \mathbf{s} \in D^G\}$. If $\mathcal{T} = \{1, 2, \dots, k\}$, and we define $Y_j(\mathbf{s}) \equiv Y(\mathbf{s}; j)$, for $j = 1, \dots, k$, then the resulting spatiotemporal process can be represented as a multivariate spatial process given by (37). Not surprisingly, the same dichotomy of approach to modeling statistical dependence (i.e., joint versus conditional) occurs for spatiotemporal processes as it does for multivariate spatial processes.

Describing all possible covariances between Y at any spatiotemporal “location” $(\mathbf{s}; t)$ and any other one $(\mathbf{u}; v)$, amounts to treating “time” as simply another dimension to be added to the d -dimensional Euclidean space, $\mathbb{R}^d$. Taking this approach, spatiotemporal statistical dependence can be expressed in $(d+1)$ -dimensional space through the covariance function,

$$C(\mathbf{s}; t, \mathbf{u}; v) \equiv \text{cov}(Y(\mathbf{s}; t), Y(\mathbf{u}; v)); \quad \mathbf{s}, \mathbf{u} \in D^G, t, v \in \mathcal{T}. \quad (39)$$

Of course, the time dimension has different units than the spatial dimensions, and its interpretation is different since the future is unobserved. Hence, the joint modeling of space and time based on (39) must be done with care to account for the special nature of the time dimension in this *descriptive approach* to spatiotemporal modeling.

From current and past spatiotemporal data $\mathbf{Z}$, predicting past values of Y is called *smoothing*, predicting unobserved values of the current Y is called *filtering*, and predicting future values of Y is called *forecasting*. The Kalman filter (Kalman 1960) was developed to provide fast predictions of the current state using a methodology that recognizes the ordering of the time dimension. Today's filtered values become "old" the next day when a new set of current data are received. Using a dynamical approach that we shall now describe, the Kalman filter updates yesterday's optimal filtered value with today's data very rapidly, to obtain a current optimal filtered value.

The best way to describe the dynamical approach is to discretize the spatial domain. The previous section "[Spatial Discretization](#)", describes a number of ways this can be done; here we shall consider the discretization that is most natural for storing the attribute and location information in computer memory, namely, a fine-resolution lattice D^L of pixels or voxels (short for "volume elements"). Replace $\{Y(\mathbf{s};t) : \mathbf{s} \in D^G, t = 1, 2, \dots\}$ with $\{Y(\mathbf{s};t) : \mathbf{s} \in D^L, t = 1, 2, \dots\}$, where $D^L \equiv \{\mathbf{s}_1, \dots, \mathbf{s}_m\}$ are the centroids of elements of small area (or small volume) that make up D^G . Often the areas of these elements are equal, having been defined by a regular grid. As we explain below, this allows a *dynamical approach* to constructing a statistical model for the spatiotemporal process Y on the discretized space-time cube, $\{\mathbf{s}_1, \dots, \mathbf{s}_m\} \times \{1, 2, \dots\}$.

Define $\mathbf{Y}_t \equiv (Y(\mathbf{s};t) : \mathbf{s} \in D^L)'$, which is an m -dimensional vector. Because of the temporal ordering, we can write the joint distribution of $\{Y(\mathbf{s};t) : \mathbf{s} \in D^L, t = 1, \dots, k\}$ from $t = 1$ up to the present time $t = k$, as

$$[\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_k] = [\mathbf{Y}_1] \times [\mathbf{Y}_2 | \mathbf{Y}_1] \dots [\mathbf{Y}_k | \mathbf{Y}_{k-1}, \dots, \mathbf{Y}_2, \mathbf{Y}_1], \quad (40)$$

which has the same form as (35). Note that this conditional modeling of space and time is a natural approach, since time is completely ordered. The next step is to make a Markov assumption, and hence (40) can be written as

$$[\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_k] = [\mathbf{Y}_1] \prod_{j=2}^k [\mathbf{Y}_j | \mathbf{Y}_{j-1}]. \quad (41)$$

This is the same Markov property that we previously discussed in section "[Lattice Processes](#)", except it is now applied to the completely ordered one-dimensional domain, $\mathcal{T} = \{1, 2, \dots\}$, and $\mathcal{N}(j) = j - 1$. The Markov assumption makes our approach dynamical: It says that the present, conditional on the past, in fact only depends on the "most recent past." That is, since $\mathcal{N}(j) = j - 1$, the factor $[\mathbf{Y}_j | \mathbf{Y}_{j-1}, \dots, \mathbf{Y}_2, \mathbf{Y}_1] = [\mathbf{Y}_j | \mathbf{Y}_{j-1}]$, which results in the model (41).

For further information on the types of models used in the descriptive approach given by (39) and the types of models used in the dynamical approach given by (41), see Cressie and

Wikle (2011, Chapters 6–8) and Wikle et al. (2019, Chapters 4 and 5). The statistical analysis of observations from these processes is known as *spatiotemporal statistics*. Inference (estimation and prediction) from spatiotemporal data using R software can be found in Wikle et al. (2019).

Conclusion

Spatial-statistical methods distinguish themselves from spatial-analysis methods found in the geographical and environmental sciences, by providing well-calibrated quantification of the uncertainty involved with estimation or prediction. Uncertainty in the scientific phenomenon of interest is represented by a spatial *process model*, $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$, defined on possibly random D in $\mathbb{R}^d$, while measurement uncertainty in the observations $\mathbf{Z}$ is represented in a *data model*. In section "[Introduction](#)", we saw how these two models are combined using Bayes' Rule (10), or the simpler version (11), to calculate the overall uncertainty needed for statistical inference.

There are three main types of spatial process models: geostatistical processes where uncertainty is in the process Y , which is indexed continuously in $D = D^G$; lattice processes where uncertainty is also in Y , but now Y is indexed over a countable number of spatial locations $D = D^L$; and point processes where uncertainty is in the spatial locations $D = D^P$. Multiple spatial processes can interact with each other to form a multivariate spatial process. Importantly, processes can vary over time as well as spatially, forming a spatiotemporal process.

With some exceptions, spatial-statistical models (1) rarely consider the case of measurement error in the locations in $\mathcal{D}$. Here we focus on a spatial-statistical model for the location error (Cressie and Kornak 2003): Write the observed locations as $D^* \equiv \{\mathbf{u}_i : i = 1, \dots, n\}$; in this case, a part of the data model is $[D^* | D]$, and a part of the process model is $[D]$. Finally then, the data consist of both locations and attributes and are $\mathbf{Z} \equiv \{(\mathbf{u}_i, Z(\mathbf{u}_i)) : i = 1, \dots, n\}$, the spatial process model is $[Y, D]$, and the data model is $[\mathbf{Z} | Y, D]$. Then Bayes' Rule given by (10) is used to infer the unknowns Y and D from the predictive distribution $[Y, D | \mathbf{Z}]$.

As the size of spatial datasets have been increasing dramatically, more and more attention has been devoted to scalable computation for spatial-statistical models. Of particular interest are methods that use spatial discretization to approximate a continuous spatial domain, D^G . There are other recent advances in spatial statistics that we feel are important to mention, but their discussion here is necessarily brief.

Physical barriers can sometimes interrupt the statistical association between locations in close spatial proximity. Barrier models (Bakka et al. 2019) have been developed to account for these kinds of discontinuities in the spatial correlation function. Other methods for modeling nonstationarity,

anisotropy, and heteroskedasticity in spatial process models are an active area of research.

It can often be difficult to select appropriate prior distributions for the parameters of a stationary spatial process, for example its correlation-length scale. Penalized complexity (PC) priors (Simpson et al. 2017) are a way to encourage parsimony by favoring parameter values that result in the simplest model consistent with the data. The likelihood function of a point process or of a non-Gaussian lattice model can be both analytically and computationally intractable. Surrogate models, emulators, and quasi-likelihoods have been developed to approximate these intractable likelihoods (Moore et al. 2020).

Copulas are an alternative method for modeling spatial dependence in multivariate data, particularly when the data are non-Gaussian (Krupskii and Genton 2019). One area where non-Gaussianity can arise is in modeling the spatial association between extreme events, such as for temperature or precipitation (Tawn et al. 2018; Bacro et al. 2020).

As a final comment, we reflect on how the field of geostatistics has evolved, beginning with applications of spatial stochastic processes to mining: In the 1970s, Georges Matheron and his Centre of Mathematical Morphology in Fontainebleau were part of the Paris School of Mines, a celebrated French tertiary-education and research institution. To see what the geostatistical methodology of the time was like, the interested reader could consult Journel and Huijbregts (1978), for example. Over the following decade, geostatistics became notationally and methodologically integrated into statistical science and the growing field of spatial statistics (e.g., Ripley 1981; Cressie 1993). It took one or two more decades before geostatistics became integrated into the hierarchical-statistical-modeling approach to spatial statistics (e.g., Banerjee et al. 2003; Cressie and Wikle 2011, Chapter 4). The presentation given in our review takes this latter viewpoint and explains well-known geostatistical quantities such as the variogram and kriging in terms of this advanced, modern view of geostatistics. We also include a discussion of uncertainty in the spatial index set as part of our review, which offers new insights into spatial-statistical modeling. Probabilistic difficulties with geostatistics, of making inference on a possibly non-countable number of spatial random variables from a finite number of observations, can be finessed by discretizing the process. In a modern computing environment, this is key to carrying out spatial-statistical inference (including kriging).

Cross-References

- Bayesian Inversion in Geoscience
- Cressie, Noel A.C.
- Krige, Daniel Gerhardus

- Matheron, Georges
- Geostatistics
- High-Order Spatial Stochastic Models
- Hypothesis Testing
- Interpolation
- Kriging
- Markov Random Fields
- Multiple Point Statistics
- Multivariate Analysis
- Point Pattern Statistics
- Spatial Analysis
- Spatial Autocorrelation
- Spatial Data
- Spatiotemporal Analysis
- Spatiotemporal Modeling
- Statistical Computing
- Stochastic Geometry in the Geosciences
- Tobler, Waldo
- Uncertainty Quantification
- Variogram

Acknowledgments Cressie's research was supported by an Australian Research Council Discovery Project (Project number DP190100180). Our thanks go to Karin Karr and Laura Cartwright for their assistance in typesetting the manuscript.

Bibliography

- Bacro JN, Gaetan C, Opitz T, Toulemonde G (2020) Hierarchical space-time modeling of asymptotically independent exceedances with an application to precipitation data. *J Am Stat Assoc* 115(530):555–569. <https://doi.org/10.1080/01621459.2019.1617152>
- Baddeley A, Rubak E, Turner R (2015) *Spatial point patterns: Methodology and applications with R*. Chapman & Hall/CRC Press, Boca Raton
- Bakka H, Vanhatalo J, Illian JB, Simpson D, Rue H (2019) Non-stationary Gaussian models with physical barriers. *Spat Stat* 29: 268–288. <https://doi.org/10.1016/j.spasta.2019.01.002>
- Banerjee S, Carlin BP, Gelfand AE (2003) *Hierarchical modeling and analysis for spatial data*. Chapman & Hall/CRC Press, Boca Raton
- Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. *J R Stat Soc Ser B Stat Methodol* 36(2):192–236
- Besag J, York J, Mollié A (1991) Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math* 43:1–20
- Chilès JP, Delfiner P (2012) *Geostatistics: Modeling spatial uncertainty*, 2nd edn. Wiley, Hoboken
- Cressie N (1990) The origins of kriging. *Math Geol* 22(3):239–252. <https://doi.org/10.1007/BF00889887>
- Cressie N (1993) *Statistics for spatial data*, Revised edn. Wiley, Hoboken
- Cressie N, Davidson JL (1998) Image analysis with partially ordered Markov models. *Comput Stat Data Anal* 29(1):1–26
- Cressie N, Johannesson G (2008) Fixed rank kriging for very large spatial data sets. *J R Stat Soc Ser B Stat Methodol* 70(1):209–226. <https://doi.org/10.1111/j.1467-9868.2007.00633.x>
- Cressie N, Kornak J (2003) Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Stat Sci* 18(4):436–456. <https://doi.org/10.1214/ss/1081443228>

- Cressie N, Wikle CK (2011) *Statistics for spatio-temporal data*. Wiley, Hoboken
- Cressie N, Zammit-Mangion A (2016) Multivariate spatial covariance models: A conditional approach. *Biometrika* 103(4):915–935. <https://doi.org/10.1093/biomet/asw045>
- Datta A, Banerjee S, Finley AO, Gelfand AE (2016) Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *J Am Stat Assoc* 111(514):800–812. <https://doi.org/10.1080/01621459.2015.1044091>
- Diggle PJ (2013) *Statistical analysis of spatial and spatio-temporal point patterns*, 3rd edn. Chapman & Hall/CRC Press, Boca Raton
- Gelfand AE, Schliep EM (2018) Bayesian inference and computing for spatial point patterns. In: NSF-CBMS regional conference series in probability and statistics, vol 10. Institute of Mathematical Statistics and the American Statistical Association, Alexandria, VA, pp i–125
- Gelfand AE, Schmidt AM, Banerjee S, Sirmans C (2004) Nonstationary multivariate process modeling through spatially varying coregionalization. *TEST* 13(2):263–312
- Journel AG, Huijbregts CJ (1978) *Mining geostatistics*. Academic Press, London
- Kaiser MS, Cressie N (2000) The construction of multivariate distributions from Markov random fields. *J Multivar Anal* 73(2):199–220
- Kalman RE (1960) A new approach to linear filtering and prediction problems. *Trans ASME J Basic Eng* 82:35–45
- Katzfuss M, Guinness J, Gong W, Zilber D (2020) Vecchia approximations of Gaussian-process predictions. *J Agric Biol Environ Stat* 25(3):383–414. <https://doi.org/10.1007/s13253-020-00401-7>
- Krupskii P, Genton MG (2019) A copula model for non-Gaussian multivariate spatial data. *J Multivar Anal* 169:264–277
- LeSage J, Pace RK (2009) *Introduction to spatial econometrics*. Chapman & Hall/CRC Press, Boca Raton
- Lindgren F, Rue H, Lindström J (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J R Stat Soc Ser B Stat Methodol* 73(4):423–498. <https://doi.org/10.1111/j.1467-9868.2011.00777.x>
- Matheron G (1963) Principles of geostatistics. *Econ Geol* 58(8):1246–1266
- Matheron G (1975) *Random sets and integral geometry*. Wiley, Hoboken
- Møller J, Waagepetersen RP (2003) *Statistical inference and simulation for spatial point processes*. Chapman & Hall/CRC Press, Boca Raton
- Moores MT, Pettitt AN, Mengersen KL (2020) Bayesian computation with intractable likelihoods. In: *Case studies in applied Bayesian data science*. Springer, Berlin, pp 137–151
- Nguyen H, Cressie N, Braverman A (2012) Spatial statistical data fusion for remote sensing applications. *J Am Stat Assoc* 107(499):1004–1018. <https://doi.org/10.1080/01621459.2012.694717>
- Ripley BD (1981) *Spatial statistics*. Wiley, Hoboken
- Rue H, Held L (2005) *Gaussian Markov random fields: Theory and applications*. Chapman & Hall/CRC Press, Boca Raton
- Simpson D, Rue H, Riebler A, Martins TG, Sørbye SH (2017) Penalising model component complexity: A principled, practical approach to constructing priors. *Stat Sci* 32(1):1–28
- Tawn J, Shooter R, Towe R, Lamb R (2018) Modelling spatial extreme events with environmental applications. *Spat Stat* 28:39–58
- Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. *Econ Geogr* 46(suppl):234–240
- Upton G, Fingleton B (1985) *Spatial data analysis by example, volume 1: point pattern and quantitative data*. Wiley, Hoboken
- Wikle CK, Zammit-Mangion A, Cressie N (2019) *Spatio-temporal statistics with R*. Chapman & Hall/CRC Press, Boca Raton
- Winkler G (2003) *Image analysis, random fields and Markov Chain Monte Carlo methods: A mathematical introduction*, 2nd edn. Springer, Berlin
- Zammit-Mangion A, Cressie N (2021) FRK: an R package for spatial and spatio-temporal prediction with large datasets. *J Stat Softw*, vol 98, pp 1–48

Spatiotemporal

Sandra De Iaco¹, Donald E. Myers² and Donato Posa³

¹Department of Economics, Section of Mathematics and Statistics, University of Salento, National Biodiversity Future Center, Lecce, Italy

²Department of Mathematics, University of Arizona, Tucson, AZ, USA

³Department of Economics, Section of Mathematics and Statistics, University of Salento, Lecce, Italy

Synonyms

Space-time, $\mathbb{R}^d \times \mathbb{R}$ domain, product space

Definition

The term spatio-temporal is associated with a wide area of research, which aims to provide methods and tools for analyzing the evolution of a phenomenon over a spatial and temporal domain. There are many variables that can be viewed as phenomena with a spatio-temporal variability, such as meteorological readings (temperature, humidity, pressure), hydrological parameters (permeability and hydraulic conductivity), and many measures of air, soil, and water pollution. In this context, methods for analyzing and explaining the joint spatio-temporal behavior of the above variables are necessary in order to get different goals: optimization of sampling design network, estimation at unsampled spatial locations or unsampled times and computation of maps of predicted values, assessing the uncertainty of predicted values starting from the experimental measurements, trends' detection in space and time, particularly important to cope with risks coming from concentrations of hazardous pollutants. Hence, more and more attention is given to the spatial-temporal analysis in order to sort out these issues. The term spatio-temporal and space-time are generally used interchangeably in literature.

Geostatistical Approach

Geostatistics is a branch of spatial and spatio-temporal statistics that uses the formalism of random functions to analyze data characterized by spatial or spatio-temporal structure. It is assumed that the data relate to a phenomenon with an underlying continuous evolution, and thus other types of phenomena, e.g., lattice or point processes, are excluded from geostatistical analysis and are not the object of this contribution. Geostatistics was born to solve mining engineering problems; however, the domain of geostatistical applications

is much broader to date. Important contributions to the current knowledge about space-time models can be also found in Kyriakidis and Journel (1999) as well as in Christakos (2017), which skillfully combine theory and modeling aspects. A recent text by Cressie and Wikle (2011) presents a statistical approach to spatio-temporal data modeling from the perspective of Bayesian Hierarchical Modeling.

The main advantages of geostatistical techniques in modeling spatio-temporal variables are related to the possibility of:

- Choosing an appropriate spatio-temporal model for the non-constant mean (or deterministic component), if any. Note that there are no restrictions in the methods or in the models to be used to define the deterministic component.
- Analyzing both the spatial distribution and the temporal evolution of the data (or the trend residuals) through a spatio-temporal correlation measure (variogram or covariance function).

This paper provides an introduction to the spatio-temporal random function (*STRF*) theory, together with modeling tools and prediction methods based on kriging. Note that the terms random function and random field are equivalent, although this last terminology is less common.

Space-Time Domain

The natural domain for a geostatistical space-time model is $\mathbb{R}^d \times \mathbb{R}$, where $\mathbb{R}^d$ stands for the d -dimensional spatial domain ($d \in \mathbb{N}_+$) and $\mathbb{R}$ for the temporal domain. From a purely mathematical perspective, $\mathbb{R}^d \times \mathbb{R} = \mathbb{R}^{d+1}$, thus there are no differences between the coordinates. However, physically there is a clear-cut separation between the spatial and time dimensions, and a realistic statistical model will take account of it. It is worth to underline that all technical results on spatial covariance functions or on least-squares prediction, or kriging, in Euclidean spaces apply directly to space-time problems, simply by separating the vector $\mathbf{u}$ (representing a point in space-time), into its spatial and temporal components, that is, $\mathbf{s}$ and t , respectively.

Other spatio-temporal domains are also relevant in practice even if they are not used hereafter. Monitoring data are frequently observed at fixed temporal lags, and it might be useful to model a random function on $\mathbb{R}^d \times \mathbb{Z}$, with time considered discrete. In atmospheric and geophysical applications, the spatial domain of interest is frequently expansive or global and the curvature of the Earth has to be considered; thus, the spatio-temporal domain is defined as $\mathbb{S} \times \mathbb{R}$ or $\mathbb{S} \times \mathbb{Z}$, where $\mathbb{S}$ denotes a sphere in the three-dimensional space.

Problems in Space-Time

Although spatio-temporal estimation techniques can be viewed as an extension of the spatial ones or, alternatively, of the temporal ones, a number of theoretical and practical problems must be addressed. These problems are related to the following aspects:

- Differences between spatial and temporal phenomena
- Data characteristics
- Metric in space-time

Among the fundamental differences between spatial and temporal phenomena, it is worth mentioning that, while there is the notion of past, present, and future in time, conversely, in a spatial context, there is in general no ordering; in particular, only directional dependence, known as anisotropy, may be found.

A typical characteristic of the space-time data is related to the data sampling. Usually, a restricted number of survey stations are established in space, while an intense sampling in time is available at each spatial point. A poor monitoring network is often due to the relatively high cost of the sampling devices and the accessibility of sampling sites. On the other hand, long times series can be easily obtained at each spatial location. Thus, the most common data sets in a space-time context are sparse in space and dense in time. As a consequence, the spatial and temporal correlations are estimated with different degrees of reliability and accuracy.

Moreover, temporal periodicity is often associated with non-stationarity in space, and, in many cases, while the temporal periodicity is limited to the first moment, the non-stationarity in space is extended to higher order moments. With respect to time, periodic functions are often combined (with deterministic or stochastic coefficients) in order to describe seasonal components with a certain period (i.e. daily, weekly, monthly cycles according to the features of the examined variable and temporal aggregation) and different amplitudes and phases. In the presence of quasi-periodic temporal behavior with varying periods, amplitudes, and phases, it is common to assume linear or quadratic trends, unless long time series allow to detect and model a more complex functional form for the mean component. As regards the spatial non-stationarity, a wide class of drift functions is represented by polynomials or, in general, by linear combinations of basic functions, such as powers, exponentials, sigmoids, sines, and cosines in the position coordinates. Some difficulties rise in distinguishing the drift from the spatial correlated residuals, then in modeling the same drift. The intrinsic random functions of order k may be used to solve the problem when polynomial drift functions are

reasonable. Problems may also originate from relevant disparity in the variances at different locations. In this case, it is advisable to divide the spatial domain into more homogeneous regions or remove the outliers sites, if any.

One of the best approaches in extending the usual geostatistical tools to space-time domain is to treat time as another dimension. After that, spatial and temporal coordinates can be kept separate or can be used to introduce a metric into the higher dimensional space. However, since the units for space and time are disparate, e.g., meters and hours, one might ask if it makes any sense to define a spatio-temporal metric,

$$d(\mathbf{u}_1, \mathbf{u}_2) = \left[a(x_1 - x_2)^2 + b(y_1 - y_2)^2 + c(t_1 - t_2)^2 \right]^{1/2},$$

with $\mathbf{u}_1 = (x_1, y_1, t_1)$, $\mathbf{u}_2 = (x_2, y_2, t_2)$ where $(x_1, y_1), (x_2, y_2) \in D \subseteq \mathbb{R}^2$, and $t_1, t_2 \in T \subseteq \mathbb{R}$, where D and T are the spatial and temporal domains, respectively.

Basic Theoretical Framework

In the literature, geostatistical methods have been widely used to analyze spatial and, more recently, spatio-temporal processes in nearly all the areas of applied sciences and especially in Earth Sciences, such as Hydrogeology, to study the movement, distribution and management of water; Environmental engineering, to study concentrations of pollutants in different environmental media, water/air/soil; Meteorology for the analysis of variations in atmospheric temperature, density, and moisture contents. In these contexts, geostatistical tools are flexible enough to analyze the spatio-temporal evolution of a wide range of phenomena, where their spatio-temporal pattern presents a systematic structure at the macroscopic level and a random behavior at the microscopic level. For this reason, suitable stochastic models for spatio-temporal processes should be considered, and a theory of *STRFs* is needed in order to proceed with the rigorous mathematical modeling of processes that jointly change in space and time.

The observed values of a spatio-temporal process are considered as a finite realization of a *STRF* $\{Z(\mathbf{s}, t); \mathbf{s} \in D, t \in T\}$, where $D \subseteq \mathbb{R}^d$, $d \leq 3$, is the spatial domain and $T \subseteq \mathbb{R}_+$ is the temporal domain. The *STRF* Z can be decomposed as follows:

$$Z(\mathbf{s}, t) = \mu(\mathbf{s}, t) + Y(\mathbf{s}, t), \quad (1)$$

where $\mu(\mathbf{s}, t) = E[Z(\mathbf{s}, t)]$ represents a macro scale component, which can be nonconstant in space and time, and Y represents the residual stochastic component with $E[Y(\mathbf{s}, t)] = 0$, which describes the random fluctuations, around

μ , on a small scale. The mean, or deterministic component, $\mu(\cdot)$ is defined as a linear combination of linearly independent functions of the space-time coordinates; it is also called drift, while the term trend, erroneously associated to $\mu(\cdot)$, pertains to data not to the *STRF* model. Under second order stationarity hypothesis, the expected value of Z is assumed to be constant over the domain, i.e., $E[Z(\mathbf{s}, t)] = \mu$, and its second order moments, such as the spatio-temporal variogram and the covariance function, depend on the spatio-temporal lag $(\mathbf{h}_s, h_t)$, for any $(\mathbf{s}, t), (\mathbf{s}', t')$ in the spatio-temporal domain, with $\mathbf{h}_s = \mathbf{s} - \mathbf{s}'$ and $h_t = t - t'$. Then, given the hypothesis of second order stationarity, the spatio-temporal variogram and the covariance function are defined, respectively, as follows:

$$\gamma(\mathbf{h}_s, h_t) = 0.5E[Z(\mathbf{s}, t) - Z(\mathbf{s} + \mathbf{h}_s, t + h_t)]^2, \quad (2)$$

$$C(\mathbf{h}_s, h_t) = E[Z(\mathbf{s}, t)Z(\mathbf{s} + \mathbf{h}_s, t + h_t)] - \mu^2. \quad (3)$$

The variogram is a measure of dissimilarity, in the sense that it increases when the distance between $(\mathbf{s}, t)$ and $(\mathbf{s} + \mathbf{h}_s, t + h_t)$ increases; thus, it is usually characterized by small values for short distances in space and time. On the other hand, the covariance function is a measure of similarity, in the sense that it usually decreases when the distance between $(\mathbf{s}, t)$ and $(\mathbf{s} + \mathbf{h}_s, t + h_t)$ increases; thus, it is often characterized by high values for short distances in space and time. Note that the term covariogram is associated with the graphical representation of the covariance function. In a 2D domain, the covariogram is obtained through a Cartesian diagram, with the stationary covariance function on the vertical axis and the distance on the horizontal axis.

A space-time variogram function γ must be conditional negative definite, namely

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j \gamma(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j) \leq 0, \quad (4)$$

for any $n \in \mathbb{N}_+$, any $(\mathbf{s}_i, t_i) \in D \times T$, where the coefficients $a_i \in \mathbb{R}$ respect the condition:

$$\sum_{i=1}^n a_i = 0. \quad (5)$$

The function γ is conditional strictly negative definite if the quadratic form in (4) is strictly less than zero for any $n \in \mathbb{N}_+$, any choice of distinct points $(\mathbf{s}_i, t_i)$, and any $a_i \in \mathbb{R}$, not all zero, under the condition (5).

Similarly the covariance function must be positive definite, namely:

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j C(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j) \geq 0, \quad (6)$$

for any $n \in \mathbb{N}_+$, any $(\mathbf{s}_i, t_i) \in D \times T$ and any $a_i \in \mathbb{R}$. It is easy to show that given a linear combination of spatial-temporal random variables,

$$C = \sum_{i=1}^n a_i Z(\mathbf{s}_i, t_i),$$

the above conditions ensure the variance to be non-negative. The function C is strictly positive definite (SPD) if it is excluded the chance that the quadratic form in (6) is equal to zero for any choice of distinct points $(\mathbf{s}_i, t_i)$, $\forall a_i \in \mathbb{R}$ (with at least one different from zero) and $\forall n \in \mathbb{N}_+$. Note also that strict positive definiteness is desirable, since it ensures the invertibility of the kriging coefficient matrix (De Iaco and Posa 2018).

Given the difficulties to verify the conditions (4) and (6), it is advisable for users to look for the best model among the wide parametric families whose members are known to respect the above-mentioned admissibility conditions. It is worth highlighting that, although a number of studies provide theoretical results in terms of the covariance, the space-time correlation is usually analyzed through the variogram, which is preferred to the covariance for several reasons (Cressie and Grondona 1992).

A space-time geostatistical analysis can be carried out through the following steps:

1. Look for a model of the *STRF* from which data could reasonably be derived
2. Estimate the variogram or the covariogram, as a measure of the spatio-temporal correlation exhibited by the data
3. Fit a suitable model to the estimated variogram or covariogram;
4. Use this last model in the kriging system for spatio-temporal prediction.

Regarding the first point, it is common to use the model for the *STRF* given in (1), which can include a component μ and the stationary residual component Y . In presence of a non-constant component with large-scale variation, this is commonly estimated (e.g., through least squares regression) and removed from the observed data; then the spatio-temporal correlation analysis is conducted on the residuals, which can be reasonably assumed to be a realization of a second order stationary random field. After analyzing and modeling both components of model (1), predictions of the variable under study can be computed by adding the estimated drift to the predicted residuals, obtained through ordinary kriging.

Alternatively, other forms of kriging can be used in case of non-stationarity, as will be clarified in the section dedicated to prediction.

Structural Analysis

Structural analysis is executed on the spatio-temporal realizations of a second-order stationary random function. These realizations correspond directly to the observed values (if the macro scale component can be reasonably supposed constant over the domain) or to the residuals otherwise. The two steps in the structural analysis consist in estimating the space-time variogram or covariogram and fitting a model. Then, geostatistical analysis proceeds with model validation.

Sample Variogram and Covariogram

Under second order stationarity, structural analysis begins with the estimation of the space-time variogram or covariogram of the random function Z .

Given the set $A = \{(\mathbf{s}_i, t_i), i = 1, 2, \dots, n\}$ of data locations in space-time, the variogram can be estimated through the sample space-time variogram $\hat{\gamma}$ as follows:

$$\hat{\gamma}(\mathbf{r}_s, r_t) = \frac{1}{2|L(\mathbf{r}_s, r_t)|} \sum_{L(\mathbf{r}_s, r_t)} [Z(\mathbf{s} + \mathbf{h}_s, t + h_t) - Z(\mathbf{s}, t)]^2, \quad (7)$$

where $|L(\mathbf{r}_s, r_t)|$ is the cardinality of the set.

$L(\mathbf{r}_s, r_t) = \{(\mathbf{s} + \mathbf{h}_s, t + h_t) \in A, (\mathbf{s}, t) \in A : \mathbf{h}_s \in \text{Tot}(\mathbf{r}_s) \text{ and } h_t \in \text{Tot}(r_t)\}$, and $\text{Tot}(\mathbf{r}_s)$, $\text{Tot}(r_t)$ are, respectively, some specified tolerance regions around $\mathbf{r}_s$ and r_t . Similarly, the space-time sample covariogram $\hat{C}$ is defined as follows:

$$\hat{C}(\mathbf{r}_s, r_t) = \frac{1}{|L(\mathbf{r}_s, r_t)|} \sum_{L(\mathbf{r}_s, r_t)} [Z(\mathbf{s} + \mathbf{h}_s, t + h_t) - \mu][Z(\mathbf{s}, t) - \mu], \quad (8)$$

where μ is substituted by the sample mean $\bar{Z}$, if μ is an unknown constant. Note that the estimator of the space-time variogram is not influenced by the expected value μ or its estimator $\bar{Z}$ and is unbiased with respect to γ . It is important to underline that no space-time metric is needed for the computation of both estimators. In fact, the pairs of points separated by $(\mathbf{h}_s, h_t)$ are detected by computing, separately, the purely spatial and purely temporal distances; thus, the pairs of realizations, $z(\mathbf{s}, t)$ and $z(\mathbf{s} + \mathbf{h}_s, t + h_t)$, correspond to the observations at the points that are simultaneously separated by $\mathbf{h}_s$, in the space domain, and h_t , in the time domain.

Models for Variogram or Covariance Function

The second step of structural analysis consists in fitting a theoretical valid model to the sample space-time variogram or covariogram. Nowadays, various classes of space-time covariance functions or variograms are available, such as the metric class of models, the sum model, the product model, and a wide range of families of non-separable space-time models, such as the product-sum and its integrated version, the classes proposed by Cressie-Huang, Gneiting, Ma, and Porcu to cite a few (De Iaco et al. 2013):

- The Cressie-Huang class of models (Cressie and Huang 1999)

$$C(\mathbf{h}_s, h_t) = \int_{\mathbb{R}^d} e^{i\mathbf{h}_s^T \boldsymbol{\omega}} \rho(h_t; \boldsymbol{\omega}) k(\boldsymbol{\omega}) d\boldsymbol{\omega}, \quad (9)$$

where $\rho(\cdot; \boldsymbol{\omega})$ is a continuous integrable correlation function for all $\boldsymbol{\omega} \in \mathbb{R}^d$, and $k(\cdot)$ is a positive function, which is integrable on $\mathbb{R}^d$;

- The Gneiting class of models (Gneiting 2002)

$$C(\mathbf{h}_s, h_t) = \frac{\sigma^2}{[\psi(h_t^2)]^{d/2}} \phi\left(\frac{\|\mathbf{h}_s\|^2}{\psi(h_t^2)}\right), \quad (10)$$

where $\phi(t)$, $t \geq 0$, is a completely monotone function, $\psi(t)$, $t \geq 0$ is a positive function with completely monotone derivative and σ^2 is the variance;

- The Ma class of models (Ma 2002)

$$C(\mathbf{h}_s, h_t) = \sum_{k=0}^{\infty} \{C_S(\mathbf{h}_s)C_T(h_t)\}^k p_k, \quad (11)$$

where $C_S(\cdot)$ and $C_T(\cdot)$ are purely spatial and purely temporal covariance functions on D and T , respectively; $\{p_k, k \in \mathbb{N}\}$ is a discrete probability function;

- The Porcu, Mateu, and Gregori class (Porcu et al. 2007; Mateu et al. 2007)

$$C(\mathbf{h}_s, h_t) = \int_0^\infty \int_0^\infty \exp\left[-\sum_{i=1}^d \psi_i(|h_i|)v_1 - \psi_t(|h_t|)v_2\right] dF(v_1, v_2), \quad (12)$$

where F is a bivariate distribution function ψ_i and ψ_t are Bernstein functions (positive functions on $\mathbb{R}_+$ with completely monotone derivatives), with $\psi_i(0) = \psi_t(0) = 1$;

- The class of integrated models (De Iaco et al. 2002)

$$C(\mathbf{h}_s, h_t) = \int_V \left[k_1 C_S(\mathbf{h}_s; x) C_T(h_t; x) + k_2 C_S(\mathbf{h}_s; x) + k_3 C_T(h_t; x) \right] d\mu(x), \quad (13)$$

where $\mu(x)$ is a positive measure on $U \subseteq \mathbb{R}$, $C_S(\mathbf{h}_s; x)$ and $C_T(h_t; x)$ are covariance functions defined on $D \subseteq \mathbb{R}^d$ and $T \subseteq \mathbb{R}$, respectively, for all $x \in V \subseteq U$, $k_1 > 0$, $k_2, k_3 \geq 0$. If $k_2 = k_3 = 0$, the above class of models is called integrated product models; otherwise, it is called integrated product-sum models. As specified in the following, this class is very flexible, since it is able to describe all types of non-separability.

The choice of an appropriate class of models can be supported by testing the main properties (symmetry/asymmetry, separability/non-separability, type of non-separability) of the spatio-temporal sample variogram/covariance function (Cappello et al. 2018); the related computational aspects were developed in Cappello et al. (2020). Although the selection of a suitable class of models can be based on its geometric features and theoretical properties, in practice, the generalized product-sum model, thanks to its versatility, is largely used in different areas, ranging from environmental sciences to medicine and from ecology to hydrology. This last model, written in terms of variogram, has the following form:

$$\gamma(\mathbf{h}_s, h_t) = \gamma(\mathbf{h}_s, 0) + \gamma(\mathbf{0}, h_t) - k\gamma(\mathbf{h}_s, 0)\gamma(\mathbf{0}, h_t), \quad (14)$$

where $\gamma(\mathbf{h}_s, 0)$ and $\gamma(\mathbf{0}, h_t)$ are the spatial and temporal marginal variograms, respectively (De Iaco et al. 2001). The power of the model lies in the flexibility of the fitting process, which uses the sample marginals and only one parameter, which depends on the global sill (i.e., the space-time variogram upper bound). The parameter k is selected in such a way to ensure that the global sill is fitted. Admissible values for this parameter are dependent on the sill values of the spatial and temporal marginals $\gamma(\mathbf{h}_s, 0)$ and $\gamma(\mathbf{0}, h_t)$, as specified in De Iaco et al. (2001). Moreover, the generalized product-sum model enables modeling processes characterized by different variability along space and time.

Model Validation

After modeling the space-time empirical variogram/covariogram surface, the subsequent step is to evaluate the reliability of the fitted model through the application of spatio-temporal cross-validation and jackknife techniques. Then, if the model performance is satisfactory, the same model can be used to predict the variable under study over

the spatial domain and for future time points by space-time kriging.

Cross-validation allows a comparison to be made between estimated values and observed ones, simply using the sample information: the value measured at a fixed space-time point is temporarily removed, and at the same point, the value of the variable of interest is estimated by using all the other available values and the fitted space-time variogram/covariogram model. This process is repeated for all the sample points. Finally, statistical tools, such as the linear correlation coefficient between the observed values and the estimated ones, can be used to evaluate the goodness of the fitted model. In the jackknife technique, the variogram/covariogram model is used to estimate the variable under study measured at some points, which are usually not coincident with the original sample points. Then, the data available at the points, where the estimation is required, are compared with the estimated values. In other terms, the validation is based on a new data set, not previously used in the structural analysis and thus not considered in the construction of the covariance/variogram model to be validated.

Prediction

Geostatistical spatio-temporal prediction techniques can be viewed as an extension of the spatial ones. Indeed, given the observed data $\{z(\mathbf{u}_i) = z(\mathbf{s}, t), i = 1, \dots, n\}$, the problem is to predict $Z(\mathbf{u}) = Z(\mathbf{s}, t)$ at the unsampled space-time location $\mathbf{u} = (\mathbf{s}, t)$. For this aim, the following class of linear predictors is defined:

$$\hat{Z}(\mathbf{u}) = \sum_{i=1}^n \lambda_i(\mathbf{u}) Z(\mathbf{u}_i). \quad (15)$$

The weights $\lambda_i(\cdot)$, $i = 1, \dots, n$ are obtained, through the kriging methods, by requiring that (15) is unbiased and the mean squared prediction error is minimized. In particular, under the second-order stationarity, one can choose between simple kriging, if the expected value μ is constant and known, and ordinary kriging, if the expected value μ is constant and unknown. In this last case, the following linear system, often called *ordinary kriging system*, is obtained:

$$\begin{cases} \sum_{i=1}^n \lambda_i(\mathbf{u}) \gamma(\mathbf{u}_i - \mathbf{u}_j) - \omega(\mathbf{u}) = \gamma(\mathbf{u}_j - \mathbf{u}), j = 1, \dots, n, \\ \sum_{i=1}^n \lambda_i(\mathbf{u}) = 1, \end{cases} \quad (16)$$

where γ represents the space-time variogram of Z and ω is the Lagrange multiplier. The above system can be also expressed

in terms of a space-time covariance model. Note that the kriging system requires the knowledge of the space-time covariance/variogram model to be solved in terms of the weights $\lambda_i(\cdot)$, $i = 1, \dots, n$.

On the other hand, if the expected value of the random field Z is known, then the estimator of Z at the unsampled space-time location $\mathbf{u} = (\mathbf{s}, t)$ is equivalent to the estimator of the zero-mean random field Y at $\mathbf{u}$ and the weights $\lambda_i(\cdot)$, $i = 1, \dots, n$ are the solutions of the following linear system of n equations, called *simple kriging system*:

$$\sum_{i=1}^n \lambda_i(\mathbf{u}) \gamma(\mathbf{u}_i - \mathbf{u}_j) = \gamma(\mathbf{u}_j - \mathbf{u}), \quad j = 1, \dots, n, \quad (17)$$

This systems has a unique solution if the coefficient matrix (or the variogram/covariogram matrix) is non-singular, and this is guaranteed if and only if the covariance function is strictly positive definite (or the variogram is conditionally strictly negative definite) and if all sample points are distinct. Note that the simple and ordinary kriging are free-distribution predictors and are considered to be the “best” because they minimize the estimation variance among all the linear unbiased estimators. In case of a Gaussian random field, the simple kriging predictor coincides with the conditional expectation $E(Z | Z_1, \dots, Z_n)$, thus it is the best estimator of Z in the mean square sense.

Moreover, if the unknown expected value cannot be assumed to be constant, the drift can be modeled (even by using external variables) and removed, then the above stochastic methods for spatio-temporal prediction can be proposed for the residuals. If the expected value of the random function can be expressed in a polynomial functional form, kriging with a trend model, known as universal kriging, can be applied. Alternatively, the intrinsic random functions of order k , introduced by Matheron (1973), can be also considered; in this case the drift component is interpreted as a deterministic component, modeled with local polynomial forms. In both options, these kriging predictors can be built to have minimum error variance subject to the unbiasedness constraint.

Computational Aspects

Spatio-temporal data analysis can be supported by using various packages, such as the ones in the R programming language. Among these, it is worth mentioning the package space-time (Pebesma 2012) for dealing with spatio-temporal data, the package gstat for geostatistical modelling and interpolation (Gräler et al. 2016), where different choices of spatio-temporal covariance models have been implemented, such as the separable, product-sum, metric, and sum-metric

models. In addition, the R package RandomFields (Schlather et al. 2015) supports kriging, conditional simulation, covariance functions, and maximum likelihood function fitting for a wide range of spatio-temporal covariance models. Padoan and Bevilacqua (2015) built the R package CompRandFld for the analysis of spatial and spatio-temporal Gaussian and binary data, and spatial extremes through composite likelihood inferential methods. A more complete list of R packages related to spatio-temporal topics is available on CRAN Task Views “SpatioTemporal” (Pebesma 2020). Other contributions concern specialized routines and packages which perform spatio-temporal geostatistical analysis (De Cesare et al. 2002). The package covatest can help to cope with the problem of selecting a suitable class of space-time covariance functions for a given data set (Cappello et al. 2020).

Key Research Findings

Major applications of spatio-temporal modeling and kriging continue to be found in Earth Sciences, Engineering, Mining as well as in Environmental Health and Biodiversity. Specific case studies along these lines are too numerous to be mentioned here, and thus interested readers are invited to consult some recent review books (Montero et al. 2015) for more details. Rather we provide a brief overview of a selected number of key research findings with regards to recent contributions on some properties (such as symmetry/asymmetry, separability/non-separability, type of non-separability) of the space-time covariance functions (Gneiting et al. 2007; De Iaco et al. 2019; Cappello et al. 2020), which can support the model selection. Indeed, one might look for the class of models whose properties are consistent with respect to the characteristics of the empirical space-time covariance surface estimated from the data. For this reason, as clarified in De Iaco et al. (2013), finding an answer to the following questions is essential: (a) How do the spatial and/or the temporal variogram/covariance marginals behave at the origin? (b) Does the space-time data set present different variability along space and time? (c) Which kind of spatial anisotropy is required by the data? (d) Is there a space-time interaction? (e) Is the empirical space-time variogram/covariance function fully symmetric?

Anisotropy in Space-Time

A covariance function is said to be anisotropic in space if at least two directional spatial covariance functions differ. In the classical literature, two types of anisotropy are usually defined: the first type is the *geometric anisotropy*, which is associated with an isotropic covariance function where the coordinates are linearly transformed; the second one is the

stratified or zonal anisotropy, which is associated with the sum of covariance models on factor spaces.

Space and time cannot be directly comparable: hence, the definition of isotropy has no meaning for *STRFs*, although several textbooks and papers have tried to provide a definition of isotropy or geometric anisotropy in this context. An attempt to introduce a geometric anisotropy in space-time was given with the metric covariance model, where the spatial and temporal distances, h_s and h_t , are re-scaled through the use of appropriate spatial and temporal ranges, a_s and a_t , so that the terms h_s/a_s and h_t/a_t are a -dimensional. This allows to work with spatial and temporal coordinates with different physical dimensions (and units). In some cases (diffusion and transport processes), this anisotropy might describe rather well the reality, but the main problem is how to reasonably define the scale parameters on the basis of a physical interpretation. It is also worth highlighting that a second order stationary *STRF* Z , defined on $\mathbb{R}^d \times \mathbb{R}$, is said to be *spatially isotropic* in the weak sense, if its covariance function is such that

$$C_{WT}(\mathbf{h}_s, h_t) = C(\|\mathbf{h}_s\|, |h_t|). \quad (18)$$

On the other hand, the zonal anisotropic models can be built as the sum of different covariance models, each defined on different subspaces of the spatio-temporal domain. For example, zonal anisotropic covariance functions in a space-time domain $\mathbb{R}^d \times \mathbb{R}$ can be expressed as sum of spatial and temporal covariance models C_S and C_t as follows:

$$C(\mathbf{h}_s, h_t) = C_S(\mathbf{h}_s) + C_T(h_t),$$

where $\mathbf{h}_s \in \mathbb{R}^d$ and $h_t \in \mathbb{R}$. However, such models can lead to non-invertible kriging matrices, i.e., even though C_S and C_t are each SPD on the respective subspaces, their sum is only positive definite on the higher dimensional space (Myers and Journel 1990). Thus, an alternative zonal anisotropic covariance model can be obtained by summing to a weak isotropic component, as in (18), another component which can be a purely spatial covariance function, i.e.,

$$C(\mathbf{h}_s, h_t) = C_{WT}(\|\mathbf{h}_s\|, |h_t|) + C_S(\mathbf{h}_s).$$

In the spatio-temporal literature, various classes of stationary non-separable covariance functions with spatial anisotropy are available (Porcu et al. 2006).

Separability

The *STRF* Z on $\mathbb{R}^d \times \mathbb{R}$ is said to have a space-time separable covariance, if it can be written as the product of a purely

spatial covariance function C_S and a purely temporal covariance function C_T , defined on $\mathbb{R}^d$ and $\mathbb{R}$, respectively, i.e.,

$$C(\mathbf{h}_s, h_t) = C_S(\mathbf{h}_s)C_T(h_t), \quad (19)$$

or equivalently

$$C(\mathbf{h}_s, h_t) = \frac{C(\mathbf{h}_s, 0)C(0, h_t)}{C(0, 0)}, \quad (20)$$

where $\mathbf{h}_s \in \mathbb{R}^d$ and $h_t \in \mathbb{R}$. This implies that there is no interaction in space-time. It is clear that the above-mentioned spatio-temporal separability is essentially a partial separability. In case of non-separability, another interesting aspect concerns the type of non-separability, that is: *pointwise* or *uniform* and positive or negative (De Iaco and Posa 2013). The type of non-separability can be easily detected by computing, for each lag, the ratio between the empirical space-time covariance and the product of the sample spatial and temporal covariance marginals $\hat{C}(\mathbf{h}_s, 0)$ and $\hat{C}(0, h_t)$, if these last are positive. Indeed, if the sample ratio is overall greater (less) than 1, uniformly positive (negative) non-separable covariance classes are required. On the other hand, if the sample ratio is greater (less) than 1 for some lags and it is less (greater) than 1 for other lags, space-time covariance classes which are non-uniformly non-separable should be considered. Regarding this aspect, a very flexible class of models is the integrated family given in (13), since *pointwise* or *uniform* positive/negative non-separable models can be derived from it.

Symmetry

Given a stationary space-time covariance function, it is fully symmetric if

$$\begin{aligned} C(\mathbf{h}_s, h_t) &= C(-\mathbf{h}_s, h_t) = C(\mathbf{h}_s, -h_t) \\ &= C(-\mathbf{h}_s, -h_t), \quad \forall (\mathbf{h}_s, h_t) \in \mathbb{R}^d \times \mathbb{R}. \end{aligned} \quad (21)$$

Full symmetry can be an unrealistic assumption for space-time covariance functions encountered in many cases when there is a dominant flow direction over time (for example, in atmospheric processes). Note that the definition of *full symmetry* for a space-time covariance function corresponds to the definition of *reflection or axial symmetry* on $\mathbb{R}^2$. Although the two definitions are referred to different domains, in both cases the domains are partitioned into two factor domains (i.e., $\mathbb{R} \times \mathbb{R}$ or $\mathbb{R}^d \times \mathbb{R}$). Taking into account the above definitions, it is important to point out that symmetry properties represent weaker hypotheses than isotropy, since isotropy requires that C is only a function of the norm of the lag vector. Isotropic covariance functions are a subset of symmetric

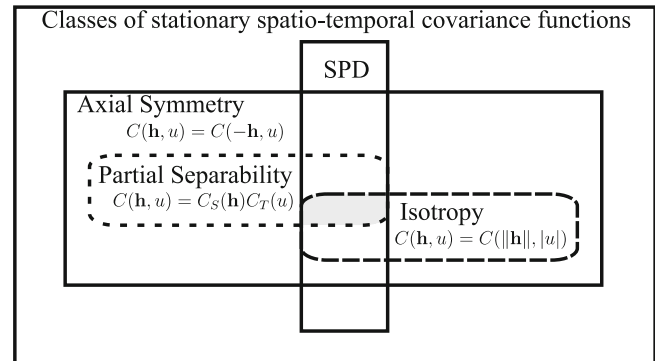
covariance functions, thus methods for testing isotropy can be often used to test symmetry properties. Note that the definitions of strict positive definiteness, symmetry, and separability can be given for *STRFs*, without assuming any form of stationarity. On the other hand, the definition of isotropy is appropriate only for second order stationary spatial random functions on $\mathbb{R}^d$ and can be extended to a space-time context only in a weak sense.

Figure 1 offers a graphical representation on relations among separability, symmetry, isotropy, and SPD condition for stationary random functions in space-time. Assuming full symmetry is equivalent to axial symmetry with respect to the temporal axis, and separability between space and time is equivalent of introducing a partial separability. Regarding isotropy, a spatio-temporal random function is said spatially isotropic in the weak sense, if condition (18) is satisfied. Thus, in a space-time context, it is worth pointing out that isotropy, as in (18), or partial separability, as in (19), implies axial (or full) symmetry, as in (21), the converse is not true. A covariance model can be isotropic and separable if and only if all the factors of the product model are Gaussian covariance functions (gray areas in Fig. 1), as clarified in De Iaco et al. (2020). Strict positive definiteness is also a transverse characteristic among the other mentioned properties (De Iaco and Posa 2018).

Conclusions

At the end, it is worth pointing out that nowadays there are important areas of research that involve random functions in space-time (Porcu et al. 2012) and relevant computational challenges that involve big spatial and spatiotemporal data sets. Some significant lines of research can be found in:

- Multivariate spatio-temporal analysis which includes methods for modeling, predicting and simulating two or



Spatiotemporal, Fig. 1 Relationships among some common second-order properties for spatio-temporal covariance functions (gray areas represent the class of Gaussian covariance functions)

more variables in space and time (Apanasovich and Genton 2010; Genton and Kleiber 2015; Cappello et al. 2022b)

- Complex-valued random functions which have applications in modeling wind fields, sea currents, electromagnetic fields (Posa 2020, 2021; De Iaco 2022; Cappello et al. 2022a)
- Bayesian methods which are gaining popularity both in spatial and spatio-temporal statistics and in physics (Diggle and Ribeiro 2007; Xu et al. 2015; Gelfand and Banerjee 2017)
- Machine learning methods developed by the computational science researchers which can also be applied to spatio-temporal data (Hristopoulos 2015; McKinley and Atkinson 2020; De Iaco et al. 2022)
- Modeling a spatio-temporal random field that has non-stationary covariance structure in both space and time domains (Porcu 2007; Zimmerman and Stein 2010; Shand and Li 2017)
- Parametric covariance modeling techniques for space-time processes, where space is the sphere representing our planet (Jun and Stein 2012; Porcu et al. 2016, 2018)

Most of the modeling advances have been contributing to face interpolation or prediction problems, perform stochastic simulation, and deal with classification issues.

In conclusion, this scientific field has been very active in the last 30 years, as confirmed by the extensive literature and the dedicated sessions in worldwide conferences, and still presents broad margins of progress.

Cross-References

- [Copula in Earth Sciences](#)
- [Kriging](#)
- [Spatial Analysis](#)
- [Spatial Autocorrelation](#)
- [Spatial Statistics](#)
- [Time Series Analysis in the Geosciences](#)
- [Variance](#)

References

- Apanasovich TV, Genton MG (2010) Cross-covariance functions for multivariate random fields based on latent dimensions. *Biometrika* 97(1):15–30
- Cappello C, De Iaco S, Posa D (2018) Testing the type of non-separability and some classes of space-time covariance function models. *Stoch Environ Res Risk Assess* 32:17–35
- Cappello C, De Iaco S, Posa D (2020) Covatest: an R package for selecting a class of space-time covariance functions. *J Stat Softw* 94(1):1–42
- Cappello C, De Iaco S, Maggio S, Posa D (2022a) Modeling spatio-temporal complex covariance functions for vectorial data. *Spat Stat* 47:100562
- Cappello C, De Iaco S, Palma M (2022b) Computational advances for spatio-temporal multivariate environmental models. *Comput Stat* 37:651–670
- Christakos G (2017) *Spatiotemporal random fields*, 2nd edn. Elsevier, Amsterdam
- Cressie N, Grondona M.O (1992) A comparison of variogram estimation with covariogram estimation. In: Mardia, K.V. (ed.) *The art of statistical science*. Wiley, Chichester
- Cressie N, Huang H (1999) Classes of nonseparable, Spatio-temporal stationary covariance functions. *J Am Stat Assoc* 94:1330–1340
- Cressie N, Wikle CL (2011) *Statistics for Spatio-temporal data*. Wiley, New York
- De Cesare L, Myers DE, Posa D (2002) Fortran programs for space-time modeling. *Comput Geosci* 28(2):205–212
- De Iaco S (2022) New spatio-temporal complex covariance functions for vectorial data through positive mixtures. *Stoch Environ Res Risk Assess* 36:2769–2787
- De Iaco S, Posa D (2013) Positive and negative non-separability for space-time covariance models. *J Stat Plan Infer* 143:378–391
- De Iaco S, Posa D (2018) Strict positive definiteness in geostatistics. *Stoch Environ Res Risk Assess* 32:577–590
- De Iaco S, Myers DE, Posa D (2001) Space-time analysis using a general product sum model. *Stat and Probab Letters* 52(1):21–28
- De Iaco S, Myers DE, Posa D (2002) Nonseparable space-time covariance models: some parametric families. *Math Geol* 34(1):23–41
- De Iaco S, Palma M, Posa D (2019) Choosing suitable linear coregionalization models for spatio-temporal data. *Stoch Environ Risk Assess* 33:1419–1434
- De Iaco S, Posa D, Myers DE (2013) Characteristics of some classes of space-time covariance functions. *J Stat Plan Infer* 143(11):2002–2015
- De Iaco S, Hristopoulos DT, Lin G (2022) Special issue: geostatistics and machine learning. *Math Geosci* 54:459–465
- De Iaco S, Posa D, Cappello C, Maggio S (2019) Isotropy, symmetry, separability and strict positive definiteness for covariance functions: a critical review. *Spat Stat* 29:89–108
- De Iaco S, Posa D, Cappello C, Maggio S (2020) On some characteristics of Gaussian covariance functions. *Int Stat Rev* 89(1):36–53
- Diggle PJ, Ribeiro Jr. PJ (2007) *Model-based Geostatistics*, Springer Science+Business Media, New York, NY, USA
- Gelfand AE, Banerjee S (2017) Bayesian Modeling and Analysis of Geostatistical Data. *Annu Rev Stat Appl* 4:245–266
- Genton MG, Kleiber W (2015) Cross-covariance functions for multivariate geostatistics. *Stat Sci* 30(2):147–163
- Gneiting T (2002) Nonseparable, stationary covariance functions for space-time data. *J Am Stat Assoc* 97(458):590–600
- Gneiting T, Genton MG, Guttorp P (2007) Geostatistical space-time models, stationarity, separability and full symmetry. In: Finkenstaedt B, Held L, Isham V (eds) *Statistics of Spatio-temporal systems. Monographs in statistics and applied probability*. Chapman & Hall/CRC Press, Boca Raton, pp 151–175
- Gräler B, Pebesma E, Heuvelink G (2016) Spatio-temporal interpolation using gstat. *The R Journal* 8(1):204–218
- Hristopoulos DT (2015) Stochastic local interaction (SLI) model: Bridging machine learning and geostatistics. *Comput Geosci* 85(Part B):26–37
- Jun M, Stein ML (2012) An Approach to Producing Space-Time Covariance Functions on Spheres. *Technometrics* 49:468–479
- Kyriakidis PC, Journel AG (1999) Geostatistical space-time models: a review. *Math Geol* 31:651–684
- Ma C (2002) Spatio-temporal covariance functions generated by mixtures. *Math Geol* 34(8):965–975
- Mateu J, Porcu E, Gregori P (2007) Recent advances to model anisotropic space-time data. *Stat Method Appl* 17(2):209–223

- McKinley JM, Atkinson PM (2020) A special issue on the importance of geostatistics in the era of data science. *Math Geosci* 52:311–315
- Montero JM, Gema Fernández-Avilés G, Mateu J (2015) *Spatial and Spatio-temporal Geostatistical modeling and kriging*. Wiley, Chichester
- Myers DE, Journel AG (1990) Variograms with zonal anisotropies and non-invertible kriging systems. *Math Geol* 22(7):779–785
- Padoan SA, Bevilacqua M (2015) Analysis of random fields using *CompRandFld*. *J Stat Softw* 63(9):1–27
- Pebesma EJ (2012) *spacetime: Spatio-Temporal Data in R*. *J Stat Softw* 51(7):1–30
- Pebesma EJ (2020) CRAN task view: handling and analyzing Spatio-temporal data. Version 2020-03-18, URL <https://CRAN.R-project.org/view=SpatioTemporal>
- Porcu E (2007) Covariance functions that are stationary or nonstationary in space and stationary in time. *Stat Neerl* 61(3):358–382
- Porcu E, Gregori P, Mateu J (2006) Nonseparable stationary anisotropic space-time covariance functions. *Stoch Environ Res Risk Assess* 21: 113–122
- Porcu E, Gregori P, Mateu J (2007) La descente et la montée étendues: the spatially d-anisotropic and spatio-temporal case. *Stoch Environ Risk Assess* 21(6):683–693
- Porcu E, Montero J, Schlather M (eds) (2012) *Advances and challenges in space-time modelling of natural events*. Lecture notes in statistics, vol 207. Springer, Heidelberg
- Porcu E, Bevilacqua M, Genton MG (2016) Spatio-temporal covariance and cross-covariance functions of the great circle distance on a sphere. *J Am Stat Assoc* 111(514):888–898
- Porcu E, Alegria A, Furrer R (2018) Modeling temporally evolving and spatially globally dependent data. *Int Stat Rev* 86(2):344–377
- Posa D (2020) Parametric families for complex valued covariance functions: some results, an overview and critical aspects. *Spat Stat* 9:100473
- Posa D (2021) Models for the difference of continuous covariance functions. *Stoch Environ Res Risk Assess* 35:1369–1386
- Schlather M, Malinowski A, Menck PJ, Oesting M, Strokorb K (2015) Analysis, simulation and prediction of multivariate random fields with package *RandomFields*. *J Stat Softw* 63(8):1–25
- Shand L, Li B (2017) Modeling nonstationarity in space and time. *Biometric* 73(3):759–768
- Xu G, Liang F, Genton MG (2015) A bayesian spatio-temporal geostatistical model with an auxiliary lattice for large datasets. *Stat Sin* 25(1):61–79
- Zimmerman DL, Stein M (2010) Constructions for nonstationary spatial processes. In: Gelfand AE, Diggle P, Guttorp P, Fuentes M (eds.) *Handbook of Spatial Statistics*, CRC Press, Boca Raton, p. 119–127

Spatiotemporal Analysis

Shrutilipi Bhattacharjee¹, Johannes Madl², Jia Chen² and Varad Kshirsagar³

¹National Institute of Technology Karnataka (NITK), Surathkal, India

²Technical University of Munich, Munich, Germany

³Birla Institute of Technology and Science, Pilani, Hyderabad, India

Synonyms

Remote sensing; Spatiotemporal analysis; Trace gases

Definition

The trace gases are those constituents of the atmosphere that exist in very small amounts, in total approximately around 0.1% of the atmosphere. Many common trace gases, such as water vapor, carbon dioxide (CO₂), methane (CH₄), and ozone (O₃), are the greenhouse gases (GHGs) which are increasing rapidly in Earth's atmosphere due to anthropogenic activity, therefore raising significant attention for its measurement and analysis. On the other hand, nitrogen dioxide (NO₂) is not a GHG but is important for the formation of tropospheric ozone, therefore indirectly contributing to the GHG concentration hike. The main sources of NO₂ are cars, trucks, power plants, and other industrial facilities that are burning fossil fuels (How to Find and Visualize 2021). The gas in high concentration can cause serious issues in the human respiratory system, even exposure over a short time may cause difficulty in breathing, aggravate respiratory diseases, particularly asthma, etc. In this chapter, a brief overview of spatio-temporal analysis of a few atmospheric trace gases is discussed with the inclusion in the recent developments of their remote sensing-based measurements. In particular, the applications of visualization, validation of trace gas measurements, and different analysis methods, such as prediction for missing data, atmospheric transport modeling, inverse modeling, machine learning, etc., are presented. Their spatiotemporal analysis leads to a broad spectrum of applications. Different analysis approaches are discussed in association with the recent investigations of three common atmospheric trace gases, CO₂, CH₄, and NO₂.

Remote Sensing-based Measurement of the Trace Gases

Many satellites are currently observing the atmospheric trace gas concentrations, mainly the GHGs and air pollutants, and measuring their spatiotemporal concentration data across the globe. Examples include the Orbiting Carbon Observatory-2 (OCO-2) (Crisp 2015) from NASA launched in 2014, and its successor OCO-3 (Eldering et al. 2019) which mainly estimates the column-averaged dry-air mole fraction of carbon dioxide (XCO₂). Another complementary example is the Sentinel 5p satellite (Judd et al. 2020), the TROPOMI spectrometer of which estimates many trace gases, including O₃, CH₄, CO, NO₂, sulphur dioxide (SO₂), etc. and aerosols in the atmosphere. These spectrometers on board the satellite are measuring the emitted or reflected radiation from the Earth, to extract different parameters related to the Earth's surface and atmosphere. The satellite retrievals are not the direct measurements; the instruments on satellites are measuring electromagnetic energy radiating from the Earth and atmosphere, which gets converted into geophysical parameters. Due to

great achievements in enhancing measurement techniques and instruments aboard, it is possible to gain better resolutions in time and space of satellite-borne trace gas data. A good example of this can be seen by comparing the Dutch-Finnish Ozone Monitoring Instrument (OMI) on the NASA Aura spacecraft in 2004 having a spatial resolution of $13 \text{ km} \times 24 \text{ km}$ at nadir, and its successor, the Tropospheric Monitoring Instrument (TROPOMI) on the satellite Sentinel 5p, launched by ESA in 2017, having a spatial resolution of $3.5 \text{ km} \times 5.5 \text{ km}$ approximately at nadir (Judd et al. 2020). Besides good improvements in the spatial resolution, the temporal resolution of satellites generally ranges between 1 – 16 days, depending on the orbit, the sensor's characteristics, and the swath width (What is Remote Sensing? 2022). In the case of the OCO-2 satellite, the temporal resolution is 16 days, which means that a huge portion of the Earth is unmeasured in a particular day. Whereas the Sentinel 5P provides the daily global coverage. The other well-known satellites for the trace gas measurements are the Exploratory Satellite for Atmospheric CO₂ (TanSat), Japanese Greenhouse Gases Observing SATellite (GOSAT), Greenhouse Gas Satellite-Demonstrator (GHGSat-D), etc.

Combining all of the desirable features into a single remote sensor is very difficult and there are trade-offs. To acquire observations with high spatial resolution, a narrower swath is required, which requires more time between observations of a given area, causing lower temporal resolution (What is Remote Sensing? 2022). Further, the missing data are preventing a complete understanding of the global cycles of trace gases, mechanisms that control their spatial and temporal variability, and other important features (Bhattacharjee and Chen 2020). In order to get the full picture of the Earth's atmospheric conditions, trace gases, and analyze them for different applications, spatiotemporal methods need to be applied to the collected data. An appropriate spatiotemporal analysis is necessary to learn about the current atmospheric trace gas conditions and to tackle global climate change by developing innovative techniques according to the observed data. A brief overview of the spatiotemporal analysis, including a few applications and methods, applied in the field of remote sensing of trace gases, will be discussed in the next section.

Spatiotemporal Analysis

Many applications and methods of spatiotemporal data analysis have been reported for satellite-borne trace gas data. These can be broadly categorized depending on their research and application objectives. In this entry, mainly three broad analysis approaches for different research objectives are described in brief, the spatiotemporal characteristics by visualization, validation of remotely-sensed trace gas data, as well as important methods of analyzing trace gases. Other research

objectives could be spatiotemporal decision-making, prediction, inversion, etc., hence the abovementioned ones are not exhaustive for the remotely-sensed trace gas data.

Visualization

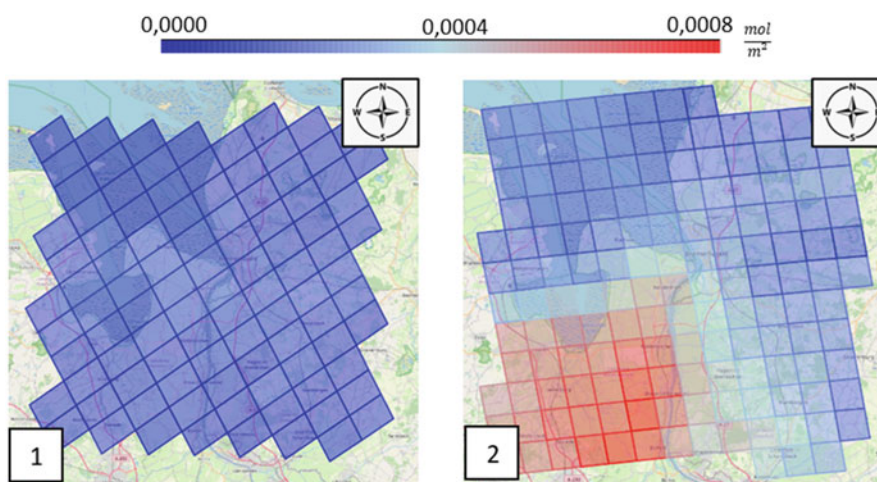
Visualization plays a central role in the process of spatiotemporal analysis of trace gases and can be considered as one of the initial steps for further analysis. Some practical uses of the visualizations are detecting emission point sources of the trace gases, like power plants or tar sand, and monitoring changes in pollution over time due to anthropogenic activities, climate change policies, and regulations. A recent example is the measured reduction of NO₂ in early 2020 with NASA's satellite Aura in 2020 over China. The NO₂ values were 10 – 30% lower than the normal range observed for this time period, which can be partly attributed to the economic slowdown after the coronavirus outbreak (How to Find and Visualize 2021). Any temporal window can be chosen to visualize and compare the trace gas concentrations depending on the need of the applications. For example, a temporal window of 1 day is chosen in Fig. 1 and the NO₂ concentration data of two consecutive days of April 28 and 29, 2021, measured by TROPOMI, are shown here around the harbor in Bremerhaven, Germany. It is clearly seen that there is a burst in the NO₂ concentration in the lower-left corner which is mostly caused by the nearby power plant in Wilhelmshaven and is subjected to further investigations. ESA's TROPOMI is estimating the tropospheric column density of NO₂ in molec/cm². It is not feasible to measure surface NO₂ with the current satellite instruments, but it is possible to estimate tropospheric NO₂ vertical column density, which correlates well with the values on the surface in industrialized regions. Both databases of TROPOMI and OMI NO₂ data can be accessed via Earthdata Search (2021), typically in hierarchical data format. With the freely available tool Panoply from NASA, this data can be easily explored and visualized. Another possible way to explore the NO₂ data is delivered by the application called Giovanni. The data can be viewed as a seasonal or time-averaged map, scatter plot, or as a time-series for the chosen temporal window. Similarly, another tool to visualize the global carbon cycle is the Global Carbon Atlas, established by the Global Carbon Project (Global Carbon Atlas 2021). It provides the estimates of the emission generated from the natural processes and anthropogenic activities and updates the database annually. Pulse GHGSat (2021) is a visualization tool to check global methane concentrations around the world, established by GHGSat. It is updated weekly and shows monthly concentrations, averaged with the resolution of $2 \text{ km} \times 2 \text{ km}$ over land.

Validation of Datasets

After the launch of a satellite, the instruments onboard have to be recalibrated regularly to make sure that they are providing

Spatiotemporal Analysis,

Fig. 1 The two images are illustrated with the measurement of TROPOMI's column-averaged NO_2 on April 28 (LHS image 1) and 29, 2021 (RHS image 2), respectively, in the region around the harbor in Bremerhaven, Germany. The second image shows a burst in the NO_2 measurements in the lower-left corner which is mostly caused by the nearby power plant and subjected to further investigations



accurate measurements. Complete and accurate satellite measurements of the trace gases are heavily influenced by the poor signal-to-noise ratio and dense cloud cover over the observed area, in consequence of which missing/erroneous measurements can be found in some locations. It is not possible to directly measure the near-surface concentrations below the clouds. There are bad weather conditions, faults in the sensors, or several other factors, due to which the uncertainty of the remotely-sensed trace gas measurement can be high. As rightly pointed out by Loew et al. (Loew et al. 2017), the uncertainties present in the satellite data are the known unknown. Further, for the general quality assurance purposes also, validation is an integral part of the satellite measurements. The spatiotemporal validation process of the remotely-sensed trace gases is typically based on a comparison with more reliable ground-based networks. In (Verhoelst et al. 2021), Sentinel 5p TROPOMI's first 2 years' NO_2 measurements are validated against multiple ground-based instruments worldwide, such as Multi-Axis DOAS, NDACC Zenith-Scattered-Light DOAS, and 25 PGN/Pandora instruments. Another example is the Total Column Observing Network (TCCON), which is a global network of ground-based Fourier transform spectrometers. It provides accurate measurements of different trace gases, such as CO_2 , CH_4 , and others (Hardwick and Graven 2016). The instruments of TCCON offer a spectral resolution of 0.02 cm^{-1} and temporal resolution of about 90s. Another global network for GHG measurements was started by Karlsruhe Institute of Technology, Germany, using the EM27/SUN spectrometer, named as COCCON (Frey et al. 2019). A similar ground-based urban GHG network (MUCCnet) was established in 2019 by the Technical University of Munich, Germany, which consists of five fully automated and highly

precise Fourier-transform infrared spectroscopy (FTIR) monitoring systems, distributed in and around the city of Munich (Dietrich et al. 2021). It measures the column concentrations of CO_2 , CH_4 , and CO throughout the year by using the sun as light source. Such local networks are essential in providing more reliable ground-based measurements for local/regional level validation and general calibration applied to the satellite instruments.

Spatiotemporal Analysis Methods

A brief overview of the common methods applied for different applications is presented here. The geostatistical interpolation methods like kriging and many of its variants, such as simple kriging, ordinary kriging, universal kriging, spatial block kriging, fixed rank kriging, etc., can be applied for predicting missing data, that occur due to cloud cover, poor signal due to bad weather conditions, etc. (Bhattacharjee et al. 2013). A recent work (Bhattacharjee and Chen 2020) has analyzed and made use of the trace gas emission estimates, which in turn contributes to their atmospheric concentrations. It has adopted the multivariate interpolation method to predict local column-averaged CO_2 distribution by combining the auxiliary emission estimates and land use/land cover distribution of the study areas. As mentioned before, the prediction results are validated against the TCCON data.

Wind is having a significant influence on the atmospheric transport pattern of the trace gases (Zhao et al. 2019), and these local spatiotemporal processes in turn have a high impact on the atmospheric concentrations of trace gases like CO_2 . Wind develops as a result of spatial differences in atmospheric pressure, which occurs at sea-land transitions due to the uneven absorption of solar radiation (Jacob 1999). The rapidly changing wind speed and wind directions

can only be taken properly into account when the scope of the spatiotemporal analysis changes into a smaller time period. Anthropogenic CO₂ emissions are perturbing the natural balance of CO₂ and causing its atmospheric concentration to rise. The emitted particles of CO₂ get transported by the wind from one place to another, which can change the concentration in a particular location, depending on the wind speed and direction. Therefore, we should take the wind information into account for short-term spatiotemporal analysis. One way to incorporate this information is to use the footprint information, which can be created using the Lagrangian particle dispersion model (LPDM) driven by meteorology models. In combination with surface fluxes, it is possible to determine concentration changes at any region (Bhattacharjee et al. 2020). Similarly, the high-resolution Weather Research and Forecasting model, in combination with GHG modules (WRF-GHG), can model precise meso-scale atmospheric GHG transport, especially in urban areas. In (Zhao et al. 2019), the WRF-GHG model is implemented to simulate the photosynthetic activity of vegetation, emission, and transport of CO₂ and CH₄ in the city of Berlin in 1 km resolution.

Inverse analysis (Jacob et al. 2016) is another important and widely used approach for flux inversion to understand the source and sink distribution at the Earth's surface. The basic idea is to combine global atmospheric transport models with trace gas concentration measurements for estimating the relationship between flux and tracer distributions with different inverse techniques, such as Bayesian inversion, adjoint approach, Markov Chain Monte Carlo (MCMC), etc. The accuracy of the inverse modeling technique depends on the quality of the transport field, the prior flux information used, etc. Many studies have used satellite observations for flux inversion to facilitate modeling very large atmospheric datasets. The inversion technique has been widely applied for CO₂, CH₄, NO₂, and other trace gas measurements from different satellite observations including OCO-2, GOSAT, Sentinel-5P, etc.

Finally, different machine learning approaches are also currently being used for the spatiotemporal analysis of trace gases for different applications. In (Lary et al. 2018), the potential of the machine learning methods for the mandatory bias correction and cross-calibration of the measurement instruments is discussed. Studies are reported to infer the CO₂ flux over land region using an LSTM recurrent neural network using OCO-2 data combined with CO₂ flux tower data (Nguyen and Halem 2018). Machine learning classifiers are being used to identify bad pixels from the satellite measurements (Marchetti et al. 2019), to estimate surface concentration from the vertical column densities (Chan et al. 2021), etc.

Summary or Conclusions

Several aspects of spatiotemporal analysis of trace gases have been discussed, including visualization, validation, and different spatiotemporal analysis methods, such as missing data handling, atmospheric transport modeling, inverse modeling, machine learning methods, etc. Each one of them explores the characteristics of atmospheric trace gases like CO₂, CH₄, NO₂, etc., in different application domains and help to understand the global and local atmospheric processes worldwide. Satellite-borne trace gas data, combined with various ground-based monitoring networks, are the foundation that enables a broad spectrum of their spatiotemporal analysis. Different investigations around the globe have been mentioned here in order to show traditional methods for the spatiotemporal analysis of trace gases and investigate the recent extensions created with data fusion approaches in the future. Though the discussion is not exhaustive, it gives the initial pointers for further exploration.

Cross-References

- [Bayesian Inversion in Geoscience](#)
- [Data Visualization](#)
- [Forward and Inverse Stratigraphic Models](#)
- [Interpolation](#)
- [Inversion Theory](#)
- [Kriging](#)
- [Machine Learning](#)
- [Markov Chain Monte Carlo](#)
- [Remote Sensing](#)
- [Spatial Analysis](#)

Bibliography

- Bhattacharjee S, Chen J (2020) Prediction of satellite-based column CO₂ concentration by combining emission inventory and LULC information. *IEEE Trans Geosci Remote Sens* 58(12):8285–8300
- Bhattacharjee S, Mitra P, Ghosh SK (2013) Spatial interpolation to predict missing attributes in GIS using semantic kriging. *IEEE Trans Geosci Remote Sens* 52(8):4771–4780
- Bhattacharjee S, Chen J, Jindun L, Zhao X (2020) Kriging-based mapping of space-borne CO₂ measurements by combining emission inventory and atmospheric transport modeling. In: EGU General Assembly Conference Abstracts, p 10076
- Chan KL, Khorsandi E, Liu S, Baier F, Valks P (2021) Estimation of surface NO₂ concentrations over Germany from TROPOMI satellite observations using a machine learning method. *Remote Sens* 13(5): 969
- Crisp D (2015) Measuring atmospheric carbon dioxide from space with the Orbiting Carbon Observatory-2 (OCO-2). In: James JB, Xiaoxiong X, Xingfa G (eds) *Earth observing systems xx*, vol 9607. International

- Society for Optics and Photonics, SPIE, pp 1–7. <https://doi.org/10.1117/12.2187291>
- Dietrich F, Chen J, Voggenreiter B, Aigner P, Nachtigall N, Reger B (2021) MUCCnet: Munich Urban Carbon Column network. *Atmos Meas Tech* 14(2):1111–1126
- Earthdata Search. Available at: <https://search.earthdata.nasa.gov/search>. Accessed on: 27 May 2021
- Eldering A, Taylor TE, O'Dell CW, Pavlick R (2019) The OCO-3 mission: measurement objectives and expected performance based on 1 year of simulated data. *Atmos Meas Tech* 12(4):2341–2370
- Frey M, Sha MK, Hase F, Kiel M, Blumenstock T, Harig R, Surawicz G et al (2019) Building the COLlaborative Carbon Column Observing Network (COCCON): long-term stability and ensemble performance of the EM27/SUN Fourier transform spectrometer. *Atmos Meas Tech* 12(3):1513–1530
- Giovanni the bridge between data and science v 4.36. Available at: <https://giovanni.gsfc.nasa.gov/giovanni/>. Accessed on: 24 May 2022
- Global Carbon Atlas. Available at: <http://www.globalcarbonatlas.org/en/content/welcome-carbon-atlas>; Accessed on: 29 May 2021
- Hardwick S, Graven H (2016) Satellite observations to support monitoring of greenhouse gas emissions. Grantham Institute Research Paper No 16. Imperial College London. <https://www.imperial.ac.uk/media/imperial-college/grantham-institute/public/publications/briefing-papers/Satellite-observations-to-support-monitoring-of-greenhouse-gas-emissions-Grantham-BP-16.pdf>. Accessed on: 24 May 2022
- How to Find and Visualize Nitrogen Dioxide Satellite Data. Available at: <https://earthdata.nasa.gov/learn/articles/feature-articles/health-and-air-quality-articles/find-no2-data>. Accessed on: 27 May 2021
- Jacob DJ (1999) Introduction to atmospheric chemistry. Princeton University Press, Princeton
- Jacob DJ, Turner AJ, Maasakkers JD, Sheng J, Sun K, Liu X, Chance K, Aben I, McKeever J, Frankenberg C (2016) Satellite observations of atmospheric methane and their value for quantifying methane emissions. *Atmos Chem Phys* 16(22):14371–14396
- Judd LM, Al-Saadi JA, Szykman JJ, Valin LC, Janz SJ, Kowalewski MG, Eskes HJ et al (2020) Evaluating Sentinel-5P TROPOMI tropospheric NO₂ column densities with airborne and Pandora spectrometers near New York City and Long Island Sound. *Atmos Meas Tech* 13(11):6113–6140
- Lary DJ, Zewdie GK, Liu X, Wu D, Levetin E, Allee RJ, Malakar N et al (2018) Machine learning applications for earth observation. In: *Earth observation open science and innovation*, vol 165. Springer Cham, Switzerland
- Loew A, Bell W, Brocca L, Bulgin CE, Burdanowitz J, Calbet X, Donner RV, Ghent D, Gruber A, Kaminski T, Kinzel J (2017) Validation practices for satellite-based Earth observation data across communities. *Rev Geophys* 55(3):779–817
- Marchetti Y, Rosenberg R, Crisp D (2019) Classification of anomalous pixels in the focal plane arrays of Orbiting Carbon Observatory-2 and-3 via machine learning. *Remote Sens* 11(24):2901
- Nguyen P, Halem M (2018) Prediction of CO₂ flux using Long Short Term Memory (LSTM) Recurrent Neural Networks with data from Flux towers and OCO-2 remote sensing. In *AGU Fall Meeting Abstracts*, vol 2018, pp T31E-0364
- Pulse GHGSat. Available at: <https://ghgsat.com/en/pulse>. Accessed on: 30 May 2021
- Verhoelst T, Compennolle S, Pinardi G, Lambert J-C, Eskes HJ, Eichmann K-U, Fjæraa AM et al (2021) Ground-based validation of the Copernicus Sentinel-5p TROPOMI NO₂ measurements with the NDACC ZSL-DOAS, MAX-DOAS and Pandora global networks. *Atmos Meas Tech* 14(1):481–510
- What is Remote Sensing?. Available at: <https://earthdata.nasa.gov/learn/backgrounders/remote-sensing>. Accessed on: 14 July 2022
- Zhao X, Marshall J, Hachinger S, Gerbig C, Frey M, Hase F, Chen J (2019) Analysis of total column CO₂ and CH₄ measurements in Berlin with WRF-GHG. *Atmos Chem Phys* 19(17):11279–11302

Spatiotemporal Modeling

Shrutilipi Bhattacharjee¹, Johannes Madl², Jia Chen² and Varad Kshirsagar³

¹National Institute of Technology Karnataka (NITK), Surathkal, India

²Technical University of Munich, Munich, Germany

³Birla Institute of Technology and Science, Pilani, Hyderabad, India

Synonyms

Environmental analysis; Spatiotemporal modeling; Outlier detection

Definition

Spatiotemporal modeling covers a broad spectrum of algorithms dealing with spatiotemporal data and a wide range of applications in many fields. The generic idea behind these modeling approaches is to analyze the temporal pattern of target variables within a spatial domain of interest. Therefore, it is needless to say that such modeling approaches are usually complex in nature due to heterogeneous behavior in space and time domains, diverse and voluminous data to deal with, the associated uncertainties, etc. These models can be used for different purposes, such as estimation or prediction, spatiotemporal trend analysis, pattern recognition, and more. In this chapter, we have discussed some applications in the area of environmental data analytics and the spatiotemporal modeling approaches. We further discuss spatiotemporal outlier detection approaches, taking it as a case study, as it is one of the basic data-preprocessing approaches applied in several spatiotemporal analyses. These models can help detect anomalies present in the data which changes over time as well as space, hotspot detection, and understanding their effects on different applications.

Introduction

With the ongoing trend toward big data and growing computational abilities, the prior computational limits of complex spatiotemporal analysis are being narrowed down further. New spatiotemporal approaches are investigated which are upgrading the strong foundations of well-established spatiotemporal analysis methods. The new possibilities are targeting many research areas that can be very different, widely spread, and of current interests, but must address the challenge of spatial data modeling that changes over time. Among many applications of spatiotemporal data modeling,

one significant example can be investigating geospatial processes on our planet Earth.

Geospatial data is a good example of spatiotemporal data. It changes with time as well as space, and some changes even might span centuries. There are many initiatives on a global scale to capture and monitor geospatial data in many forms. Satellite missions have been measuring geospatial environmental variables, for example, TIROS-1, launched by NASA in 1960 (Schnapf 1982), is considered to be the first successful weather satellite. Modeling these data has been an area of interest for a long time, mainly due to the global problem of climate change, global warming, and related subproblems. Geospatial data usually comes with spatial and temporal uncertainties due to the difficulty of capturing it. Therefore, once this data has been captured, it needs to be processed to remove erroneous readings and converted into data products. These data products are then used for modeling to infer implicit patterns. One of the major data-preprocessing stages that could be involved in any modeling processes is the outlier detection.

Outliers can reveal significant information regarding the datasets and the applications. As these are the points that do not follow a specific pattern or are highly isolated from similar data points, investigating the reason behind these points being outliers can lead us to infer important conclusions about the specific problem domain. For example, outliers can reveal information about hotspots, coldspots, or any other point sources of the target variables. Here, the spatiotemporal modeling approaches and its characteristics, one application domain, one modeling approach, i.e., the outlier detection method in general, and a few algorithms for the spatiotemporal outlier detection are discussed.

Spatiotemporal Modeling

Spatiotemporal modeling covers a broad spectrum of approaches and can include different research objectives. It refers to methodological formalization, which will accurately describe the behavior of a certain set of target variables that vary over space and time. The research and application objectives of spatiotemporal modeling can include multiple aspects as follows (Song et al. 2017):

- Description of spatiotemporal characteristics
- Exploration of potential features and spatiotemporal prediction
- Modeling and simulation of spatiotemporal processes
- Spatiotemporal decision-making

Similarly, many models are associated with the abovementioned objectives. For example, models such as spatiotemporal kriging build variogram models that describes

underlying spatiotemporal characteristics of the data. Similarly, Bayesian Maximum Entropy (BME) model (Christakos and Li 1998) can be used for incorporating prior information in case of limited data availability. Various types of regression models, such as spatiotemporal multiple regression, geographically and temporally weighted regression (GTWR), spatiotemporal Bayes hierarchical model (BHM), etc., are used for exploring potential factors and spatiotemporal prediction. On the other hand, popular modeling and simulation models include cellular automaton (CA) and the geographical agent-based model (ABM).

Spatiotemporal decision-making processes use models such as the spatiotemporal multicriteria decision-making model (MCDM).

Applications of Spatiotemporal Modeling

These abovementioned objectives of the spatiotemporal modeling approaches are common for a wide range of applications. For example, the analysis and modeling of meteorological and atmospheric data, which falls under the broad spectrum of environmental applications, have implemented these models for different applications.

In Qin et al. (2017), the geographically and temporally weighted regression (GTWR) model has been considered for predicting ground-level concentration of NO₂ over central-eastern China, based on tropospheric NO₂ columns provided by the Ozone Monitoring Instrument (OMI) combined with ambient monitoring station measurements and meteorological data. The approach showed better or comparable performance with respect to chemistry-based transport models. It is also observed that the people from densely populated areas are more prone to be affected by high NO₂ pollution. Another application for spatiotemporal modeling can be found within the prediction of wildfire propagation and extinction, as shown in Clarke et al. (1994). According to Pechony and Shindell (2010), global fire activity is increasing and there are indications that the future climate will be the deciding factor of global fire trends, compared to the direct human intervention effects on the wildfires. Clarke *et al.* (Clarke et al. 1994) have used a cellular automaton model with additional environmental data, such as wind speed and magnitude, in order to predict the behavior of wildfire and how it will expand. On the other hand, for estimating the sparse column-averaged dry-air mole fraction of carbon dioxide (XCO₂) measured by the Orbiting Carbon Observatory-2 (OCO-2) satellite, kriging-based spatiotemporal and cokriging-based spatial interpolation have been used for generating a Level-3 XCO₂ mapping (Bhattacharjee et al. 2020; Bhattacharjee and Chen 2020). The used method combined the satellite-borne data with auxiliary emission data and land use land cover data, which enhanced the prediction accuracy of XCO₂. Lee

et al. (2018) have proposed an approach using support vector regression for predicting the weather station observations from a spatial perspective. Excluding potential measurement errors within the observations of automatic weather stations is important for increasing the quality of the meteorological observations and for further weather forecasting, disaster warning, or policy formulations. In this manuscript, all abnormal data points will be detected and excluded based on the difference between their actual and estimated values. Marchetti et al. (2019) have presented a machine-learning approach to identify the damaged or unusable pixels in the focal plane array (FPA) of NASA's OCO-2 and Orbiting Carbon Observatory-3 (OCO-3). This task is mainly challenging because of the voluminous data and the diversity of anomalous behavior. Therefore, identifying abnormal factors or data points to increase the accuracy of the measurements or to identify the hotspots is one of the crucial tasks in the field of spatiotemporal modeling. In the next section, the spatiotemporal outlier detection methods and related algorithms are discussed as a representative of widely used methods in spatiotemporal applications.

Spatiotemporal Outlier Detection Method

A spatial outlier is a data point or an object whose nonspatial attribute values are distinctly different from its neighboring data points. A point can be an outlier for a certain neighborhood even if it is typical considering the entire population. Based on these backgrounds, in statistics and data science, the three common categorizations of the outliers (in general) are *global*, *contextual*, and *collective* outliers (Han et al. 2012). When a data point or an object is significantly different compared to the entire sample population, is referred to as the global outlier. On the other hand, the data point deviations given a selected context are the contextual outliers. For example, a high greenhouse gas concentration may be typical for an industrial area, but could be a hotspot for a vegetation cover. The collective outliers are a subset of data points if they significantly deviate from the entire population as a group, but not individually in a global or contextual sense. The same

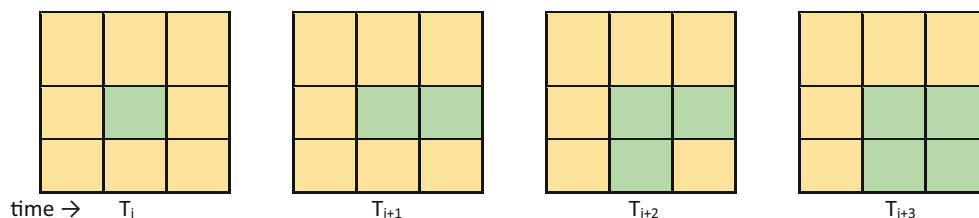
definition can be extended for spatiotemporal outliers. In the case of spatial and spatiotemporal outliers, the difference exists in defining the *neighborhood*. For a spatial outlier, the neighborhood is the set of points within a certain distance of the point under consideration at a single timestamp, whereas, spatiotemporal outliers are data points which exhibit abnormal behavior within a region of interest at some timestamps over a consecutive period (Cheng and Li 2006). Checking and identifying the anomalous points in both spatial and/or time-temporal dimensions makes the spatiotemporal outlier detection approaches more complex compared to purely spatial or temporal outlier detection.

A sample point might seem like an outlier at a certain timestamp (spatial outlier) but may not actually be one in a spatiotemporal context. For example, in the below map of a region with nine pixels, the distribution of a variable *land use land cover* at different timestamps is shown (Anbaroglu 2009). At time T_i , the vegetation (green) pixel in the middle might be an outlier compared to the surrounding built-up area pixels (yellow), and are subjected to further investigations. However, if future timestamps are considered, it is realized that the vegetation is growing and forming a pattern temporally, hence not a spatiotemporal outlier considering four timestamps.

According to Agarwal et al. (2017), the spatiotemporal outliers can be detected in two possible ways:

- Combining the spatial and temporal outliers that are found in two separate processes
- Using both the spatial (for example, latitude, longitude, etc.) and temporal (for example, timestamp) attributes as contextual attributes and find the outliers

The authors have mentioned that the second approach is more *general*, *integrated*, and *flexible* because both spatial and temporal locality can be used to detect the outlier and both the dimensions can be weighed differently based on the application requirement. However, depending on the applications and the datasets used, either of the methods can produce the best results.



Spatiotemporal Modeling, Fig. 1 The change/growing of “vegetation” (green) pixel over the time

Spatiotemporal Outlier Detection Algorithms

The algorithm used to detect spatiotemporal outliers varies from one use case to another as it highly depends on applications, temporal frequency of data collection, number of available timestamps, size of the study region, etc. Some spatiotemporal outlier detection algorithms, used in different literature, will be discussed as follows.

Cheng and Li (Cheng and Li 2006) propose a multiscale technique for spatiotemporal outlier detection. In the preliminary stage of the approach, spatiotemporal objects are classified into clusters using some clustering algorithms adapted to the characteristics of the data. Then, in the second phase, more spatiotemporal objects are aggregated in each cluster by increasing the window or scale of the cluster. An object, which was not classified into any cluster before might be assigned to one now, and objects might change clusters as well. This step is followed by a comparison of the results within the first and the second phases to detect potential spatial outliers. Finally, the potential spatial outliers will be observed over multiple time periods via human observations. If they show consistent behavior over time, they are no longer considered as the outliers. The remaining set of objects is then concluded to be spatiotemporal outliers. This method is more suitable for large geographical data with a very low temporal frequency. As the final phase is carried out by human observations, too many frames or maps will make it difficult to identify patterns or trends. In the manuscript, the authors have used a dataset with a temporal frequency of once per year.

Wu et al. (2008) demonstrate the use of a grid-based approach for outlier detection in the South American precipitation data set obtained from the NOAA. Here, the top k outliers are detected and extracted from each of the time periods, and a sequence of outliers are stored using the *Outstretch* algorithm. At the end, all the possible extracted sequences are identified as the spatiotemporal outliers. Despite the large dataset that has been used during the investigation, the running time of the algorithm seems to be admirable.

Rodgers et al. (2009) propose a statistical method to identify spatiotemporal outliers using the *strangeness-based outlier detection* algorithm (*StrOUD*). The “strangeness” measure of each object is the summation of the distances from all its nearest neighbors. Here, the distance between two points are measured as a weighted distance combining their geographical, temporal, and feature vectors distance. By comparing the strangeness factor of an individual object to the baseline strangeness of the normal objects, spatiotemporal outliers can be detected. If the difference is significant, then the object is said to be an outlier.

When focusing purely on spatial outlier detection algorithms, Chen et al. (2008) have proposed a Mahalanobis distance-based approach. Here, the neighborhood of each of

the data points is defined with k nearest neighbor clustering method, using the spatial attributes, such as latitude and longitude. The generated cluster defines the spatial region to which each data point is compared. All nonspatial attributes within the neighborhood will be summarized with the median value. Based on the Mahalanobis distance between the nonspatial attributes of each data point and its difference from the median, individual outliers can be detected. Once the calculated distance becomes too large and exceeds a predefined threshold, the corresponding data point will be treated as an outlier.

Another approach for detecting spatial outliers within a given data set is the *density-based local outlier factor (LOF)* algorithm is proposed by Breunig et al. (2000). Here, the isolation of a data point with respect to its surrounding neighborhood decides its anomaly score. Hence, the samples within the areas of high density are less likely to be anomalous, since they are validated by a large number of similar data samples. Using domain knowledge or further investigations, outliers can be detected based on their assigned local outlier factors.

In general, for the spatiotemporal applications, the purely spatial outlier detection algorithms, applied for each timestamp independently, might not be very suitable, as the temporal correlation needs to be taken care of by the models. The efficiency of the spatiotemporal models highly depends on the resolutions (spatial and temporal) of the application dataset, their autocorrelation, measurement uncertainty, etc. In some cases, adding additional context of the study region to the process may help to define an outlier more accurately.

Summary or Conclusions

Spatiotemporal dataset has three components in general, *attributes*, *space*, and *time*. Modeling approaches of spatiotemporal data have covered a broad spectrum of applications in many fields, including environmental applications, crime hotspot analysis, healthcare informatics, transportation modeling, social media, and many others. Clustering, predictive learning, frequent pattern mining, anomaly detection, change detection, and relationship mining are the few broad categories of the modeling approaches irrespective of the applications (Atluri et al. 2018). This chapter discusses some modeling approaches used for environmental applications in general. Further, one spatiotemporal modeling approach of outlier detection is chosen and presented here. Outlier detection within the application data is an essential preprocessing step for most of the spatiotemporal applications. Some important literature on spatiotemporal outlier detection methods is also discussed. Though the applications and methods presented here are not exhaustive, this chapter

gives the initial pointers for further exploration of the spatio-temporal models.

Cross-References

- [Bayesian Inversion in Geoscience](#)
- [Data Visualization](#)
- [Forward and Inverse Stratigraphic Models](#)
- [Interpolation](#)
- [Inversion Theory](#)
- [Kriging](#)
- [Machine Learning](#)
- [Markov Chain Monte Carlo](#)
- [Remote Sensing](#)
- [Spatial Analysis](#)
- [Spatiotemporal Analysis](#)

Bibliography

- Aggarwal CC (2017) An introduction to outlier analysis. In: Outlier analysis. Springer, Cham, pp 1–34
- Anbaroğlu TCB (2009) Spatio-temporal outlier detection in environmental data. *Spatial and Temporal Reasoning for Ambient Intelligence Systems I*
- Atluri G, Karpatne A, Kumar V (2018) Spatiotemporal data mining: a survey of problems and methods. *ACM Comput Surveys (CSUR)* 51(4):1–41
- Bhattacharjee S, Chen J (2020) Prediction of satellite-based column CO₂ concentration by combining emission inventory and LULC information. *IEEE Trans Geosci Remote Sens* 58(12):8285–8300
- Bhattacharjee S, Dill K, Chen J (2020) Forecasting interannual space-based CO₂ concentration using geostatistical mapping approach. In: 2020 IEEE international conference on electronics, computing and communication technologies (CONECCT). IEEE, pp 1–6
- Breunig MM, Kriegel H-P, Ng RT, Sander J (2000) LOF: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD international conference on management of data. ACM, pp 93–104
- Chen D, Chang-Tien L, Kou Y, Chen F (2008) On detecting spatial outliers. *GeoInformatica* 12(4):455–475
- Cheng T, Li Z (2006) A multiscale approach for spatio-temporal outlier detection. *Trans GIS* 10(2):253–263
- Christakos G, Li X (1998) Bayesian maximum entropy analysis and mapping: a farewell to kriging estimators? *Math Geol* 30(4):435–462
- Clarke KC, Brass JA, Riggan PJ (1994) A cellular automaton model of wildfire propagation and extinction. *Photogramm Eng Rem S* 60(11):1355–1367
- Han J, Kamber M, Pei J (2012) Outlier detection. In: Data mining: concepts and techniques. Amsterdam, Boston, pp 543–584
- Lee M-K, Moon S-H, Yoon Y, Kim Y-H, Moon B-R (2018) Detecting anomalies in meteorological data using support vector regression. *Adv Meteorol* 2018
- Marchetti Y, Rosenberg R, Crisp D (2019) Classification of anomalous pixels in the focal plane arrays of orbiting carbon observatory-2 and-3 via machine learning. *Remote Sens* 11(24):2901
- Pechony O, Shindell DT (2010) Driving forces of global wildfires over the past millennium and the forthcoming century. *Proc Natl Acad Sci* 107(45):19167–19170
- Qin K, Rao L, Jian X, Bai Y, Zou J, Hao N, Li S, Chao Y (2017) Estimating ground level NO₂ concentrations over Central-Eastern China using a satellite-based geographically and temporally weighted regression model. *Remote Sens* 9(9):950
- Rogers JP, Barbara D, Domeniconi C (2009) Detecting spatiotemporal outliers with kernels and statistical testing. In: 2009 17th international conference on geoinformatics. IEEE, pp 1–6
- Schnapf A (1982) The development of the TIROS global environmental satellite system. *Meteorol Satellites-Past Present Future* 7
- Song Y, Wang X, Tan Y, Peng W, Sutrisna M, Cheng JCP, Hampson K (2017) Trends and opportunities of BIM-GIS integration in the architecture, engineering and construction industry: a review from a spatiotemporal statistical perspective. *ISPRS Int J Geo Inf* 6(12):397
- Wu E, Liu W, Chawla S (2008) Spatiotemporal outlier detection in precipitation data. In: International workshop on knowledge discovery from sensor data. Springer, Berlin/Heidelberg, pp 115–133

Spatiotemporal Weighted Regression

Xiang Que^{1,2}, Xiaogang Ma², Chao Ma³, Fan Liu⁴ and Qiyu Chen⁵

¹Computer and Information College, Fujian Agriculture and Forestry University, Fuzhou, China

²Department of Computer Science, University of Idaho, Moscow, ID, USA

³State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation, Chengdu University of Technology, Chengdu, China

⁴Google, Sunnyvale, CA, USA

⁵School of Computer Science, China University of Geosciences, Wuhan, China

Definition

Spatiotemporal weighted regression (STWR) (Que et al. 2020) is a new time dimension extended GWR-based model for analyzing local nonstationarity in space and time. Different from geographically and temporally weighted regression (GTWR) (Huang et al. 2010; Fotheringham et al. 2015), a novel concept of time distance, which is the rate of value difference between regression and observed points rather than the time interval, is proposed for weighting the local temporal effects. STWR utilizes a new designed weighted average form of spatiotemporal kernel for combining the spatial weights and temporal weights.

Introduction

Geographically weighted regression (GWR) (Brunsdon et al. 1996; Fotheringham et al. 2003) is a useful tool for exploring the potential spatial heterogeneity in geospatial processes. It is

widely applied in many geoscience-related fields, such as climate science (Brown et al. 2012), geology (Atkinson et al. 2003), and environmental science (Mennis and Jordan 2005).

In addition to spatial heterogeneity, temporal heterogeneity also exists widely in the natural and geospatial process. Although the geographically and temporally weighted regression (GTWR) (Huang et al. 2010) can incorporate the time dimension into GWR model, the fitting and prediction performances of GTWR usually do not outperform the basic GWR model. It is probably because of the design and configuration of the spatiotemporal kernel, which might limit the performance of spatiotemporal weight matrix in GTWR.

The GTWR model (Huang et al. 2010) takes time as a new distance dimension, combining with the spatial dimension to calculate the spatiotemporal (Euclidean) distance, and then obtains the weight through the Gaussian or bi-square kernel. However, time and space are at different scales and dimensions, and their units, scales, and impacts on regression points are fundamentally different. Although GTWR uses the adjustment factor τ to synthesize the distance between time and space, directly integrating the spatial distance and time distance is easily misleading (Fotheringham et al. 2015), and the performances of fitting and prediction are sometimes inferior to traditional GWR models. A GTWR (Fotheringham et al. 2015) model uses a set of time-isolated spatial bandwidths to calculate the weights of the influence of observation points on regression points in space and time. But the calibration process is cumbersome and cannot simultaneously optimize time and space bandwidth.

Both GTWR models use the time interval as the time distance, and they use the distance decay kernel function to calculate the weight. It will cause the observation points of different spatial positions recorded at the same time in the past to have the same temporal impacts on the regression point. However, at these observed points, some values change significantly, while others may be almost unchanged. The magnitude of the value change has a different influence on the regression point. The more significant the value change during the time interval, the higher the impact it is. The rate of change of attribute values (nonstationarity in time) also has heterogeneity in spatial position. Using the time interval as the time distance to calculate the weight cannot fully capture the local spatiotemporal effects from observations to the regression point.

Besides, the spatiotemporal kernel of both GTWR models takes the form of multiplying the time kernel and the space kernel function (whose value range of the kernel function is 0–1). However, the multiple forms may cause the problem of time and space interaction, supposing the temporal weight and the spatial weight of an observation point on the regression point is equal to 0.9, while the temporal weight to 0.1. If multiple forms are adopted, the weight value of the

observation point on the regression point will eventually change by 0.09. The weight value might be improper for the spatiotemporal effects. Supposed a house B (observation point) is very close to a house A (regression point) in space. And the house price of B is observed in a relatively long-time interval from now (not exceeding the optimized temporal bandwidth). The multiplying weight values may not correctly reflect the spatiotemporal effect of the observation house price at B on the current house price at the regression point A. If the rates of house prices near A change fast, the past price of B will still have substantial impacts on the current price at A. But if adopted the time interval as the time distance, the composited weight value will be seriously underestimated. How to establish a spatiotemporal kernel to capture the actual weights of observation points to the current regression points becomes a problem worthy study.

Unlike both geographical and temporal weighted regression (GTWR) (Huang et al. 2010; Fotheringham et al. 2015) mentioned above, STWR (Que et al. 2020) proposed a novel time-distance decay weighting method, in which the time distance is the rate of value difference through a time interval rather than the time interval between an observed point and a regression point. Based on the temporal weighting method (temporal kernel), the degree of temporal impact from each observed point to a regression point can be measured, which is a highlighted feature in STWR. Besides, a weighted average form of the spatiotemporal kernel, instead of multiplication form in GTWR (Huang et al. 2010; Fotheringham et al. 2015), was proposed in STWR, which can better capture the combined effects of a nonstationary spatiotemporal process from observed data than GTWR.

These characteristics enable STWR to significantly outperform the GWR and GTWR in model fitting and prediction for the latest observed time stage. Thus, STWR is a practical tool for analyzing the local nonstationary in spatiotemporal process.

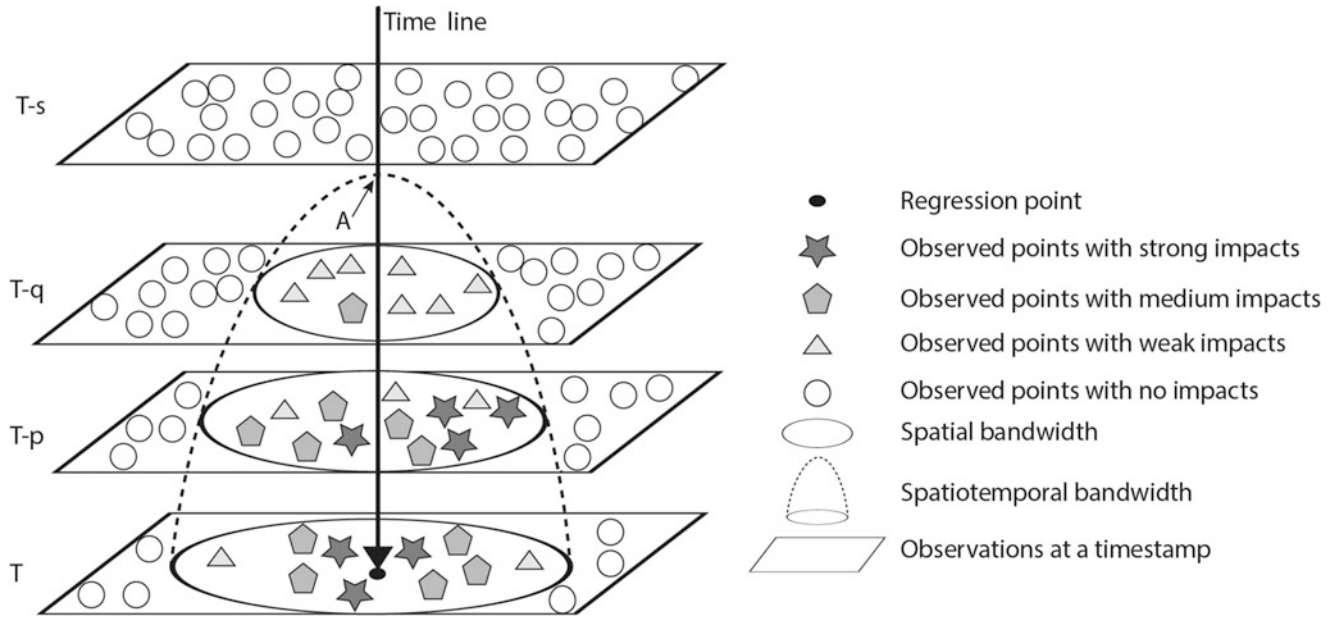
Model Formulation

Principle of GWR

GWR is the background of STWR; it is helpful to introduce the basic framework of GWR (Brunsdon et al. 1996; Fotheringham et al. 1996, 2003). The basic formulation of GWR is described in Eq. 1:

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i)x_{ik} + \varepsilon_i \quad (1)$$

In Eq. 1, (u_i, v_i) is the coordinate of a point i , and y_i is a response variable of the point i . x_{ik} is the k^{th} dependent variable, and ε_i denotes the error term, which is assumed to be independent and drawn identically from the normal



Spatiotemporal Weighted Regression, Fig. 1 Spatiotemporal impacts of observed points with different rates of value change on a regression point at time stage T. Temporal bandwidth is the length of

time from the intersection point A of the spatiotemporal bandwidth and the timeline to the regression point. Spatial bandwidth and spatiotemporal bandwidth are illustrated in the figure legend (Que et al. 2020)

distribution $N(0, \sigma^2)$, i.e., i.i.d. The main difference between GWR and the general global regression method, such as ordinary least squares (OLS), is that GWR allows the coefficient $\beta_k(u_i, v_i)$ vary spatially to identify spatial heterogeneity. The estimated coefficient $\hat{\beta}_k(u_i, v_i)$ can be expressed by Eq. 2:

$$\hat{\beta}_k(u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i) y \quad (2)$$

In Eq. 2, $W(u_i, v_i)$ denotes a diagonal matrix, whose diagonal elements represent the influence of each observation point on the regression point i . The weight values of the diagonal elements are obtained by the bi-square or Gaussian kernel function with inputting their distances. An optimal spatial bandwidth can be optimized through minimizing the errors of the regression model.

After introducing the time dimension, to calibrate the regression point, the model not only needs to “borrow points” from the nearby spatial location but also “borrow points” from the near past time to capture local spatial and temporal effects. To obtain the weight matrix $W(u_i, v_i)$, we need a spatiotemporal kernel and to optimize the temporal bandwidth and spatial bandwidth.

Time-Distance Decay Weighting Strategy Based on the Rate of Value Difference between Regression and Observed Points

Like the spatial distance decay weighting strategy, the time-distance decay weighting strategy is discussed in the previous GTWR model (Crespo et al. 2007; Huang et al. 2010; Wu et al. 2014; Fotheringham et al. 2015). Supposed T-s, T-q, T-p,

and T are the four-time stages from the past to present (Fig. 1). STWR considers the effects of rate of value difference between different observation points and the regression point. Although some data points are farther away from the regression point in space, they may have more significant influences on the regression point (because of the different temporal effects among observed points). As shown at the T-p time stage in Fig. 1, some star points are farther from the regression point in space than some pentagonal points. STWR can better capture the mixed effects of time and space. If the time of the observation point is too old or the rate of numerical difference is too low, that is, exceeding the time-bandwidth, then the weight of temporal effects from the observation point to the regression point is 0.

Spatiotemporal Kernel

Suppose a set of observation points $O_{\Delta t} = \{O_{N_t}, O_{N_{t-1}}, \dots, O_{N_{t-q}} | \Delta t = [t-q, t]\}$ is collected within a certain time interval Δt , where t denotes the current time stage and $N_t - i, i \in \{0, 1, 2, \dots, q\} (N_t = N_t - 0)$ is the number of points observed at each time stage. In STWR, the spatiotemporal takes the weighted average form:

$$w_{ijST}^t = (1 - \alpha)k_s(d_{sij}, b_{ST}) + \alpha k_T(d_{tij}, b_T), 0 \leq \alpha \leq 1 \quad (3)$$

In Eq. 3, w_{ijST}^t denotes the spatiotemporal weight on the observed location j at time stage t . k_s and k_T , both ranging from 0 to 1, are the spatial and temporal kernel, respectively. α is an

adjustable parameter used to weigh the strength of effects from local time and space. It can be optimized by using the same search strategy as bandwidth optimization. b_{ST} is the spatial bandwidth b_S at a certain time stage T , and b_T denotes

the temporal bandwidth. d_{sij} and d_{tij} denote the spatial (Euclidean) and temporal distance between the regression point i and an observed data point j , respectively. In STWR, the temporal kernel k_T is specified as below:

$$w'_{ijST} = \begin{cases} \left[\frac{2}{1 + \exp\left(-\frac{|(y_{i(t)} - y_{j(t-q)})/y_{j(t-q)}|}{\Delta t/b_T}\right)} - 1 \right], & \text{if } 0 < \Delta t < b_T \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

In Eq. 4, Δt is the time interval. $y_{i(t)} - y_{j(t-q)}$ denotes the value difference between the observed value of regression point i at t and the value of observation point j at $t - q$. The weight will be set to zero if the time interval Δt is out of the range $(0, b_T)$. Compared with GTWR models, the temporal kernel k_T of STWR can better capture the effects of temporal variation. Thus, it can represent that the faster the rate of value difference is during a certain time interval (within the temporal bandwidth), the bigger weight value it will be.

In the spatiotemporal process context, the optimized spatial bandwidth can also vary over time. Various functions can be specified for describing the variation of spatial bandwidth

over time. For convenience of model calibration, STWR assumed that the relationship is linear:

$$b_{ST} = b_{St} - \tan\theta * \Delta t, \quad -\frac{\pi}{2} < \theta < \frac{\pi}{2} \quad (5)$$

In Eq. 5, $\tan\theta$ denotes the slope of the spatial bandwidth change with time Δt . Combining with spatial kernel k_s , such as the bi-square kernel and Gaussian kernel (Fotheringham et al. 2003), and then substituting them into Eq. 3, two corresponding spatiotemporal kernel functions can be drawn:

$$w'_{ijST} = \begin{cases} \left[(1 - \alpha) * \left[1 - \left(\frac{d_{sij}}{b_{St} - \tan\theta * \Delta t} \right)^2 \right]^2 + \alpha * \left(2 / \left(1 + \exp\left(-\frac{|(y_{i(t)} - y_{j(t-q)})/y_{j(t-q)}|}{\Delta t/b_T}\right) \right) - 1 \right) \right], & \text{if } \Delta t < b_T, \text{ and } d_{sij} < (b_{St} - \tan\theta * \Delta t) \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

$$w'_{ijST} = \begin{cases} \left[(1 - \alpha) * \exp\left[-\frac{1}{2} \left(\frac{d_{sij}}{b_{St} - \tan\theta * \Delta t} \right)^2\right] + \alpha * \left(2 / \left(1 + \exp\left(-\frac{|(y_{i(t)} - y_{j(t-q)})/y_{j(t-q)}|}{\Delta t/b_T}\right) \right) - 1 \right) \right], & \text{if } \Delta t < b_T, \text{ and } d_{sij} < (b_{St} - \tan\theta * \Delta t) \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

Equations 8 and 9 are based on the bi-square and Gaussian spatial kernel, respectively. b_{St} is the initial spatial bandwidth at the given time stage t of the regression point (i.e., t is the

initial time for searching observed points in the past). The procedure of optimizing the initial bandwidth is the same as GWR model for observations at t . The STWR model needs to

optimize the parameters α and θ , given the initial spatial bandwidth at t , instead of the spatial bandwidth b_{ST} at every time stages.

Model Calibration

Compared with the general GWR, STWR needs to optimize the parameters α and θ as well as the time stage model used (temporal bandwidth). Some goodness-of-fit diagnostics like the cross-validation (CV) score (Cleveland 1979) or the Akaike information criterion (AIC) (Akaike 1973) and its corrected version AICc (Hurvich et al. 1998) can still be used for the calibration of STWR, whose process is listed below:

1. Sort the input data, and get the total number of observation periods N_t and all observation points O_{N_t} of the latest time stage.
2. Traverse the set of $BT_\lambda = \{\Delta t_1, \Delta t_2, \dots, \Delta t_\lambda\}$ (where Δt_λ denotes the time interval between t and $-\lambda$).
3. Obtain the optimized initial spatial width b_{ST}^* by utilizing the GWR model to fit the latest observation points O_{N_t} .
4. Calculate the limitation of the parameter θ with the range of $[0, \arctan \frac{b_{ST}^*}{\Delta t_\lambda}]$ (ensure the spatial bandwidth b_{ST} at each time stages is greater than 0). And set the limitation of searching spatial bandwidth (the number of the nearest neighbors) with the discrete integer set of $[k + 1, N_t]$.
5. Traverse each element in the set BT_λ and then obtain all observation points for each time interval t_i . Further traverse each possible value within the limitation range in (4).
6. Select a searching strategy, such as the equal interval or the golden section method, to figure out model goodness-of-fit value by trying all parameter α and θ available.
7. Record the model diagnostic value (CV value) with its corresponding model parameter in each step of traversal. When the traversal is completed, find the optimal CV value and its corresponding parameter values.
8. Use the optimal parameters to build the spatiotemporal weight matrix, and then estimate the coefficient $\hat{\beta}(u_i, v_i)$ based on Eq. 2.

Case Study

To test the performance of different models, we designed a study area of 25×25 grids for simulating a five-time stages spatio-temporal process with four different heterogeneous coefficient surfaces (one of which is a constant term). The initial values of the three independent variables were generated by random normal distribution $x_1^{initial} \sim N(110, 25^2)$, $x_2^{initial} \sim N(150, 15^2)$, and $x_3^{initial} \sim N(80, 10^2)$, respectively. Their variance kept unchanged, but mean values varied over time as the Eq. (8). The independent random error is assumed to be drawn from a normal distribution $Err \sim N(0, 30^2)$:

$$x_i^{t+\Delta t} = x_i^t \pm \eta * \Delta t \quad (8)$$

In Eq. 8, η is the parameter to control the rate of variation, and x_i^t denotes the value of the i th independent variable at time t . $x_i^{t+\Delta t}$ is the updated value of x_i^t during a certain time interval Δt .

In this case, we also assumed that the four coefficient surfaces of the constant term and the three independent variables were known, and they were generated by the following Eqs. 9, 10, 11, and 12, respectively.

$$\beta_0 = 3 \quad (9)$$

$$\beta_1 = 1 + 4 * \sin[1/8\pi(u + v)] \quad (10)$$

$$\beta_2 = 1 + \frac{1}{12}(u + v) \quad (11)$$

$$\beta_3 = 1 + \frac{1}{324} [36 - (6 - u/2)^2] [36 - (6 - v/2)^2] \quad (12)$$

We randomly sample 1000 points from the five-time stages described above. Our target is to evaluate the fitting and prediction performances for the last time stage t_4 . Table 1 presents the modeling results of the global OLS, GWR, GTWR, and STWR. STWR utilized observations from all-time stages, the initial spatial bandwidth is 6, and α and θ are 0.04 and 0, respectively. Although the GTWR (Huang et al. 2010) used the observations in the past time stages, its AICc and sum of squared estimate of errors (SSE) are 2721.11 and 7675859.93, which are larger than the basic GWR model of 2714.14 and 2658224.80, respectively. STWR has the highest

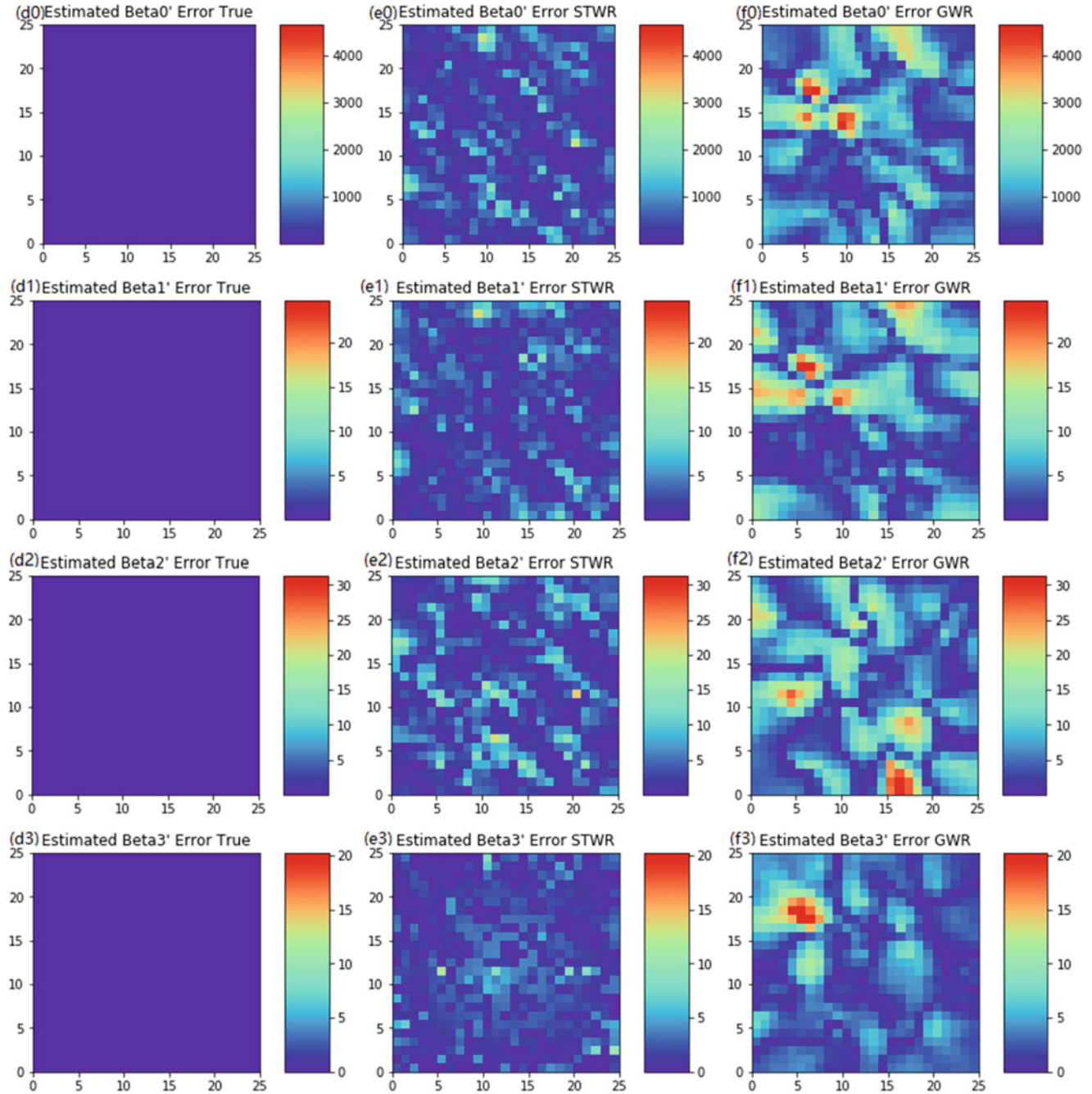
Spatiotemporal Weighted Regression,
Table 1 Comparison of model performances

Time stage t_4	AICc	R2	SSE	Sigma
OLS	2866.32	0.07	18634055.73	
GWR	2714.14	0.87	2658224.80	145.88
GTWR	2721.11	0.62	7675859.93	264.79
STWR	2320.09	0.97	565479.53	24.72

Note: the OLS and GWR model only utilized the observation data at t_4 , that is, because they could not process the temporal effects from observations in the past to the regression points at t_4

R-squared (R^2) 0.97, which is greater than GWR of 0.87 and GTWR of 0.62. Besides, both the AICc and SSE of STWR are the lowest of all these models, which indicates that STWR significantly outperforms other models. Also, the estimated standard error (Sigma) of STWR is only 24.72, which is smaller than 145.88 of GWR and 264.79 of GTWR.

In this case, we also tested the prediction performance of the coefficient surfaces by GWR and STWR model (GTWR add-in package of GTWR does not provide prediction functions). Figure 2 shows the results of absolute predicted errors of the regression coefficient surfaces by GWR and STWR. Four absolute predicted errors of estimated coefficient



Spatiotemporal Weighted Regression, Fig. 2 Comparison of the absolute predicted errors of the regression coefficient surfaces between STWR and GWR models, where (e0), (e1), (e2), and (e3) are four absolute predicted errors of estimated coefficient surfaces (corresponding to β_0 , β_1 , β_2 , and β_4) by STWR

(i.e., $|\hat{\beta}_{0_stwr} - \beta_0|$, $|\hat{\beta}_{1_stwr} - \beta_1|$, $|\hat{\beta}_{2_stwr} - \beta_2|$ and $|\hat{\beta}_{3_stwr} - \beta_3|$). (f0), (f1), (f2), and (f3) are four absolute predicted errors of estimated coefficient surfaces by GWR (i.e., $|\hat{\beta}_{0_gwr} - \beta_0|$, $|\hat{\beta}_{1_gwr} - \beta_1|$, $|\hat{\beta}_{2_gwr} - \beta_2|$ and $|\hat{\beta}_{3_gwr} - \beta_3|$)

surfaces by STWR are much smaller than GWR, which indicates that the STWR outperforms GWR in recovering the coefficient surfaces.

Conclusions

Spatiotemporal weighted regression (STWR) is a novel GWR-based model for exploring local nonstationary spatiotemporal processes in nature and socioeconomics. Value change rate is introduced in the temporal kernel which presents significant model fitting and accuracy in our case study. STWR fully incorporates the observed data in past time and outperforms the geographic temporal weighted regression (GTWR) and geographic weighted regression (GWR) models. The temporal bandwidths and the parameter α and θ in the spatiotemporal kernel of STWR can help to analyze the local effects that form space and time. The model may be appropriate for the local spatiotemporal nonstationary processes in more fields.

Cross-References

- [Geographically Weighted Regression](#)
- [Regression](#)

Bibliography

- Akaike H (1973) Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* 60:255–265
- Atkinson PM, German SE, Sear DA, Clark MJ (2003) Exploring the relations between riverbank erosion and geomorphological controls using geographically weighted logistic regression. *Geogr Anal* 35: 58–82
- Brown S, Versace VL, Laurenson L, Ierodionou D, Fawcett J, Salzman S (2012) Assessment of spatiotemporal varying relationships between rainfall, land cover and surface water area using geographically weighted regression. *Environ Model Assess* 17:241–254
- Brunsdon C, Fotheringham AS, Charlton ME (1996) Geographically weighted regression: a method for exploring spatial nonstationarity. *Geogr Anal* 28(4):281–298. <https://doi.org/10.1111/1467-9884.00145>
- Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc* 74:829–836
- Crespo R, Fotheringham S, Charlton M (2007) Application of geographically weighted regression to a 19-year set of house price data in London to calibrate local hedonic price models. In *Proceedings of the 9th International Conference on Geocomputation*. National University of Ireland Maynooth. https://mural.maynoothuniversity.ie/5816/1/MC_application.pdf
- Fotheringham AS, Brunsdon C, Charlton M (2003) *Geographically weighted regression: the analysis of spatially varying relationships*. Wiley, Hoboken
- Fotheringham AS, Crespo R, Yao J (2015) Geographical and temporal weighted regression (GTWR). *Geogr Anal* 47:431–452
- Huang B, Wu B, Barry M (2010) Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *Int J Geogr Inf Sci* 24:383–401
- Hurvich CM, Simonoff JS, Tsai CL (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J R Statist Soc Series B (Statist Methodol)* 60:271–293
- Mennis JL, Jordan L (2005) The distribution of environmental equity: exploring spatial nonstationarity in multivariate models of air toxic releases. *Ann Assoc Am Geogr* 95:249–268
- Que X, Ma X, Ma C, Chen Q (2020) A spatiotemporal weighted regression model (STWR v1. 0) for analyzing local nonstationarity in space and time. *Geosci Model Dev* 13(12):6149–6164. <https://doi.org/10.5194/gmd-13-6149-2020>
- Wu B, Li R, Huang B (2014) A geographically and temporally weighted autoregressive model with application to housing prices. *Int J Geogr Inf Sci* 28:1186–1204

Spectral Analysis

Katherine L. Silversides

Australian Centre for Field Robotics, Rio Tinto Centre for Mining Automation, The University of Sydney, Sydney, NSW, Australia

Definition

Spectral analysis is a method of transforming sequenced data to extract or filter information. It is frequently used as a preliminary step to simplify further processing. While spectral analysis was initially developed using time series data, it can also be applied to any sequence of data with at least one independent variable. In geoscience this variable is typically distance or depth. The aim is to identify geologically significant short or long term variations from the overall signal. This can be done using either parametric or nonparametric methods.

Introduction

Spectral analysis was initially developed to analyze time-dependent functions and time series, aiming to identify underlying spectral content. This typically involved identifying frequencies within the series that contain significant amplitudes. Nonparametric or parametric methods are applied to a time series, or to any sequential data with at least one independent variable such as spatial coordinates. The analysis separates out spectral content such as waves or oscillations, extracting data or filtering the data to simplify future processing. The signal or data processed by spectral analysis can be either one-dimensional, such as data from a seismograph, or multidimensional, such as seismic reflection surveys (Bath 1974; Buttkus 2000; Jekeli 2017; Stoica and Moses 2005).

Theory

In spectral analysis, the correlation properties of a stationary process $x[n]$ are fully defined by the autocorrelation or Fourier transform, using the power spectral density (PSD). This can be defined as

$$r_x[k] = E[x^*[n]x[n+k]] \leftrightarrow S_x(\omega) = \mathcal{F}\{r_x[k]\}$$

where ω is angular frequency, and $\mathcal{F}$ is the Fourier transform (Spagnolini 2018). As there is only a limited set of N observations, $\mathbf{x} = [x[0], \dots, x[N-1]]^T$, this then becomes:

$$\hat{S}_x(\omega) = \hat{S}_x(\omega|\mathbf{x}).$$

For more complete mathematics, please refer to the Fourier transform and autocorrelation chapters in this book.

Spectral analysis can be applied using either a parametric or non-parametric mathematical model for the PSD. Parametric methods attempt to model the mechanism that generated the data by estimating the correct values of parameters that relate to important characteristics of the physical process. Therefore, information obtained from the model will relate to both the data and the physical processes. Nonparametric methods require no a priori knowledge of the mechanism that generated the data or the correlation between samples. It is a more generic method that has more flexibility (Koopmans 1974; Spagnolini 2018). The method chosen for spectral analysis depends on the type of data to be processed and the desired output, such as defining a cyclical pattern, removing noise, or detecting variations in the data.

Spectral analysis can also be used to filter data, removing noise or unwanted signals to increase the clarity of the desired information. For example, a low-pass filter can be used to remove high frequency data within a signal to remove noise. Other applications produce a high-pass or band-pass filter and the user selects the desired filter depending on the application and the frequency of the useful data that is being extracted (Bath 1974).

Applications

As spectral analysis can be applied to any sequence of data with at least one independent variable, it has a wide range of applications within the geosciences. Three common uses are measuring the seismic moment, seismic reflection surveys, and climate analysis. These are briefly discussed below.

Spectral analysis is often used to investigate the physical properties associated with the sources of natural and induced seismic events (Shemeta and Anderson 2010; Urbancic et al. 1996). Standard magnitude scales for earthquakes (e.g.,

Richter) are empirical and do not provide any information on the physical properties associated with the earthquake. In comparison, the moment magnitude scale is directly related to the slip on the fault and the area of the slip. However, this scale requires the accurate estimate of the seismic moment. While waveform inversion is typically used to estimate the seismic moment for large earthquakes, spectral analysis is frequently used for smaller seismic events. This is typically done using seismogram recording in either the time or frequency domain. In the time domain, the seismic moment can be calculated using the differences in the arrival times of the P- and S-waves, assuming a constant rock density and a known (or calculated) source-receiver distance. This calculation can be affected by the accuracy of the arrival times, the signal to noise ratio, and artifacts in the signal conversion from the seismogram. In the frequency domain, a Fourier transform allows the seismic moment to be directly measured from the amplitude spectra if the low frequencies are sufficiently available. If the low frequencies are not sufficient, a source model spectrum can be fitted to the amplitude spectra instead to estimate the seismic moment (Stork et al. 2014).

Seismic reflection surveys were developed to explore sedimentary basins that have a layered structure and contrasting units. Typically these surveys are used for coal or oil and gas exploration (Taner et al. 1979). The surveys are collected using a seismic source and a series of receivers that collect the signal that is reflected from the subsurface geology. This is done on either a two-dimensional line or a three-dimensional grid and produces a large amount of data that requires both filtering and signal processing. The emitted signal is partially reflected when it encounters geological contacts between two units that have different rock densities and therefore seismic velocities. This may happen multiple times for a single signal, resulting in both a weaker signal and multiple signals received. The time it takes for a signal to reach a receiver depends on both the distance it has travelled and the velocity of the individual rock units. Each individual signal from the receivers needs to be processed using several steps such as filtering, amplitude processing, prestack analysis, deconvolution, and velocity analysis, before the signals can be stacked and processed together. The combined signals are generally filtered using a band-pass method. Most of these steps can be done using spectral analysis and the Fourier transform is commonly used (Gadallah and Fisher 2009).

In climate science, spectral analysis can be used to determine short and long term variations from noisy data. For example, a Fourier transformation into the frequency domain can be used to identify cyclical patterns in a climate system (Mudelsee 2010; Rao and Hsu 2008). This requires estimating the spectral density function and testing for harmonic signals. This typically requires spectral smoothing, and the user must balance the estimation variance and frequency resolution. This could be done using a method such as the

multitaper smoothing method for evenly spaced time series or the Lomb–Scargle method for unevenly spaced data (Mudelsee 2010).

Summary

Spectral analysis uses methods such as the Fourier transform or autocorrelation to analyze time series data and other sequential data with at least one independent variable. The analysis can be used in multiple ways, from extracting waves or oscillations to filtering the data. Parametric methods can be used to obtain information about physical properties, for example, finding the seismic moment from the seismogram record of a small seismic event. Nonparametric methods can be applied more generally, without requiring a priori knowledge. They can be used for filtering and processing signals such as seismic reflections surveys. Therefore, spectral analysis has a wide range of applications to geoscience data and problems.

Cross-References

- [Autocorrelation](#)
- [Fast Fourier Transform](#)

Bibliography

- Bath BM (1974) *Spectral analysis in geophysics*. Elsevier Scientific Publishing Company, Amsterdam, 580p
- Buttkus B (2000) *Spectral analysis and filter theory in applied geophysics*. Springer, Berlin Heidelberg, 667p
- Gadallah MR, Fisher R (2009) *Exploration geophysics*. Springer-Verlag, Berlin/Heidelberg, 262p
- Jekeli C (2017) *Spectral methods in geodesy and geophysics*. CRC Press, Boca Raton, 415p
- Koopmans LH (1974) The spectral analysis of time series. In: Bimbaum ZW, Lukacs E (eds) *Probability and mathematical statistics*, vol 22. Elsevier Science & Technology, New York
- Mudelsee M (2010) *Climate time series analysis: classical statistical and bootstrap methods*. Springer, Dordrecht, 454p
- Rao AR, Hsu EC (2008) *Hilbert–Huang transform analysis of hydrological and environmental time series*. Springer, Dordrecht, 248p
- Schuster GT (2017) Seismic inversion, society of exploration geophysicists, USA, <https://library.seg.org/doi/book/10.1190/1.9781560803423>
- Shemeta J, Anderson P (2010) It's a matter of size: magnitude and moment estimates for microseismic data. *Leading Edge* 29:296–302
- Spagnolini U (2018) *Statistical signal processing in engineering*. Wiley, Hoboken, 562p
- Stoica P, Moses R (2005) *Spectral analysis of signals*. Prentice Hall, Upper Saddle River, 427p
- Stork AL, Verdon JP, Kendall JM (2014) The robustness of seismic moment and magnitudes estimated using spectral analysis. *Geophys Prospect* 62(4):862–878
- Taner MT, Koehler F, Sheriff RE (1979) Complex seismic trace analysis. *Geophysics* 44(6):1041–1063
- Urbancic TI, Trifu C-I, Mercer RA, Feustel AJ, Alexander JAG (1996) Automatic time-domain calculation of source parameters for the analysis of induced seismicity. *Bull Seismol Soc Am* 86:1627–1633

Spectrum-Area Method

Behnam Sadeghi

EarthByte Group, School of Geosciences, University of Sydney, Sydney, NSW, Australia

Earth and Sustainability Research Centre, University of New South Wales, Sydney, NSW, Australia

Definition

Spectrum-area (S-A) fractal model is a different version of concentration-area (C-A) fractal model, in a frequency domain (Cheng et al. 2000; Zuo and Wang 2016), representing the power-law frequency of the power-spectrum density (Afzal et al. 2013). It also has been introduced as the power-spectrum-volume (P-V) fractal model in 3D (Afzal et al. 2012).

Introduction

Benoit Mandelbrot (1983) introduced fractal geometry as a new branch of nonlinear mathematics. He demonstrated the domination of fractal geometry to Euclidean geometry in modeling the nature (see the entry ► [“Fractal Geometry in Geosciences”](#)). Bölviken et al. (1992) studied the importance of fractal geometry to generate geological models such as geochemical landscapes.

As a significant objective in geochemical exploration, geoscientists always look for the highly anomalous mineralization areas to focus the exploration and mining projects on such areas. To do so, they need to classify the areas and discriminate anomalies from the background areas. This is why, during the last decades, a variety of mathematical and statistical classification models have been developed. For further details, see the entry ► [“Singularity Analysis”](#) and Sadeghi 2020, 2021. Therefore, considering the advantages and disadvantages of the developed models and based on Bölviken et al. (1992), Cheng et al. (1994) proposed the first classification fractal model in a spatial domain, named “concentration-area (C-A) fractal model” – see the entry ► [“Concentration-Area Plot”](#). A few years later, Cheng et al. (2000) demonstrated that geochemical patterns can also be studied in the frequency domain. It means the spatial domain in the C-A model can be taken into consideration as

superimposed signals of a variety of frequencies. Based on this theory, the spectrum-area (S-A) fractal model was introduced. This model discriminates geochemical anomalies from the background by spectral analysis and C-A fractal model applied to the frequency domain. In this entry, after introducing the S-A model in detail, its 3D format, called “power-spectrum-volume (P-V)” (Afzal et al. 2012), will also be discussed.

Spectrum Analysis

Spectrum analysis is a well-developed theory that has been applied in exploration geophysics to discriminate anomalies. For instance, it has been widely applied to detect various frequencies' signals (patterns) representing geological areas and their variety. This theory simply demonstrates a couple of frequency and spatial domains per signal. To convert such signals from spatial domain to frequency domain and vice versa, Fourier and inverse Fourier transformations (FT) are applied, respectively, as follows (Cheng et al. 2000):

$$F(K_x, K_y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \cos(K_x x + K_y y) dx dy$$

$$- i \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \sin(K_x x + K_y y) dx dy$$

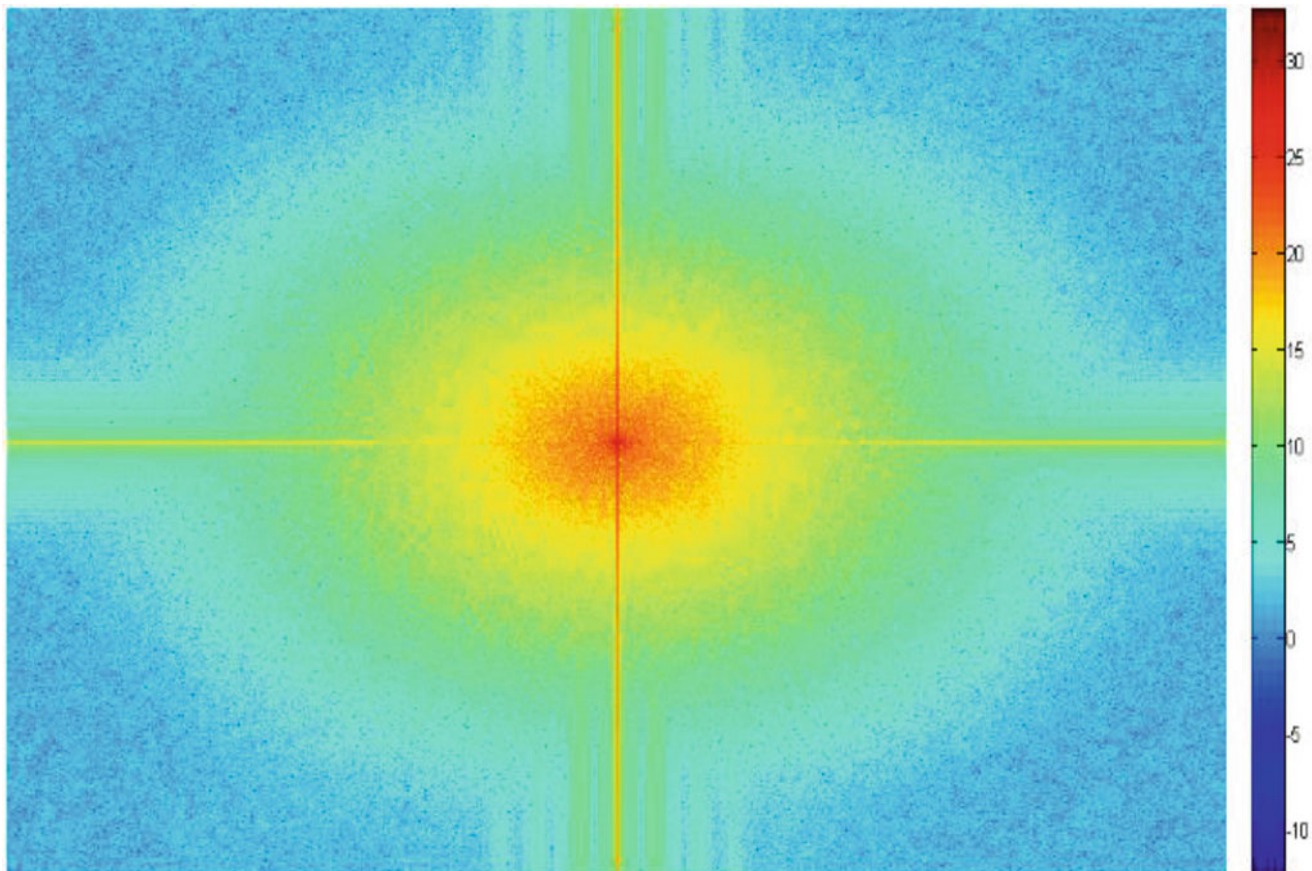
$$F(x, y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(K_x, K_y) \cos(K_x x + K_y y) dx dy$$

$$- \frac{i}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(K_x, K_y) \sin(K_x x + K_y y) dx dy$$

where K_x and K_y represent wave numbers (WN) per x and y axes; this WN is the frequency spatial equivalent (rad/sec); $f(x, y)$ is the signal in spatial domain.

Methodology: Spectrum-Area (S-A) Fractal Model

The S-A fractal model is a power-law relation between the power spectrum (Fig. 1) of S and the occupied area $A(\geq S)$ (Cheng et al. 2000):



Spectrum-Area Method, Fig. 1 Power-spectrum map generated by FT based on the Swedish till data (from Sadeghi 2020)

$$A(\geq S) \propto S^{-2d/\beta}$$

where S is power spectrum, A represents the area covered by a range of geochemical concentrations, β is an anisotropic scaling exponent, and d is the overall degree of concentration. In other words, $2d$ is associated with elliptical dimension ($d = 1$), which is related to isotropic dimensions.

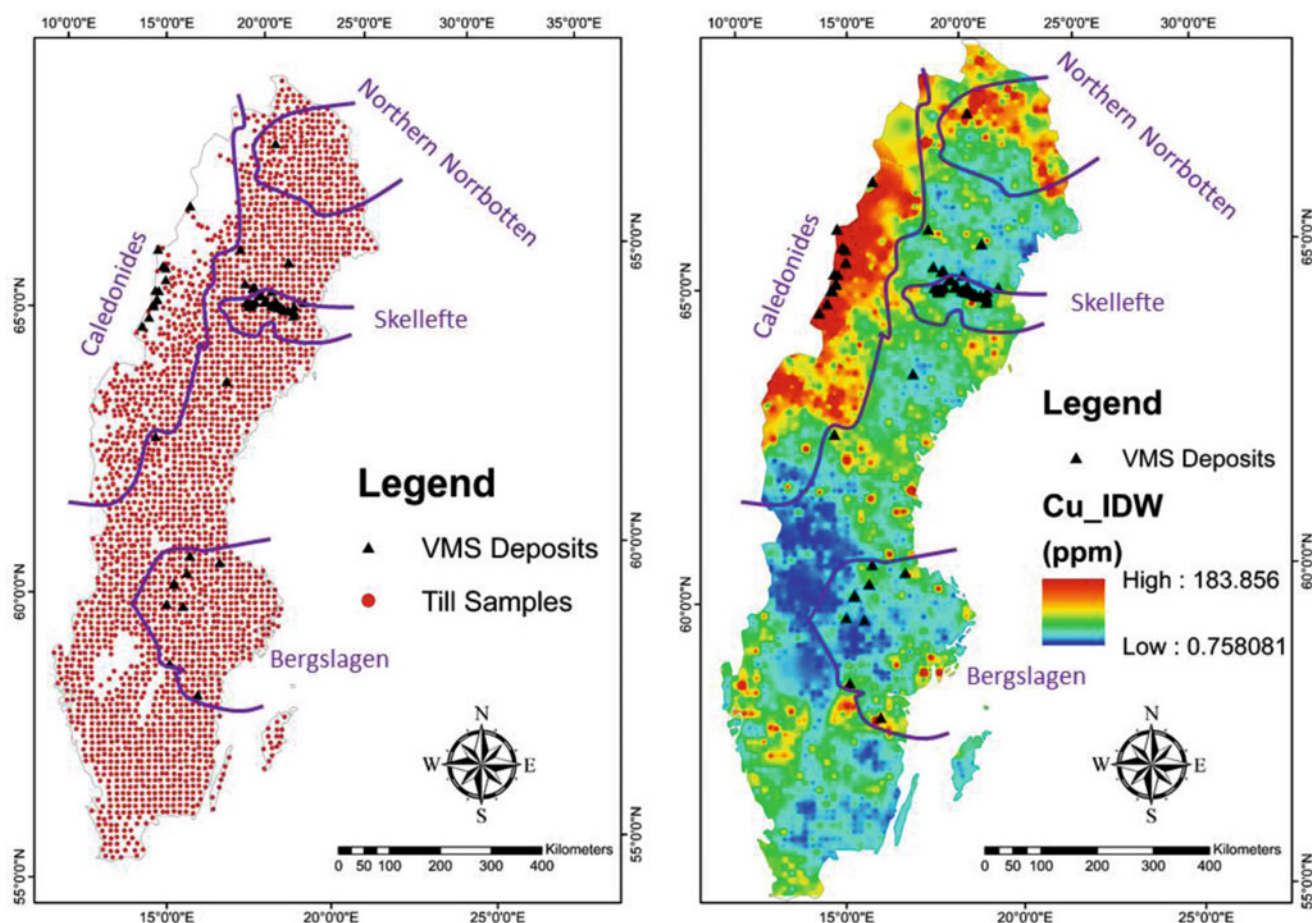
The S-A model applies FT/inverse FT transformation to convert/convert back of the spatial distribution to the frequency one. The reason why in this model frequency distribution is taken into account is because it highly simplifies some operations such as complex correlation analysis and filtering (Afzal et al. 2013).

Like the C-A model, the S-A model provides a log-log plot but based on the relevant power spectrum and area. To define different geochemical patterns, several straight lines are fit to the log-log plot using least-squares (LS) method – see the entry ► “Concentration-Area Plot” for further details. The straight lines’ cross points are the desired thresholds to discriminate “background,” “noise,” and low/high power-

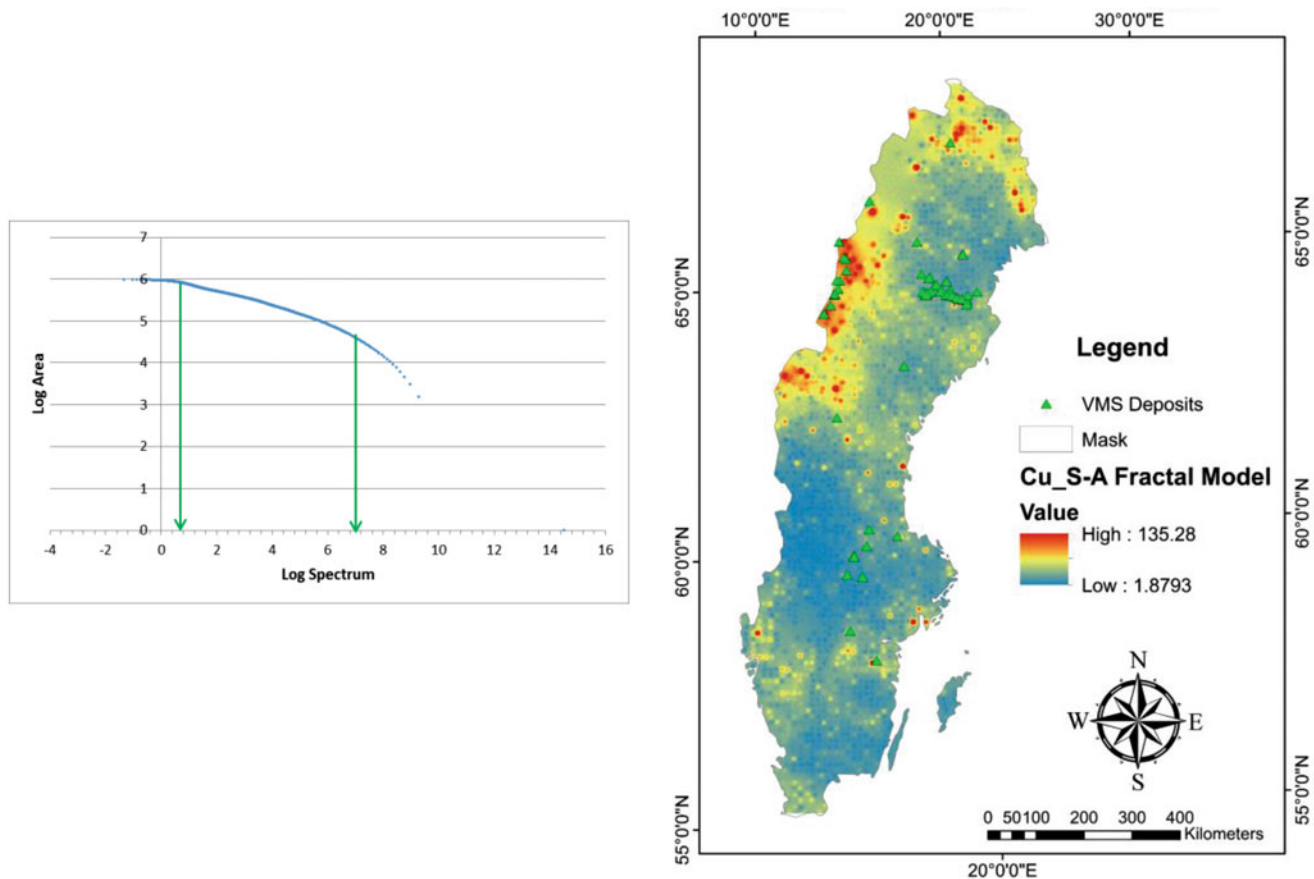
spectrum values. To define the geochemical anomalies, we need to convert the introduced patterns to the spatial domain (Cheng et al. 2000).

Case Study: Sweden

Based on a till geochemical dataset, collected by the Geological Survey of Sweden (Andersson et al. 2014) (Fig. 2), Sadeghi (2020) developed several novel fractal/multifractal models to separate Cu-related volcanic massive sulfide (VMS) geochemical anomalies through uni-, bi-, and multivariate analyses. Several other traditional fractal/multifractal models, such as C-A and S-A models, were also applied to the same data to compare the results and evaluate their significance. In Fig. 3, the Cu geochemical anomalies, the corresponding FT power-spectrum map, and the S-A log-log plot have been demonstrated. On the log-log plot, two thresholds are defined to filter geochemical background (the first population from left), geochemical anomalies (the middle



Spectrum-Area Method, Fig. 2 Till samples collected throughout Sweden and the Cu representative IDW interpolated map (from Sadeghi 2020; Sadeghi and Cohen 2021)



Spectrum-Area Method, Fig. 3 Geochemical anomalies of Cu defined by S-A fractal model based on the Swedish till data (from Sadeghi 2020)

line), and noise (the last population in the right hand). Based on the S-A model in Sweden, the main VMS Cu geochemical anomalies are mainly located in NW and Northern Sweden, which are Caledonides and Northern Norrbotten areas.

Power-Spectrum-Volume (P-V) Fractal Model: The 3D Format of S-A Model

The P-V fractal model was developed by Afzal et al. (2012) in 3D to discriminate various mineralization zones in Kahang Cu porphyry deposit in Central Iran, as a case study (Fig. 4):

$$V(\geq S) \propto S^{-2d/\beta}$$

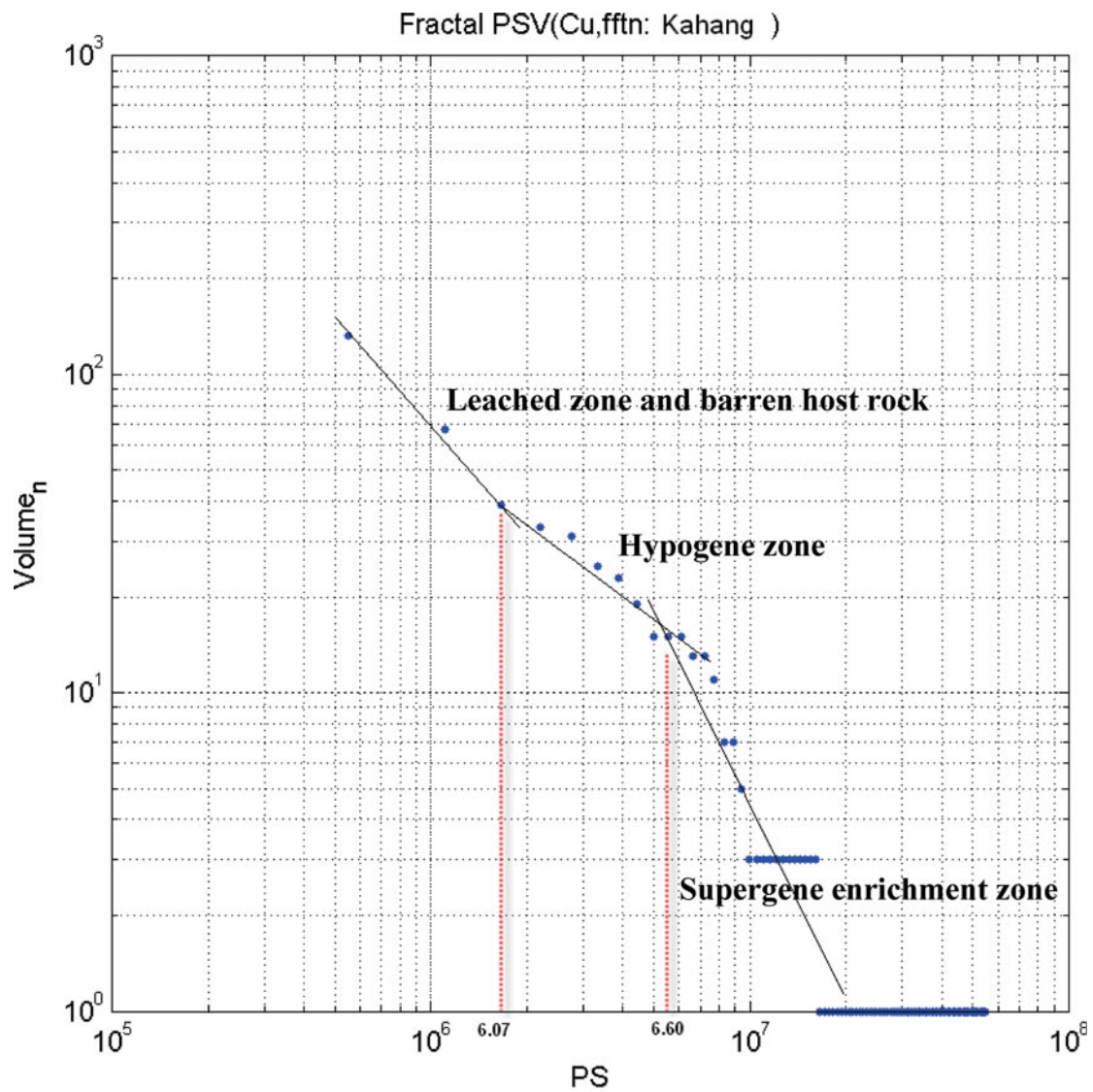
where $V(S)$ denotes the volume occupied by concentrations greater than S . This model is actually based on the relation between power-spectrum values and the volume of the voxels.

Afzal et al. (2012) demonstrated that P-V model is efficient enough to define different mineralization zones. For instance,

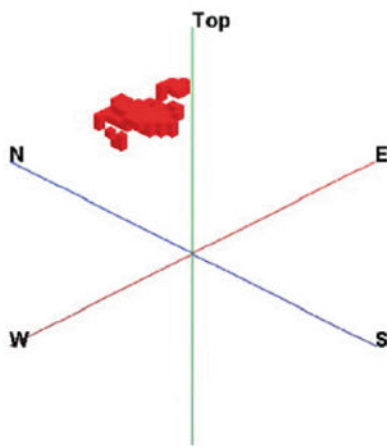
in their study, the P-V model successfully recognized supergene and hypogene enrichment zones and differentiated them from the background or the barren host rock (Fig. 4).

Summary and Conclusions

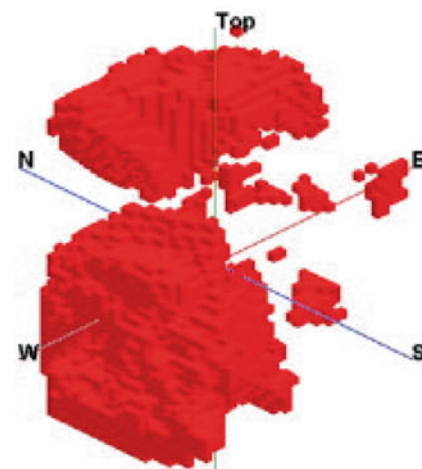
A variety of mathematical and statistical models have been developed to classify geochemical anomalies from the background and identify mineralization zones. Fractal/multifractal models are among the newest and most robust classification models. The C-A fractal model is the first fractal model applied for this sake; then, it was developed in a frequency domain, rather than spatial domain, and introduced as S-A fractal model. This model is based on FT and inverse FT to transform the data from spatial domain to the frequency domain and back. The 3D format of S-A model was introduced later as P-V model, which is also a robust model to characterize mineralization zones in depth.



a



b



c

Spectrum-Area Method, Fig. 4 (a) P-V log-log plot and its defined (b) supergene and (c) hypogene zones in Kahang Cu porphyry deposit (modified from Afzal et al. 2012)

Cross-References

- [Concentration-Area Plot](#)
- [Fractal Geometry in Geosciences](#)
- [Singularity Analysis](#)

Bibliography

- Afzal P, Fadakar Alghalandis Y, Moarefvand P, Rashidnejad Omran N, Asadi Haroni H (2012) Application of power-spectrum-volume fractal method for detecting hypogene, supergene enrichment, leached and barren zones in Kahang Cu porphyry deposit, Central Iran. *J Geochem Explor* 112:131–138
- Afzal P, Harat H, Fadakar Alghalandis Y, Yasrebi AB (2013) Application of spectrum–area fractal model to identify of geochemical anomalies based on soil data in Kahang porphyry-type Cu deposit, Iran. *Geochemistry* 73(4):533–543
- Andersson M, Carlsson M, Ladenberger A, Morris G, Sadeghi M, Uhlbäck J (2014) Geochemical Atlas of Sweden. Geological Survey of Sweden (SGU), Uppsala, 208 p
- Bölviken B, Stokke PR, Feder J, Jössang T (1992) The fractal nature of geochemical landscapes. *J Geochem Explor* 43:91–109
- Cheng Q, Agterberg FP, Ballantyne SB (1994) The separation of geochemical anomalies from background by fractal methods. *J Geochem Explor* 51:109–130
- Cheng Q, Xu Y, Grunsky E (2000) Integrated spatial and spectrum method for geochemical anomaly separation. *Nat Resour Res* 9: 43–52
- Mandelbrot BB (1983) *The fractal geometry of nature*, 2nd edn. New York: WH freeman
- Sadeghi B (2020) Quantification of Uncertainty in Geochemical Anomalies in Mineral Exploration. PhD thesis, University of New South Wales
- Sadeghi B (2021) Concentration-concentration fractal modelling: a novel insight for correlation between variables in response to changes in the underlying controlling geological-geochemical processes. *Ore Geol Rev.* <https://doi.org/10.1016/j.oregeorev.2020.103875>
- Sadeghi B, Cohen D (2021) Category-based fractal modelling: a novel model to integrate the geology into the data for more effective processing and interpretation. *J Geochem Explor.* <https://doi.org/10.1016/j.gexplo.2021.106783>
- Zuo R, Wang J (2016) Fractal/multifractal modeling of geochemical data: a review. *J Geochem Explor* 164:33–41

Splines

Bengt Fornberg
Department of Applied Mathematics, University of Colorado,
Boulder, CO, USA

Definition

A spline is a piecewise polynomial function, commonly used to interpolate or approximate one-dimensional data. At its nodes, it changes from one polynomial to another one, in

such a way that a certain number of derivatives remain continuous. (Typically one derivative less than the degree of the polynomial pieces.)

Introduction

Splines were introduced by Schoenberg in 1946 (Schoenberg 1946). Later generalizations extend to interpolating both gridded and unstructured data in any number of dimensions. General reference books on splines include (de Boor 2001; Dierckx 1993; Powell 1981; Prautzsch et al. 2002).

Interpolation in 1-D, using a single polynomial, is notoriously unstable. Figure 1 contrasts this with piecewise linear and *natural* cubic splines. The first two cases represent extremes, infinitely differentiable but with large oscillations (the Runge phenomenon), vs. not even once differentiable, respectively. Cubic splines provide an attractive compromise. The one shown here minimizes, over every possible interpolating function $s(x)$, the integral $\int s''(x)^2 dx$ over the interval spanned by the outermost nodes. Loosely speaking, it interpolates with the minimal amount of total curvature.

Some Spline Features

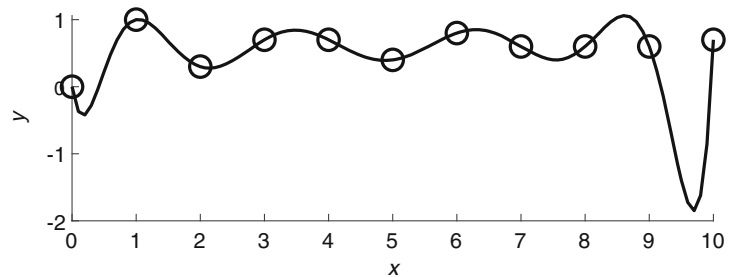
For practical use of splines, most coding environments contain a rich set of convenient library routines (including for differentiation, integration, etc.) We therefore focus here on their main concepts, sometimes in simplified cases.

Spline End Conditions

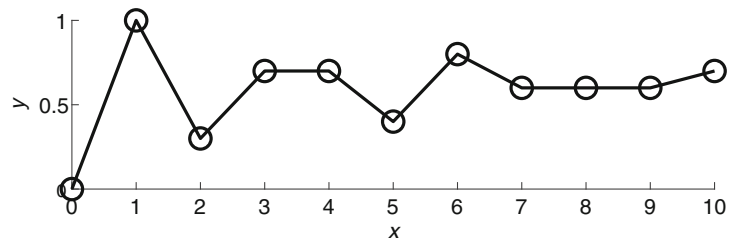
Using between nodes polynomials of degree p (each requiring $p + 1$ coefficients), imposing the function value at each node, and continuity of $p - 1$ derivatives at each interior node is readily seen to leave us short of $p - 1$ pieces of information. We thus need no extra information in the linear ($p = 1$) case (as to be expected for a piecewise linear interpolant), but we are two conditions short in the cubic ($p = 3$) case. Unless specifying otherwise, we focus in the following on this cubic case. Four common options for end conditions are:

1. Set $s'' = 0$ at each end point. This gives the *natural cubic spline*. While of theoretical interest (e.g., minimization of $\int s''(x)^2 dx$, as noted above), the information $s'' = 0$ is typically outright wrong. If nodes are spaced h apart, the error when interpolating a smooth function scales in the interval interior as $O(h^4)$, but the incorrect information reduces this to $O(h^3)$ near to the boundaries.
2. Instead of adding two conditions, one can get the counting right also by taking away two freedoms, such as

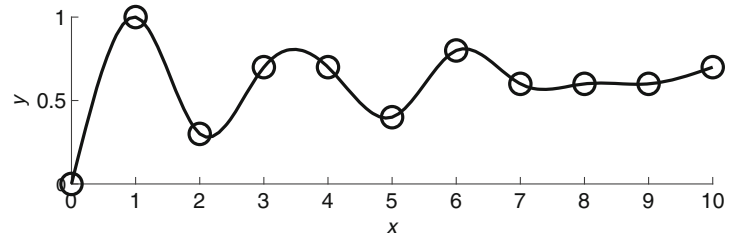
Splines, Fig. 1 Contrasting polynomial interpolation with linear and (natural) cubic splines when interpolating the same data. Note different vertical scale in top plot



(a) Single polynomial of minimal degree



(b) First order spline: piecewise linear



(c) Third order (cubic) spline

disallowing a discontinuity in s''' one step in from each boundary – known as *Not-a-Knot*.

3. Fit the slope at each end point with that of the cubic polynomial that interpolates the four points nearest to each end.
4. If one happens to know s' at each boundary, that can be imposed, giving a *clamped spline*.

B-Splines: Concept

Reminiscent of how Lagrange interpolation combines simpler *cardinal* polynomials (one at one node, and zero at the other nodes), it is natural to seek a “simplest” spline, and then represent a general spline as a linear combination of these. A good such choice is the spline that transitions from identical zero non-trivially back to identical zero in the narrowest way possible; cf., Fig. 2 (The customary normalization is that the integral evaluates to one).

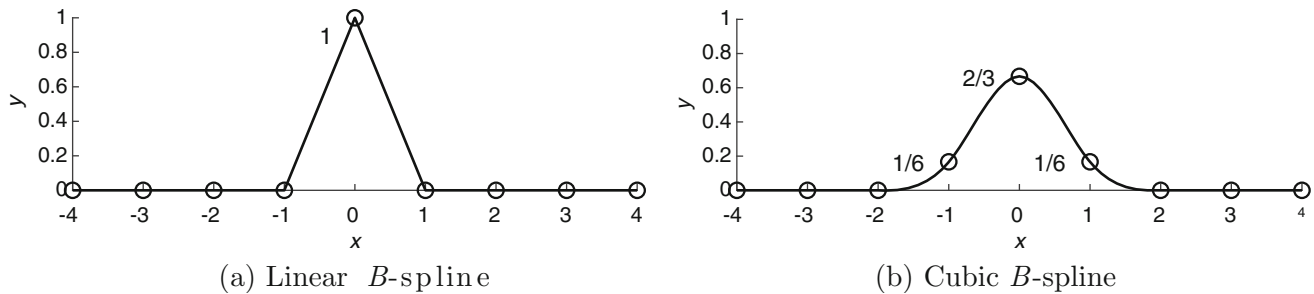
B-Splines on Equi-Spaced Nodes

We will mention just three B -spline relations here (in case of unit-spaced nodes and centered at the origin):

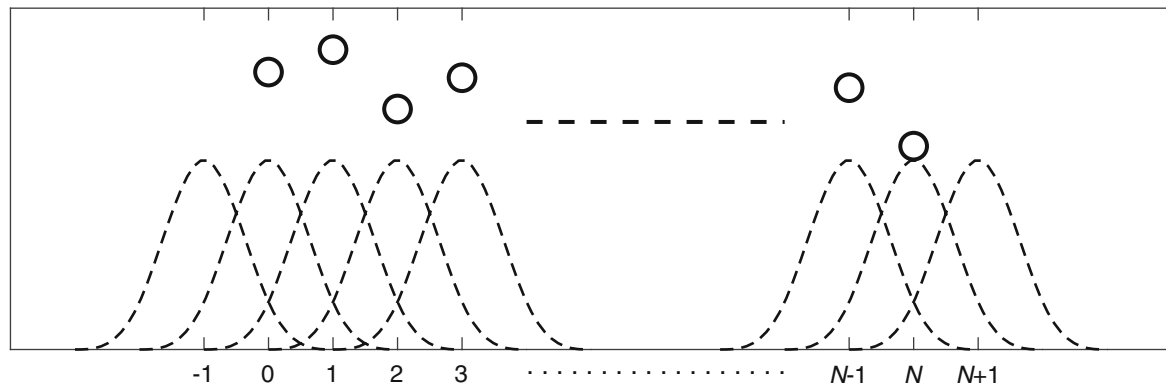
1. With $B_p(x)$ denoting the B -spline of order p , and $B_0(x) = \begin{cases} 1 & -\frac{1}{2} < x < \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$ $B_p(x)$ is the convolution of $B_{p-1}(x)$ with $B_0(x)$.
2. The Fourier transform of a B -spline is remarkably simple: $\int_{-\infty}^{\infty} B_p(x) \cos \omega x dx = \left(\frac{\sin(\omega/2)}{\omega/2} \right)^{p+1}$.
3. Derivatives (and integrals) of B -splines can be conveniently evaluated by using $B'_p(x) = B_{p-1}\left(x + \frac{1}{2}\right) - B_{p-1}\left(x - \frac{1}{2}\right)$.

Creating an Equi-Spaced Cubic Spline Interpolant with the Use of B -Splines

Figure 3 shows schematically (as circles) some data $y(i)$, $i = 0, 1, \dots, N$. Any cubic spline $s(x)$ over these nodes can be written as a linear combination of translates of the cubic B -spline: $s(x) = \sum_{i=-1}^{N+1} \alpha_i B_i(x)$ (the subscript on B now marks the location at which the B -spline is centered, which includes one extra node outside each interval end). The first task is to determine the α_i -coefficients, as the spline then can be readily evaluated at any location. In the linear system (1), the top and



Splines, Fig. 2 Examples of B -splines on unit-spaced grids



Splines, Fig. 3 Schematic illustration of data at $x = 0, 1, 2, \dots, N$ and B -splines centered at $x = -1, 0, 1, 2, \dots, N, N + 1$

bottom equations come from the extra conditions that have to be imposed (shown here for the natural spline), while the remaining equations enforce that the splines takes the correct values at all nodes:

$$\begin{bmatrix} 1 & -2 & 1 & & \\ 1 & 4 & 1 & & \\ & 1 & 4 & 1 & \\ & & \ddots & \ddots & \ddots \\ & & & 1 & 4 & 1 \\ & & & & 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} \alpha_{-1} \\ \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_N \\ \alpha_{N+1} \end{bmatrix} = 6 \begin{bmatrix} 0 \\ y_0 \\ y_1 \\ \vdots \\ y_N \\ 0 \end{bmatrix} \quad (1)$$

← Left BC

← Match

← values at

← nodes

← Right BC

The 3-group of entries in the top row (and similarly in the bottom row) are obtained from the fact that $[B_{-1}''(0) \ B_0''(0) \ B_1''(0)] = [1 \ -2 \ 1]$ (making $s'' = 0$). In the case of a Not-a-Knot spline, we similarly need to use that $B'''(x)$ for $x = \{-2, -1, 0, 1, 2\}$ evaluates to $\{1, -2, 0, 2, -1\}$.

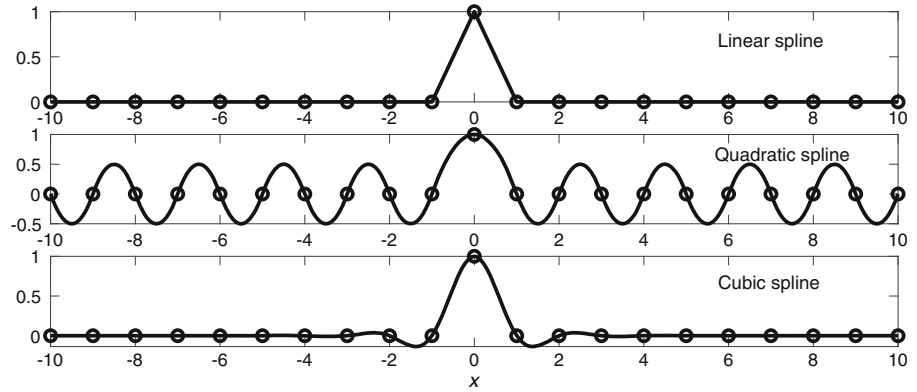
Cardinal Splines

Odd order splines are generally preferred over even orders. One reason is illustrated in Fig. 4, where we see spline interpolants to *cardinal data*. In contrast to B -splines, cardinal splines (non-zero at just one node) of degree $p > 1$ oscillate forever. In the cubic case, the amplitudes decay rapidly. (For successive extrema, the ratio of magnitudes approach $2 - \sqrt{3} \approx 0.2679$. Increasing smoothness leads to slower decay; for $p = 5, 7, 9$ the ratios become approximately 0.43, 0.53, 0.61, respectively; approaching 1 as $p \rightarrow \infty$.) In the quadratic spline case, they don't even decay. (This can however be remedied by letting successive quadratics join half-way between where the data values are imposed.) Like with Lagrange's interpolation polynomial, cardinal functions are of more theoretical than computational utility.

Splines on Non-uniformly Spaced Grids

If the nodes are not equi-spaced, the procedure in section [“Creating an Equi-Spaced Cubic Spline Interpolant with the Use of \$B\$ -Splines”](#) still works, as long as we can create non-equispaced cubic B -splines (which will still extend over 5 nodes). There are three main options for finding such:

Splines, Fig. 4 Equi-spaced cardinal splines on an infinite interval



1. Piece-by-piece construction of the cubics.
2. Represent $B(x)$ as a linear combination of the functions $|x - x_k|^3$, $k = 1, 2, \dots, 5$.
3. Use recursions similar to those in Newton's interpolation formula (detailed for ex. in (de Boor 2001), Chapter X and (Powell 1981), Section 19.4).

For **Case (1)**, one notes that, at a node $x = x_i$, the two cubics that meet can only differ by a term $\alpha_i(x - x_i)^3$. After 5 nodes, the requirement of starting and ending with all 4 cubic polynomial coefficients zero leads to a 4×5 homogeneous linear system, guaranteed to have a nontrivial solution.

For **Case (2)**, with equi-spaced nodes at $x = \{-2, -1, 0, 1, 2\}$,

$$B(x) = \frac{1}{12} \left(|x+2|^3 - 4|x+1|^3 + 6|x|^3 - 4|x-1|^3 + |x-2|^3 \right) \quad (2)$$

This function satisfies all the requirements of a B -spline: It is a cubic in each sub-interval, jumps only the third derivative at the nodes, and becomes identically zero outside $[-2, 2]$. This formula generalizes to arbitrarily spaced nodes:

$$B(x) = 2 \left(\frac{|x-x_1|^3}{\prod_{k \neq 1} (x_1 - x_k)} + \frac{|x-x_2|^3}{\prod_{k \neq 2} (x_2 - x_k)} + \dots + \frac{|x-x_5|^3}{\prod_{k \neq 5} (x_5 - x_k)} \right), \quad (3)$$

(again normalized to make the integral equal to one).

Some Generalizations

Parametric Splines

A curve in d dimensions, given by points $x_1(i), x_2(i), \dots, x_d(i)$, $i = 0, 1, \dots, N$, can be interpolated by equi-spaced splines in the joint parameter i .

Smoothing Splines

Given many points, maybe with noise present, interpolation might cause over-fitting. On a unit spaced grid, a very simplistic approach is to use as B -spline coefficients directly the function values (instead of calculating these for ex. by (1)).

In the *Modified Akima* approach (Akima 1970), the pieces are still cubics, but give up the continuity of second derivative at nodes in exchange for much reduced oscillations. This approach can be advantageous, e.g., when dealing with variables that ought not change sign (such as pressure, density, etc.).

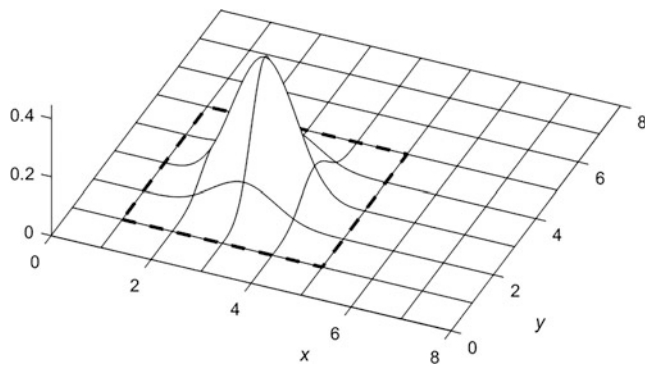
Least Square Fitting with Splines

The advantage over smoothing splines is that one can utilize that smoother curves need fewer coefficients. If one wishes to write an own implementation, one can form an over-determined linear system $A\mathbf{a} = \mathbf{y}$, where $\mathbf{a}$ will provide the B -spline coefficients of the resulting spline. The vector $\mathbf{y}$ should be the full set of data values, and the columns of A should be made up of the different B -splines, when evaluated at the respective data locations. This matrix A will be sparse, so iterative solvers will be attractive for very large problem sizes.

Generalizations to Multiple Dimensions

Structured Nodes in Multiple Dimensions

Bicubic splines on 2-D rectilinear grids were introduced in (de Boor 1962). Higher dimensional counterparts to the 1-D B -spline are then obtained by forming tensor products, as illustrated in the 2-D case in Fig. 5. Linear combinations of such can then be used similarly to the 1-D illustration in Fig. 3.



Splines, Fig. 5 Tensor product of cubic B -splines in 2-D

Unstructured Nodes in Multiple Dimensions: Radial Basis Functions (RBFs)

The following two observations lead to major opportunities for generalizing splines:

1. The interpolation accuracy of splines of order p on a grid with spacing h scales as $O(h^{p+1})$.
2. A 1-D function $s(x) = \sum_{k=1}^N \lambda_k |x - x_k|^p$ is a spline of order p (assuming p odd (If p is even, this formula instead gives a single polynomial of degree p .); cf. (2), (3)).

This can be written as

$$s(x) = \sum_{k=1}^N \lambda_k \phi(|x - x_k|), \text{ with } \phi(r) = r^p. \quad (4)$$

Regarding item #1, the order of the error $O(h^{p+1})$ is a consequence of $|r|^p$ (with p odd) having its p^{th} derivative discontinuous at the origin. Choosing instead, say, $\phi(r) = e^{-(\varepsilon r)^2}$ (where $\varepsilon > 0$ is a *shape parameter*), there is no derivative discontinuity, and the accuracy generally improves from algebraic to spectral (beyond any algebraic order).

Regarding item #2, one can generalize (4) to

$$s(\underline{x}) = \sum_{k=1}^N \lambda_k \phi(\|\underline{x} - \underline{x}_k\|), \quad (5)$$

where $\|\cdot\|$ denotes the standard Euclidean 2-norm. Here, the nodes $\underline{x}_k$ can be arbitrarily scattered in a d -dimensional space. For a wide choice of radial functions $\phi(r)$, the linear system for the $\underline{\lambda}_k$ that arises when enforcing $s(\underline{x}_k) = \underline{y}_k$ (function values at the nodes) can never be singular, no matter how distinct nodes $\underline{x}_k$, $k = 1, 2, \dots, N$ are scattered in any number of dimensions.

A wealth of theory and practical computational results and enhancements are available, such as

1. Using $\phi(r) = r^2 \log r$ and $\phi(r) = r$ generalizes to 2-D and 3-D, respectively, the cubic spline minimal curvature result.
2. As described above, the linear systems will have a full (rather than a sparse) coefficient matrix. However, the “local” RBF-FD (Radial Basis Function-generated Finite Differences) approach overcomes this. (RBF-FD for interpolation is closely related to *kriging* – a statistically motivated approach to also predict function values by weighted average of nearby known ones. Splines and kriging are compared in (Dubrule 1984).)
3. With RBF-FD, it can be advantageous to use $\phi(r) = r^3$ and $\phi(r) = r^5$ (as for cubic and quintic splines), but to supplement the RBF sum with multivariate polynomials.
4. Applications of RBF-FD extend well beyond data interpolation, to solving a wide range of PDEs in complex geometries.

RBFs in the present application of geoscience are surveyed in (Flyer et al. 2014; Fornberg and Flyer 2015).

Summary and Conclusions

Splines provide a routine approach for interpolating or for finding smooth approximations to data, especially in one dimension. A rich set of library routines are available in almost all programming environments. Available variations include reduction of data noise, adjusting smoothness, and preventing spurious oscillations (such as enforcing positivity), etc. Generalizations are also available both to gridded and to fully unstructured data in higher dimensions.

Bibliography

- Akima H (1970) A new method of interpolation and smooth curve fitting based on local procedures. J ACM 17(4):589–602
- de Boor C (1962) Bicubic spline interpolation. J Math Phys 41(1–4):212–218
- de Boor C (2001) A practical guide to splines, Revised edn. Springer, New York/Berlin/Heidelberg
- Dierckx P (1993) Curve and surface fitting with splines. Oxford University Press, Oxford
- Dubrule O (1984) Comparing splines and kriging. Comput Geosci 10(2–3):327–338
- Flyer N, Wright GB, Fornberg B (2014) Radial basis function-generated finite differences: a mesh-free method for computational geosciences. In: Freedon W, Nashed M, Sonar T (eds) Handbook of geomathematics. Springer, Berlin/Heidelberg. https://doi.org/10.1007/978-3-642-27793-1_61-1
- Fornberg B, Flyer N (2015) A primer on radial basis functions with applications to the geosciences. SIAM, Philadelphia
- Powell MJD (1981) Approximation theory and methods. Cambridge University Press, Cambridge
- Prattusch H, Boehm W, Paluszny M (2002) Bézier and B-spline techniques. Springer, Berlin
- Schoenberg IJ (1946) Contributions to the problem of approximation of equidistant data by analytic functions. Q Appl Math 4:45–99, 112–141

Standard Deviation

Alejandro C. Frery

School of Mathematics and Statistics, Victoria University of Wellington, Wellington, New Zealand

Definition

The standard deviation is a measure of spread of either a random variable or a sample. When referring to a random variable, it requires the existence of at least the second moment. When referring to a sample, it requires at least two observations.

Overview

Consider the real-valued random variable $X: \Omega \rightarrow \mathbb{R}$, in which Ω is the sample space; the distribution of X is uniquely characterized by its cumulative distribution function $F_X(t) = \Pr(X \leq t)$. Moreover, if X is continuous, then its distribution is also characterized by the probability density function $f_X(x) = dF_X(x)/dx$. The first and second moments of the continuous random variable X are, respectively, $E(X) = \int xf(x)dx$ and $E(X^2) = \int x^2f(x)dx$, provided the integrals exist. If X is discrete with values in $\Xi = \{\xi_1, \xi_2, \dots\} \subset \mathbb{R}$, the distribution of X is uniquely characterized by the probability function $p_X = (\xi_i, p_i = \Pr(X = \xi_i))$. The first and second moments of the discrete random variable X are, respectively, $E(X) = \sum_i \xi_i p_i$ and $E(X^2) = \sum_i \xi_i^2 p_i$, provided the sums exist. If both $E(X)$ and $E(X^2)$ are finite, then the variance of X is also finite and is given by $\text{Var}(X) = \sigma^2(X) = E(X^2) - (E(X))^2$. We then define the standard deviation of X as the square root of its

variance: $\text{SD}(X) = \sigma(X) = \sqrt{\text{Var}(X)}$. Since the variance cannot be negative, this quantity is always well defined.

The standard deviation is a measure of the spread of the distribution around its expected value. It is expressed in the same units of the random variable. Table 1 presents the name, notation, probability density function, and standard deviation of some commonly used continuous distributions. We denote the indicator function of the set A as $\mathbb{1}_A(x)$, i.e., $\mathbb{1}_A(x)$ is one if $x \in A$ and zero otherwise. The gamma function $\Gamma(z)$ is the extension of the factorial function to complex and, in our case, real numbers: $\Gamma(z) = \int_{\mathbb{R}_+} x^{z-1} e^{-x} dx$. Notice that the gamma distribution reduces to the exponential distribution when $\alpha = 1$. Also, the Student's t-distribution becomes the Cauchy distribution when $\nu = 1$ (Fig. 1).

Table 2 presents the name, notation, parameter space, probability function, and standard deviation of some commonly used discrete distributions.

Figure 2 shows the probability function of a Poisson random variable with mean $\lambda = 3.5$, along with the mean and the area corresponding to $\lambda \pm \sqrt{\lambda}$. The probability in this area is $\Pr(2 \leq X \leq 5) \approx 0.72$.

The Chebyshev's inequality relates the mean and standard deviation of a random variable with the probability of observing values distant from the mean:

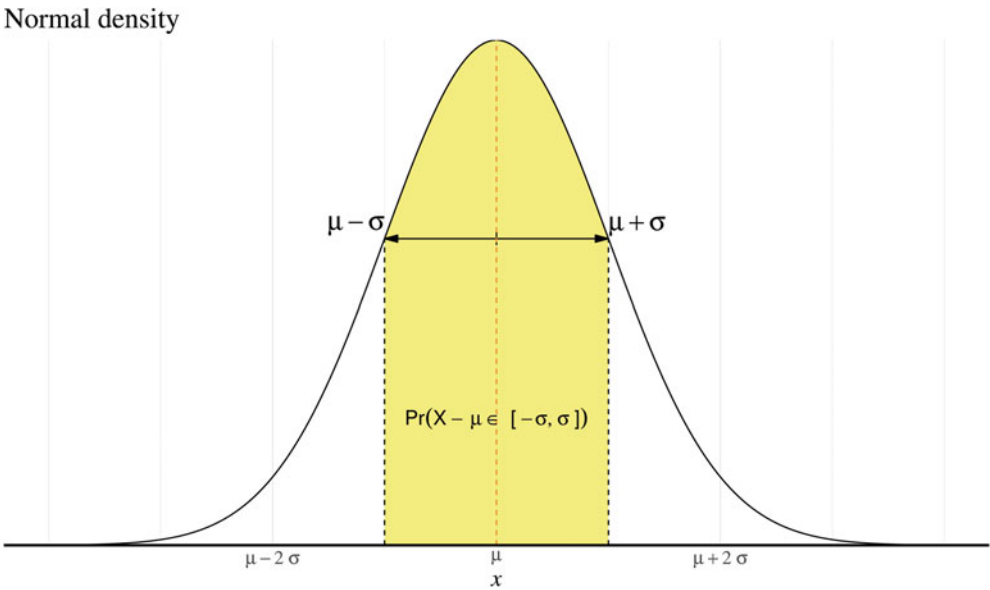
$$\Pr(|X - \mu| \geq z\sigma) \leq \frac{1}{z^2}.$$

Consider the sample n of real values $\mathbf{x} = (x_1, x_2, \dots, x_n)$. The sample mean is $\bar{\mathbf{x}} = n^{-1} \sum_{i=1}^n x_i$, and the sample standard deviation is $s(\mathbf{x}) = \sqrt{n^{-1} \sum_{i=1}^n (x_i - \bar{\mathbf{x}})^2}$.

The user should always check the accuracy of the computational platform. Simple computations as, for instance, the sample standard deviation, are prone to gross numerical errors. The reader is referred to the work by Almiron et al. (2010) for an analysis of the numerical precision of five

Standard Deviation, Table 1 Continuous distributions, parameter space, their densities, expected values, and standard deviations

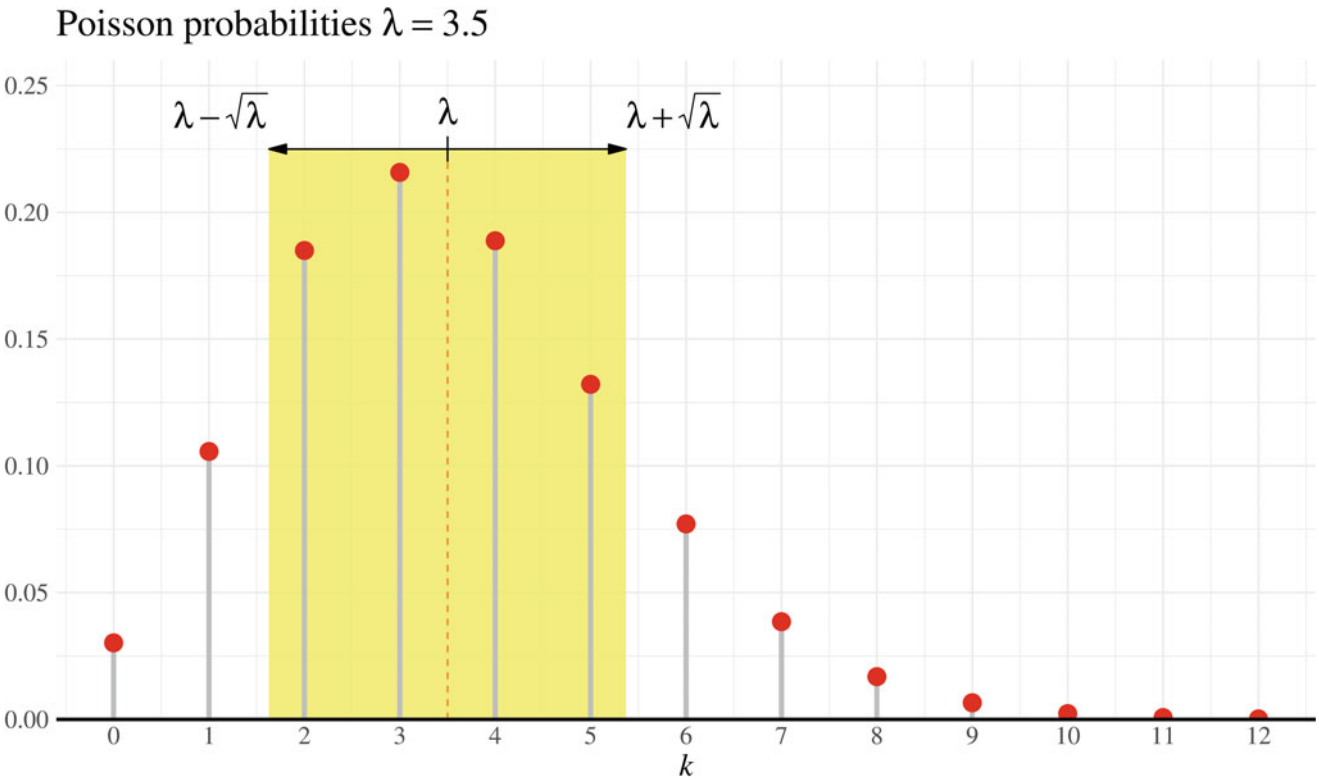
Name and notation	Parameter space	$f_X(x)$	$E(X)$	$\text{SD}(X)$
Uniform $\mathcal{U}_{(a,b)}$	$-\infty < a < b < \infty$	$(b-a)^{-1} \mathbb{1}_{(a,b)}(x)$	$(a+b)/2$	$(2\sqrt{3})^{-1}(b-a)$
Gamma $\Gamma(\alpha, \beta)$	$\alpha, \beta > 0$	$\frac{\beta^\alpha x^{\alpha-1} \exp\{-\beta x\}}{\Gamma(\alpha)} \mathbb{1}_{\mathbb{R}_+}(x)$	α/β	$\sqrt{\alpha}/\beta$
Normal $N(\mu, \sigma^2)$	$\mu \in \mathbb{R}, \sigma > 0$	$\frac{\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}}{\sqrt{2\pi}\sigma}$	μ	σ
Student's t $t(\nu)$	$\nu > 0$	$\frac{\Gamma((\nu+1)/2) \left(1+x^2/\nu\right)^{-(\nu+1)/2}}{\sqrt{\nu\pi} \Gamma(\nu/2)}$	0 if $\nu > 1$, Otherwise undefined	$\sqrt{\nu/(\nu-2)}$ if $\nu > 2$, Otherwise undefined
Lognormal $\text{LN}(\mu, \sigma^2)$	$\mu \in \mathbb{R}, \sigma > 0$	$\frac{\exp\left\{-(\ln x - \mu)^2/(2\sigma^2)\right\}}{x\sigma\sqrt{2\pi}} \mathbb{1}_{\mathbb{R}_+}(x)$	$\exp\{\mu + \sigma^2/2\}$	$\sqrt{(e^{\sigma^2} - 1) \exp\{2\mu + \sigma^2\}}$
Weibull $W(\lambda, k)$	$\lambda, k > 0$	$\frac{k\lambda^k \exp\left\{-(x/\lambda)^k\right\}}{\lambda^k} \mathbb{1}_{\mathbb{R}_+}(x)$	$\lambda\Gamma(1+1/k)$	$\lambda\sqrt{\Gamma(1+2/k) - \Gamma^2(1+1/k)}$
Beta $B(\alpha, \beta)$	$\alpha, \beta > 0$	$\frac{\Gamma(\alpha+\beta)x^{\alpha-1}(1-x)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} \mathbb{1}_{(0,1)}(x)$	$\alpha/(\alpha+\beta)$	$\sqrt{\alpha\beta/(\alpha+\beta+1)/(\alpha+\beta)}$



Standard Deviation, Fig. 1 Normal density, standard deviations, and probability

Standard Deviation, Table 2 Discrete distributions, their probability functions, parameter space, expected values, and standard deviations

Name and notation	Parameter space	$\Pr(X = k)$	$E(X)$	$SD(X)$
Binomial $\text{Bi}(p, n)$	$0 < p < 1, n \in \mathbb{N}$	$\binom{n}{k} p^k (1-p)^{n-k}$	np	$\sqrt{np(1-p)}$
Poisson $\text{Po}(\lambda)$	$\lambda > 0$	$\frac{\lambda^k e^{-\lambda}}{k!}$	λ	$\sqrt{\lambda}$
Negative binomial $\text{NBi}(p, n)$	$0 < p < 1, n \in \mathbb{N}$	$\binom{k+n-1}{n-1} (1-p)^n p^n$	$pn/(1-p)$	$\sqrt{pn/(1-p)}$



Standard Deviation, Fig. 2 Poisson probability function for $\lambda = 3.5$

spreadsheets (Calc, Excel, Gnumeric, NeoOffice, and Oleo) running on two hardware platforms (i386 and amd64) and on three operating systems (Windows Vista, Ubuntu Intrepid, and Mac OS Leopard). The article discusses a methodology for assessing their performance with datasets of varying numerical difficulty.

Summary

The standard deviation is a measure of the spread of either the probability of a distribution or the proportion of observed data around the mean.

Bibliography

- Almiron M, Vieira BL, Oliveira ALC, Medeiros AC, Frery AC (2010) On the numerical accuracy of spreadsheets. *J Stat Softw* 34(4):1–29. <https://doi.org/10.18637/jss.v034.i04>. URL <http://www.jstatsoft.org/v34/i04>
- Johnson NL, Kotz S, Kemp AW (1993) *Univariate discrete distributions*. Wiley series in probability and mathematical statistics, 2nd edn. Wiley, New York
- Johnson NL, Kotz S, Balakrishnan N (1994) *Continuous univariate distributions*. Wiley series in probability and mathematical statistics, vol 1, 2nd edn. Wiley, New York
- Johnson NL, Kotz S, Balakrishnan N (1995) *Continuous univariate distributions*, vol 2, 2nd edn. Wiley, New York

Stationarity

Francky Fouedjio
Department of Geological Sciences, Stanford University,
Stanford, CA, USA

Definition

The stationarity concept refers to the translation invariance of all or some statistical characteristics of a random function (also referred to as random field or spatial stochastic process). A random function is called strictly stationary if its spatial distribution is invariant under spatial shifts. Second-order (or weakly or wide-sense) stationarity for a random function means that its first- and second-order moments are translation invariant. A random function is said to be intrinsically stationary if its increments are second-order (or weakly or wide-sense) stationary random functions. A random function is said to be locally (or quasi) second-order stationary if its first two moments are approximately stationary at any given position of a neighborhood moved around in the study domain.

Introduction

A classical problem in geosciences is predicting a spatially continuous variable of interest over the whole study region, from measurements taken at some locations. Examples of spatially continuous variables include the elevation, the temperature, the moisture, the rainfall, the soil fertility, the permeability, the porosity, and the mineral grade. In geostatistics, the spatially continuous variable of interest is modeled as a random function (also referred to as random field or spatial stochastic process) and the observed data as the random function's unique realization. Simplifying modeling assumptions such as the stationarity are often made on the random function to make the statistical inference possible. The notion of stationarity broadly refers to the invariance by translation of all or part of the random function's spatial distribution.

The need for a stationarity assumption is based on the fact that any type of statistical inference requires some sort of data replication. Two reasons prevent us from carrying out statistical inference in all generality. On the one hand, one has only a single realization of the random function; the observed data are considered as a unique realization of the random function. This takes all its meaning because there are not multiple parallel physical worlds. On the other hand, the realization is only partially known at certain sampling locations. However, this second restriction is not as problematic as the first one. The question of inference would still arise if one knew the reality exhaustively because the spatial distribution or its moments are not regional quantities (i.e., any quantity whose value is determined by the data). To evaluate them, it would take many realizations (multiple copies) of the random function, but these are purely theoretical and do not exist in reality.

To get out of this impasse, certain restrictions on the random function are necessary. They appeal to the notion of stationarity, which describes, in a way, a form of spatial homogeneity of the random function. The basic idea is to allow statistical inference by replacing the replication on the realizations of the random function by the replication in the space. Thus, the values encountered in the different areas of the study region present the same characteristics and can be considered as different realizations of the same random function. Thus, the stationarity assumption provides some replication of the data, which makes statistical inference possible.

There are different forms of stationarity, depending on which of the random function's statistical characteristics are restricted and the working spatial scale. They include the strict stationarity, the second-order (or weakly or wide-sense) stationarity, the intrinsic stationarity, and the local (or quasi) second-order stationarity. These notions can be found in the following reference books: Cressie (2015), Wackernagel (2013), Chiles and Delfiner (2012), Olea (2012), Christakos (2012), Webster and Oliver (2007), Diggle

and Ribeiro (2007), Goovaerts (1997), and Isaaks and Srivastava (1989).

Formulation

Suppose that the spatially continuous variable of interest is modeled as a random function (also referred to random field or spatial stochastic process) $\{Z(x) : x \in D \subseteq \mathbb{R}^d\}$ defined over a fixed continuous spatial domain of interest $D \subseteq \mathbb{R}^d$, where $\mathbb{R}^d$ is d -dimensional Euclidean space, typically $d = 2$ (plane), and $d = 3$ (space).

The random function $Z(\cdot)$ is said to be strictly stationary if its spatial distribution is invariant under translations, that is to say,

$$F_{x_1, \dots, x_k}(v_1, \dots, v_k) = F_{x_1+h, \dots, x_k+h}(v_1, \dots, v_k) \quad (1)$$

for all integers $k \in \mathbb{N}^*$, collections of points $x_1, \dots, x_k$ in the spatial domain D , set of possible values $v_1, \dots, v_k$, and vectors $h \in \mathbb{R}^d$; in which F denotes finite-dimensional joint distributions defined by $F_{x_1, \dots, x_k}(v_1, \dots, v_k) = P(Z(x_1) \leq v_1, \dots, Z(x_k) \leq v_k)$.

Under the strict stationarity assumption, finite-dimensional joint distributions characterizing the spatial distribution of the random function stay the same when shifting a given set of points from one part of the study domain to another. In other words, a translation of a point configuration in a given direction does not change finite-dimensional joint distributions. Thus, the random function appears as homogeneous and self-repeating in the whole spatial domain. The strict stationarity assumption is very restrictive because it assumes an identity of all probability distributions in the space. Moreover, it is only partially verifiable since it is made for all $k \in \mathbb{N}^*$, and it does not tell us about the existence of moments of the random function.

The random function $Z(\cdot)$ is said to be second-order (or weakly or wide-sense) stationary if its first- and second-order moments exist and are both invariant under translations over the spatial domain D , i.e., $\forall x, x+h \in D$,

$$E(Z(x)) = E(Z(x+h)) \quad (2)$$

$$\text{Cov}(Z(x+h), Z(x)) = C(h) \quad (3)$$

Thus, under the second-order stationarity, the expected value (or mean) of the random function is constant, i.e., the same at any point x of the spatial domain D . The covariance function which measures the second-order association between any pair of points depends only on the separation between them in both distance and direction. When a random function is second-order stationary, its first- and second-order moments are also said to be stationary. The second-order

stationarity implies the existence not only of the mean and the covariance function but also other second-order moments such as the variance, the variogram, and the correlogram: $\forall x, x+h \in D$,

$$\text{Var}(Z(x)) = \text{Cov}(Z(x), Z(x)) = C(0) \quad (4)$$

$$\frac{1}{2} \text{Var}(Z(x+h) - Z(x)) = C(0) - C(h) = \gamma(h) \quad (5)$$

$$\frac{\text{Cov}(Z(x+h), Z(x))}{\sqrt{\text{Var}(Z(x)) \times \text{Var}(Z(x+h))}} = \frac{C(h)}{C(0)} = \rho(h) \quad (6)$$

In particular, when $h = 0$, the covariance comes back to the ordinary variance which must also be constant. Stationarity of the covariance implies stationarity of variance, and the variogram, and the correlogram. These moments also do not depend on the absolute position of the points where they are calculated but only on their separation. The second-order stationarity assumption requires the existence of the covariance function which might not exist.

The strict stationarity requires all the moments of the random function (if they exist) to be invariant under translations. In particular, the strict stationarity implies the weak stationarity when the mean and the covariance function of the random function exist. These two forms of stationarity are equivalent if the random function is Gaussian. Indeed, since a Gaussian distribution is completely defined by its first two moments, knowledge of the mean and the covariance function suffices to determine the spatial distribution of a Gaussian random function. Overall, weak stationarity is less restrictive than strict stationarity. Moreover, it turns out that the weak stationarity condition is sufficient for most of the statistical results derived for spatial data to hold. For these reasons, weak stationarity is employed more often than strict stationarity. Various authors use the term stationarity to refer to weak stationarity.

The random function $Z(\cdot)$ is called intrinsically stationary if for any vector $h \in \mathbb{R}^d$, the increment (or difference) $\{Z_h(x) = Z(x+h) - Z(x) : x \in D \subseteq \mathbb{R}^d\}$ is a second-order stationary random function. Specifically,

$$E(Z(x+h) - Z(x)) = 0 \quad (7)$$

$$\begin{aligned} \text{Var}(Z(x+h) - Z(x)) &= E\left((Z(x+h) - Z(x))^2\right) \\ &= 2\gamma(h) \end{aligned} \quad (8)$$

The intrinsic stationarity focuses on the second-order stationarity of the increments of the random function. It is less restrictive than the second-order stationary. Indeed, the second-order stationarity implies the intrinsic stationarity.

The converse is not necessarily true. The intrinsic stationarity hypothesis expands the second-order stationarity hypothesis which no longer relates to the increments of the random function. The intrinsic stationarity is the reason that the variogram has emerged as the preferred measure of spatial association in geostatistics. The variogram can still be defined for cases where the covariance function and the correlogram cannot, and it does not require knowledge of the mean. The covariance function of an intrinsic stationary random function might not exist; it exists only if the variogram is bounded, in which case one has the relation,

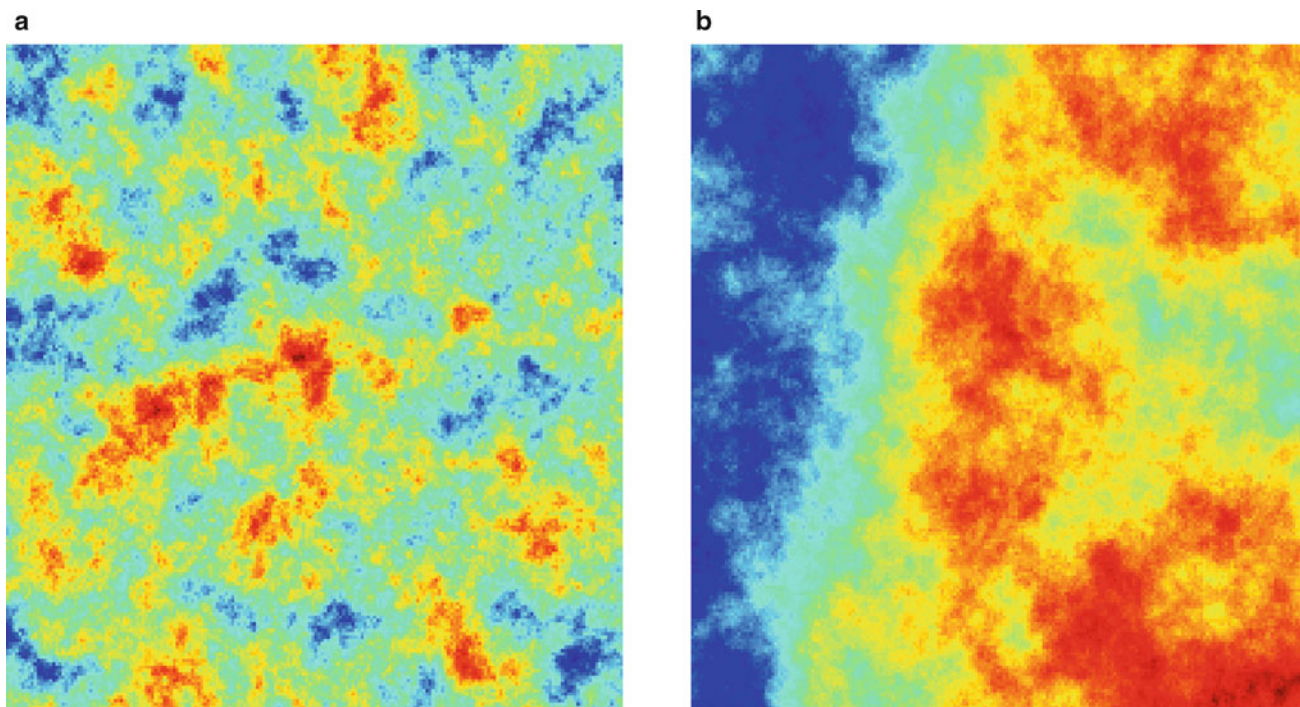
$$\gamma(h) = C(0) - C(h) \quad (10)$$

Thus, under the second-order stationarity, it is equivalent to work with the covariance function or the variogram, one deduced from the other by the above relation. Classical Brownian motion in one dimension (or Wiener process) provides an example of a Gaussian random function that is intrinsically stationary, but not second-order stationary. It has independent and stationary Gaussian increments, and its variogram is given by $\gamma(h) = |h|$ while its covariance function is expressed as follows: $\text{Cov}(Z(s), Z(t)) = \min(s, t), s, t > 0$. Examples of realizations of second-order and intrinsically stationary random functions defined on the unit square are given in Fig. 1.

The different forms of stationarity presented previously do not explicitly involve the working spatial scale, which is nevertheless an essential parameter in practice. A model valid at a particular spatial scale may no longer be valid at larger or smaller spatial scales. In most applications (especially prediction problems), it is not useful for second-order stationary or intrinsically stationary assumptions to be valid at the scale of the whole spatial domain, but only for distances less than a certain limit distance. Thus, they do not require global stationarity, but only local stationarity.

A random function is said to be locally (or quasi) second-order stationary if its mean is a very regular function varying slowly in space such that $E(Z(x+h)) \approx E(Z(x))$ if $\|h\| \leq b$ ($b > 0$), and its covariance function $\text{Cov}(Z(x+h), Z(x))$ only depends on the vector h for distances $\|h\| \leq b$ ($b > 0$). Such random functions, which are not necessarily second-order stationary at the scale of the whole spatial domain, have a smooth mean function, which can be evaluated locally as a constant and a covariance function which can be considered locally as stationary.

Under the local (or quasi) second-order stationarity assumption, the statistical inference of the first- and second-order moments requires in practice to have enough data locally. Thus, the local second-order stationarity hypothesis is therefore a trade-off between the size of the neighborhood where the random function can be considered as stationary



Stationarity, Fig. 1 (a) example of realization of a second-order stationary random function; and (b) example of realization of an intrinsically stationary random function

and the number of data necessary to perform a reliable statistical inference.

In addition to the stationarity assumption, isotropy assumption is often added, which extends the notion of translational invariance of stationarity by including rotational invariance (around the origin). In other words, the covariance function (or variogram) is simply a function of the Euclidean distance, i.e., $C(h) = C(\|h\|)$ or $\gamma(h) = \gamma(\|h\|)$. One then calls both the random function and the covariance function (or variogram) isotropic. Thus, stationarity means translation-invariance of the statistical properties of the random function, while isotropy is a corresponding rotation-invariance.

Limitations

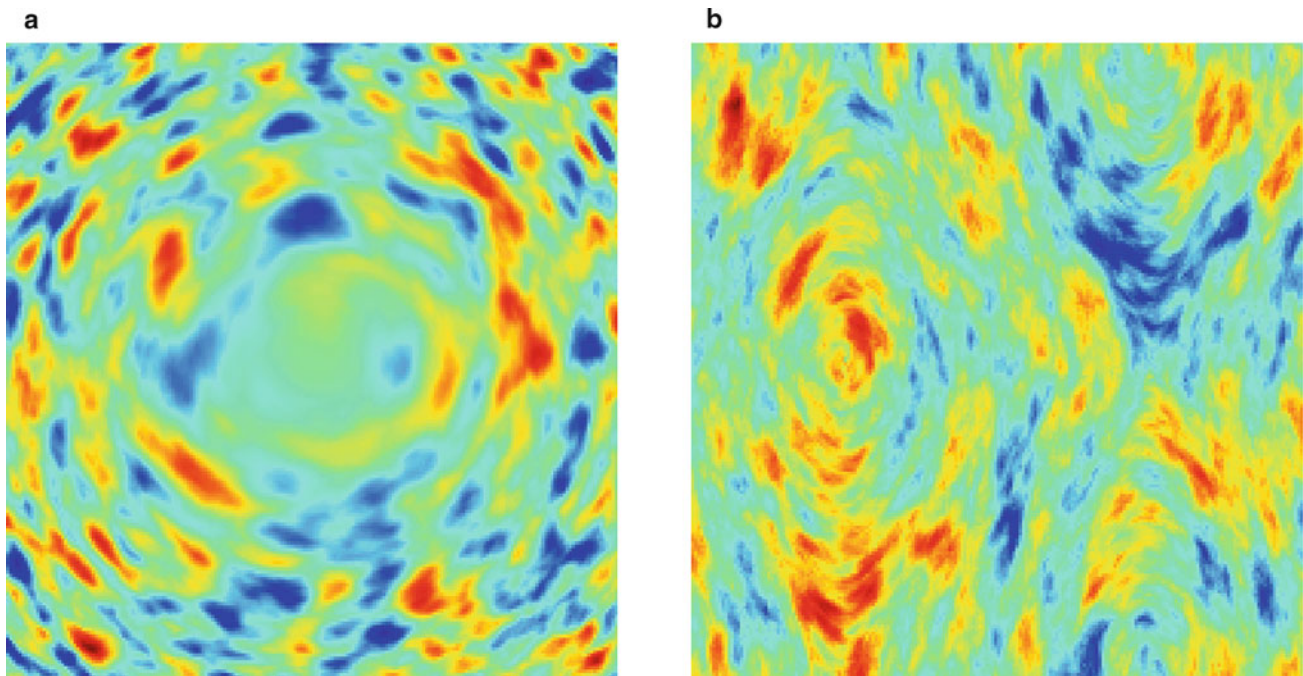
The assumption that all or some statistical characteristics of the random function are translation invariant over the whole spatial domain may be appropriate, when the latter is small in size, when there is not enough data to justify the use of a complex model, or simply because there is no other reasonable alternative. Although often justified and leading to a reasonable analysis, this assumption is often inappropriate and unrealistic given certain spatial data collected in practice. Stationarity assumption can be doubtful due to many factors, including specific landscape and topographic features of the study region or other localized effects. These local influences are reflected in the fact that observed data can obey to a

covariance function (or variogram) whose characteristics vary across the study domain. In such a setting, a stationary assumption is not appropriate because it could produce less accurate predictions, including an incorrect assessment of the prediction error. Hence, the need to go beyond the stationarity.

When the stationary assumptions do not hold, we are in the non-stationarity framework which means that some statistical characteristics of the random function are no longer translational invariant; they depend on the location. A random function that does not have the stationarity property is called nonstationary. One distinguishes different types of nonstationarity: nonstationarity in mean, nonstationarity in variance, and nonstationarity in covariance (or variogram). It is now recognized that most spatial processes exhibit a nonstationary spatial dependence structure when considering large distances. A detailed discussion on nonstationary random functions and the various models can be found in Fouedjio (2017). Figure 2 shows two examples of realizations of nonstationary random functions defined on the unit square.

Diagnostic Tools

The random function's modeling choice, either in the stationary framework or the nonstationary framework, is a fundamental decision to make during geostatistical analysis. A common practice to check for the second-order stationarity assumption is to informally assess plots of local experimental variograms computed at different subregions across the study



Stationarity, Fig. 2 Two examples of realizations of nonstationary random functions

region. If the different local experimental variograms are bounded and do not differ notably, the assumption of second-order stationarity may be reasonable. Although a useful diagnostic tool, this graphical tool is challenging to assess and open to subjective interpretations.

Hypothesis tests of the assumption of second-order stationarity may be more valuable. However, it is not straightforward to build such tests in the context of a single realization of the random function. Very few formal procedures to test for the second-order stationarity have been developed (Fuentes 2005; Jun and Genton 2012; Bandyopadhyay and Rao 2017). They are adaptations of stationarity tests for time series. Most of them are built up based on the spectral representation of the random function. They consist essentially of testing either the homogeneity or the uncorrelatedness of some variables defined in the frequency or spectral domain. Some of the testing procedures check both for stationarity and isotropy.

Summary and Conclusions

In geostatistics, the spatially continuous variable of interest is modeled as a random function and the observed data as the random function's unique realization. To infer the statistical characteristics of the spatial distribution of the random function from observed data, limiting assumptions are needed. They appeal to the notion of stationarity of the random function. It enables us to treat observed data as though they have the same degree of variation over a region of interest. Several stationary assumptions are possible: strict stationarity, second-order (or weakly or wide-sense) stationarity, intrinsic stationarity, and local (or quasi) second-order stationarity.

Although stationarity assumptions are often violated in practice, they remain fundamental. They form the basic building blocks of more complex models, such as second-order nonstationary random function models. Indeed, in most of these models, the stationarity assumption is present, either locally or globally. The majority of second-order nonstationary random function models include some stationary random function models as special cases. Hence, the concept of stationarity plays a critical role in geostatistics. Although it is challenging to accept or reject a stationarity hypothesis in the context of a single realization of the random function, some diagnostic tools can help to decide whether this assumption is appropriate or not. Moreover, the relevance of a stationarity hypothesis can depend on the working spatial scale considered.

In this entry, the concept of stationarity has been presented in the case of a scalar random function for univariate spatially continuous data. However, it extends to the case of a vector-valued random function (i.e., set of scalar random functions stochastically correlated to one another) for multivariate spatially continuous data.

Cross-References

- [Geostatistics](#)
- [Kriging](#)
- [Variogram](#)

Bibliography

- Bandyopadhyay S, Rao SS (2017) A test for stationarity for irregularly spaced spatial data. *J R Stat Soc B Stat Methodol* 79(1): 95–123
- Chilès JP, Delfiner P (2012) *Geostatistics: modeling spatial uncertainty*. Wiley
- Christakos G (2012) *Random field models in earth sciences*. Dover earth science. Dover Publications
- Cressie N (2015) *Statistics for spatial data*. Wiley
- Diggle P, Ribeiro P (2007) *Model-based geostatistics*. Springer series in statistics. Springer, New York
- Fouedjio F (2017) Second-order non-stationary modeling approaches for univariate geostatistical data. *Stoch Env Res Risk A* 31(8): 1887–1906
- Fuentes M (2005) A formal test for nonstationarity of spatial stochastic processes. *J Multivar Anal* 96(1):30–54
- Goovaerts P (1997) *Geostatistics for natural resources evaluation*. Applied geostatistics series. Oxford University Press
- Isaaks E, Srivastava R (1989) *Applied geostatistics*. Oxford University Press
- Jun M, Genton MG (2012) A test for stationarity of spatio-temporal random fields on planar and spherical domains. *Stat Sin* 22: 1737–1764
- Olea R (2012) *Geostatistics for engineers and earth scientists*. Springer
- Wackernagel H (2013) *Multivariate geostatistics: an introduction with applications*. Springer, Berlin/Heidelberg
- Webster R, Oliver M (2007) *Geostatistics for environmental scientists*. Statistics in practice. Wiley

Statistical Bias

Gilles Bourgault
Computer Modelling Group Ltd., Calgary, AB, Canada

Definitions

- | | |
|-----------------|--|
| Sampling Bias | In its basic definition, statistical bias is simply the difference in values between a statistic that is estimated from a sample and its equivalent characterizing the entire population. It is an intrinsic sampling error that affects all statistics estimated from the partial sampling of a population. |
| Estimation Bias | In a spatial context (geostatistics), a statistical estimate is localized in space, and multiple estimates are required for spatial estimation from a given sampling. Because all non- |

sampled locations are equivalent in regards of statistical estimation, it is the set of estimated values that can be considered for statistical bias. The estimated values will, in general, show a statistical bias in their distribution shape (less variance and less skewness) and in their spatial continuity (greater spatial continuity) when compared to the population values. For spatial estimation, the estimation bias is combined to the sampling bias.

Estimate and Estimator

For any population statistical parameter (θ), and its estimate (θ^*) from a sample, the statistical bias is the difference.

$$\text{Bias} = \theta^* - \theta \quad (1)$$

This difference measures the statistical bias for a given sampling. It cannot be avoided and fluctuates from sampling to sampling. The most important factor affecting this statistical bias is the sample size since the statistic θ^* will converge towards the population value as the sample size increases. But in practice, this statistical bias can never be known from a single sampling since populations are always very large and can never be sampled exhaustively. Instead of correcting the statistical bias, one tries to minimize it. It calls on the unbiasedness of the estimate when it is viewed as a random variable (estimator). The sample estimator (θ^*) is said to be unbiased, if on average over many samplings, the expected value of the statistical bias is zero.

$$E[\theta^* - \theta] = 0 \quad (2)$$

Or equivalently, the expected value of the estimator is equal to the population parameter.

$$E[\theta^*] = \theta \quad (3)$$

For example, from a sample of n values (z_i), the estimator

$$\bar{z} = \frac{\sum_{i=1}^n z_i}{n} \quad (4)$$

is unbiased for estimating the mean of the population (Z).

$$E[\bar{z}] = E[Z] \quad (5)$$

The unbiasedness of an estimator is easily demonstrated by replacing each sample value (z_i) with an equivalent random variable (Z_i), and assuming $E[Z_i] = E[Z]$ for all Z_i .

Estimator Quality

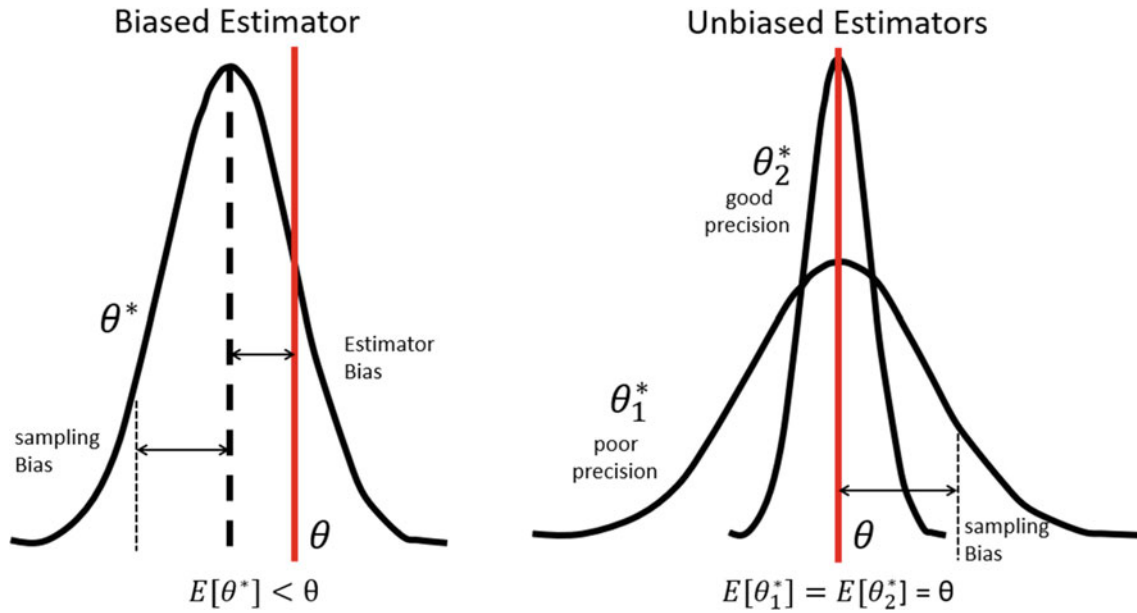
An estimator can be biased or unbiased and of variable precision. For a biased estimator, its expected value, over multiple samplings, will be different from the population value ($E[\theta^*] \neq \theta$). Most often, this type of bias would be introduced by a systematic calibration error in the instrumentations, or in the laboratory processes, used to measure each sample value. For an unbiased estimator, its sampling error may be small or large, depending on its precision. The precision is higher when the estimated statistic varies little between samplings. Usually, the precision increases with the sampling size. Some of the most important estimators' characteristics are depicted in Fig. 1. Note that a biased estimator can still be precise without being accurate. Thus, precision does not guaranty accuracy. The precision of an estimator may also depend on the distribution shape of the population values to be sampled. The precision tends to be lower as the skewness of the distribution is more pronounced.

Spatial Estimate and Spatial Estimator

The mean estimator (Eq. 4) for the global population can be adapted for local spatial estimation (geostatistics) when assuming that each location ($\mathbf{u} = [u_x, u_y, u_z]$) is associated with a random variable $Z(\mathbf{u})$. The set of all spatial locations, within a domain of interest, defines the population (Z). Kriging (Matheron 1965; Armstrong 1998) is a spatially weighted average that has been developed as the best linear unbiased estimator (BLUE) for spatial estimation. Multiple types of kriging exist, but ordinary kriging is the most used in practice.

$$z^*(\mathbf{u}_0) = \sum_{\alpha=1}^n \lambda_{\alpha} * z(\mathbf{u}_{\alpha}) \quad (6)$$

where $z^*(\mathbf{u}_0)$ is the kriging estimate at a spatial location $\mathbf{u}_0$, the $z(\mathbf{u}_{\alpha})$'s are the sample values at n locations ($\mathbf{u}_{\alpha}$), and the λ_{α} 's are the kriging weights. The weights are optimum in the sense that they account for spatial correlation (variogram) between the $n + 1$ locations involved (Isaaks and Srivastava 1989). The estimator is unbiased by ensuring that the sum of the weights equals 1, and assuming each sample value $z(\mathbf{u}_{\alpha})$ can be replaced by an equivalent random variable ($Z(\mathbf{u}_{\alpha})$), with same mean as the spatial population ($E[Z(\mathbf{u}_{\alpha})] = E[Z]$ for all locations $\mathbf{u}_{\alpha}$ in the spatial domain). Note that a statistical bias can still be induced, even with an unbiased spatial estimator, when the sampling favors certain locations in the population (i.e., high grade zones in a mineral deposit). Random or systematic samplings can be used to avoid this type of bias.



Statistical Bias, Fig. 1 Schematic representation for the distribution (over multiple samplings) of a biased estimator (left) and the distributions of two unbiased estimators (right). For the unbiased estimators, the

estimator 2 is more precise than estimator 1 since its distribution over multiple random samplings is tighter around the population value (in red)

Transfer Function Bias

Kriging is a particular estimator (Eq. 6) in the sense that it can generate multiple estimated values, one for each non-sampled location in the spatial domain. Because of estimation bias, the set of all estimated values ($\{z^*\}$) will have different statistical properties, such as lower variance and greater spatial continuity (variogram), as compared with the set of their corresponding population values ($\{z\}$). This will, in general, introduce some functional bias.

$$\text{Bias} = f(\{z^*\}) - f(\{z\}) \quad (7)$$

The bias depends on the degree of estimation bias in the set of estimated values, and on the degree of nonlinearity of the function $f()$ (i.e., spatial correlation, metal recovery function in mining, flow simulation in oil and gas). Conditional simulations (Goovaerts 1997; Deutsch and Journel 1998) can minimize this bias by restoring the data variance and the spatial variability in the estimated values. But contrary to the kriging estimates, simulation results are not unique as an infinite number of realizations can be simulated.

Conditional Bias

Spatial estimators are designed to be unbiased globally for the population, $E[Z^*(\mathbf{u}_0)] = E[Z(\mathbf{u}_0)] = E[Z]$. However, they can still be biased, on average over the spatial domain, when considering a particular estimate value. This brings about

the concept of conditional bias. Kriging is such an estimator that is globally unbiased but can be conditionally biased when each estimated location value is compared to the true value at that location (comparing $z(\mathbf{u}_0)$ with $z^*(\mathbf{u}_0)$ for all estimation locations $\mathbf{u}_0$). In the spatial context, a special quality for an estimator is to be conditionally unbiased. On average, over the spatial domain, one expects that the estimator will predict the correct value. Thus, the estimator is conditionally unbiased if

$$E[Z^* \mid Z^* = z^*] = E[Z \mid Z^* = z^*] \quad (8)$$

or equivalently

$$E[Z \mid Z^* = z^*] = z^* \quad (9)$$

where Z^* is the kriging estimator viewed as a random variable, Z the random variable representing the spatial population, and z^* a kriging estimate value. If one averages all true values, for all locations with the same kriging estimate value, then this average should be equal to that kriging estimate value. The conditional expectation is localized in the sense that the kriging estimates with the same given value z^* are localized in space. These locations represent a subset of the population. The expectation is conditional to this subset (Eq. 8). Conditional unbiasedness is very important in mining to best categorize spatial locations as being mineralized or sterile for example. Minimizing conditional bias is at the origin of kriging. It was first addressed by Krige (1951) and revisited by Bourgault (2021).

In the following, statistical biases and non-biases are illustrated for various statistics, and for the kriging estimator, in using samplings from synthetic datasets.

Global Fluctuations for Basic Statistics

The effects of the sampling size and distribution shape on statistical bias are illustrated when sampling two synthetic populations with same mean and variance, but with different distributions. One is characterized with a normal distribution (no skewness) and the other is characterized with a lognormal distribution (highly skewed). Each population has a total of 2500 values that are spatially distributed on a regular 50×50 grid. The lognormal distribution is from the primary variable of GSLIB dataset (Deutsch and Journel 1998). The normal distribution is the lognormal distribution transformed in normal scores and rescaled to have the same mean and same variance as the lognormal distribution. Fluctuations in sampling mean, variance, histogram, and variogram are illustrated for sample sizes of 50, 100, and 200. For each sample size, 11 random samplings were drawn from the population. For each sampling, the same spatial locations were sampled in both populations. Table 1 presents the fluctuations for the sample mean and the sample variance, over the 11 random samplings, for the 3 different sampling sizes, for the lognormal population. Table 2 presents the fluctuations for the sample mean and the sample variance, over the 11 random samplings, for the 3 different sampling sizes, for the normal population.

The results show that statistical bias from sampling can be significant, but on average, the mean and variance are fairly

well estimated for the two populations. The statistical bias fluctuations (precision), as measured by the range and the variance of each statistic, tend to diminish when the sampling size increases. The fluctuations are much less when the population is normally distributed as compared to lognormally distributed, especially for the variance.

Figures 2, 3, and 4 show the histogram for each sampling with the average histogram over the 11 samplings and the histogram of the lognormal population. For each sampling size, the averaged histogram matches very well the true population histogram. However, the fluctuations between the 11 histograms increase as the sample size increases. Figures 5, 6, and 7 show the histogram for each sampling with the average histogram over the 11 samplings and the histogram of the normal population. Again, for each sampling size, the averaged histogram matches very well the true histogram. Contrary to the lognormal population, the fluctuations between the 11 histograms decrease as the sample size increases.

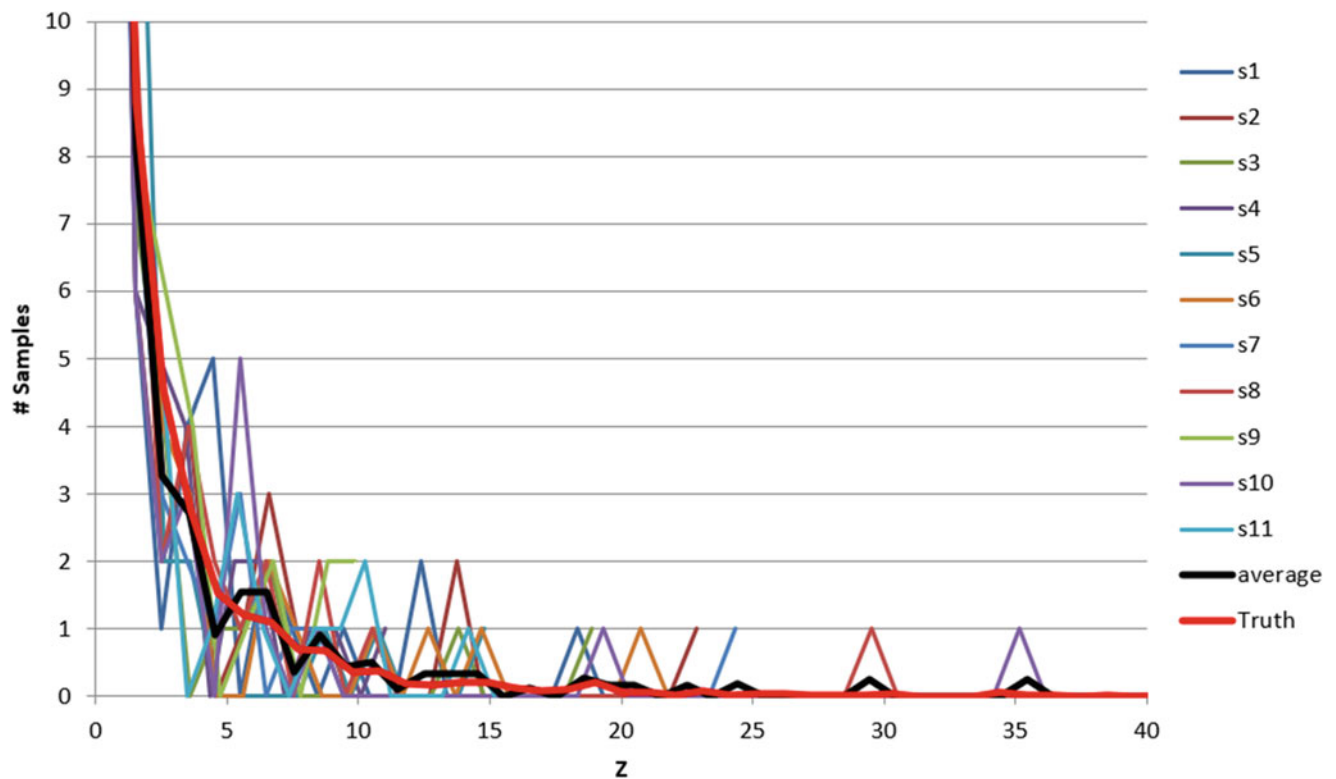
In a spatial context, not only the data histogram is important, but also the data variogram as it needs to be modelled for measuring the spatial correlation required for spatial estimation by kriging (Isaaks and Srivastava 1989). Figures 8, 9, and 10 present the variogram for each sampling with the averaged variogram over the 11 samplings, and the variogram for the lognormal population. For each sampling size, the averaged variogram matches very well the population true variogram. The fluctuations between the 11 variograms diminish somewhat when the sample size increases. However, similar to the sample variance (Table 1), there is not much reduction when increasing the sampling size from 100 to 200. Figures 11, 12, and 13 present the variogram for each sampling with the

Statistical Bias, Table 1 Average, variance, and range for the sample mean and the sample variance for 11 random samplings of different size (50, 100, 200) of the lognormal distribution. The last column provides the mean and variance for the entire population (size = 2500)

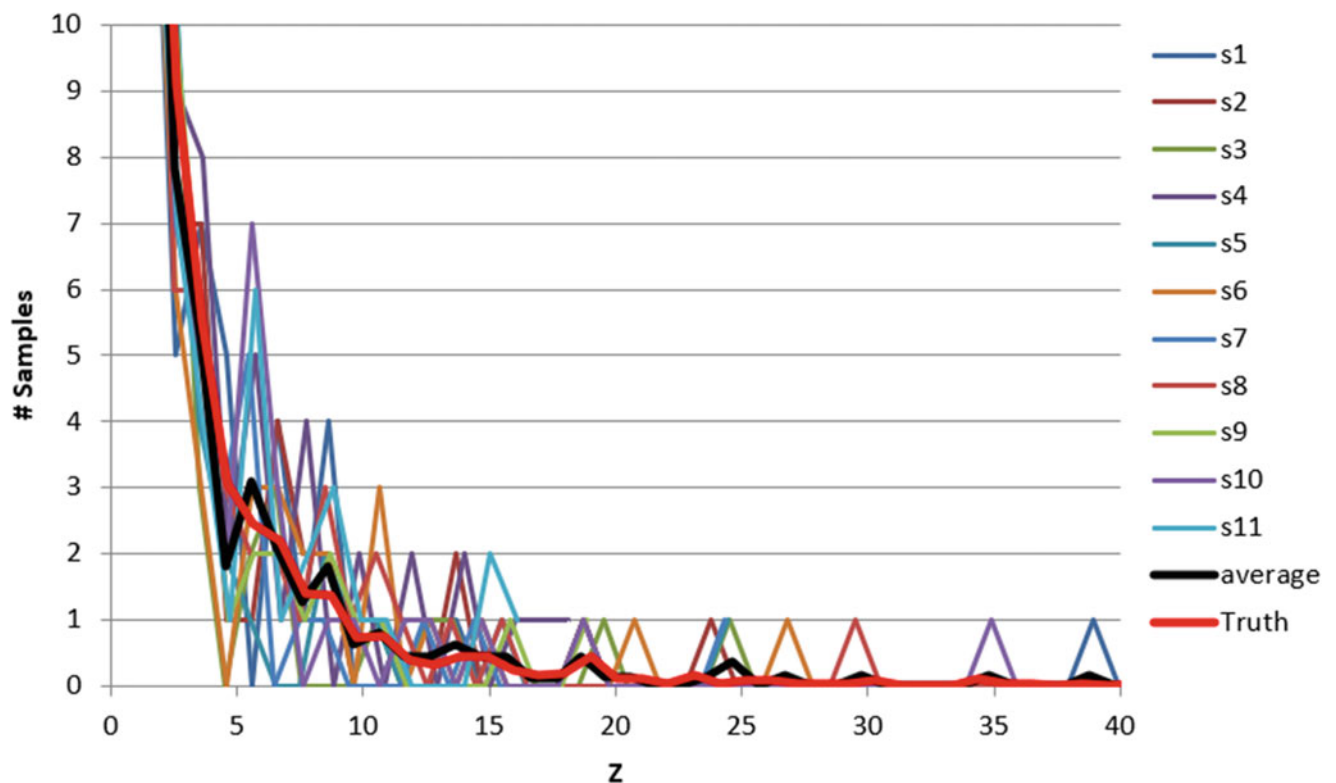
		$n = 50$	$n = 100$	$n = 200$	$n = 2500$
Mean	Average	2.60	2.57	2.68	2.58
	Variance	0.49	0.20	0.09	0
	Range	1.66–3.59	1.92–3.07	1.98–3.01	
Variance	Average	27.35	24.49	29.75	26.53
	Variance	478.60	150.78	182.62	0
	Range	6.55–69.48	9.98–40.44	9.09–63.45	

Statistical Bias, Table 2 Average, variance, and range for the sample mean and the sample variance for 11 random samplings of different size (50, 100, 200) of the normal distribution. The last column provides the mean and variance for the entire population (size = 2500)

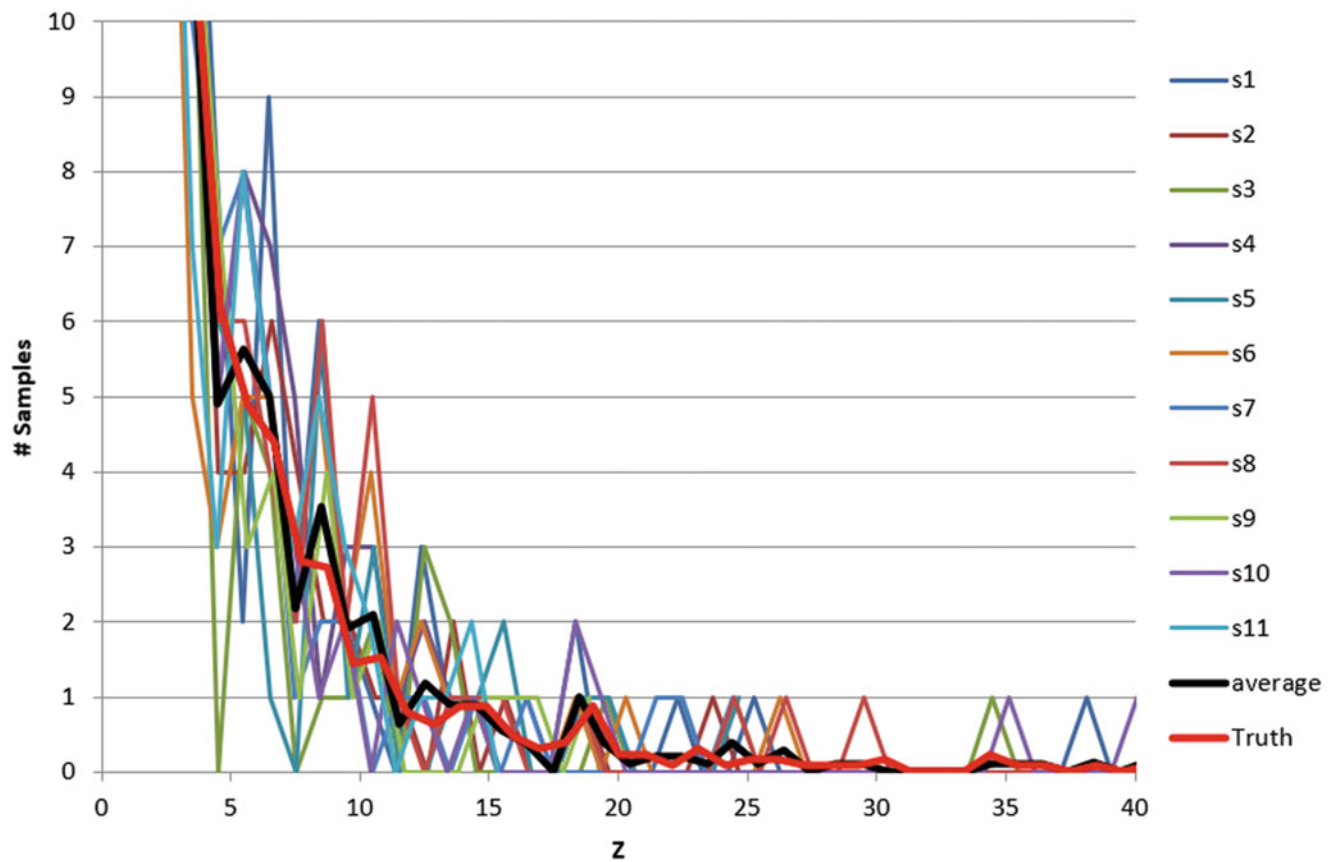
		$n = 50$	$n = 100$	$n = 200$	$n = 2500$
Mean	Average	2.28	2.46	2.60	2.58
	Variance	0.25	0.17	0.12	0
	Range	1.33–3.19	1.79–3.25	2.0–3.13	
Variance	Average	28.71	27.11	27.39	26.53
	Variance	47.07	18.62	4.85	0
	Range	18.46–38.85	20.02–33.77	24.17–31.08	



Statistical Bias, Fig. 2 Histograms for 11 samplings (s1–s11) of 50 sample values each for the lognormal population. The averaged histogram (in black) can be compared to the true histogram (in red)



Statistical Bias, Fig. 3 Histograms for 11 samplings (s1–s11) of 100 sample values each for the lognormal population. The averaged histogram (in black) can be compared to the true histogram (in red)



Statistical Bias, Fig. 4 Histograms for 11 samplings (s1–s11) of 200 sample values each for the lognormal population. The averaged histogram (in black) can be compared to the true histogram (in red)

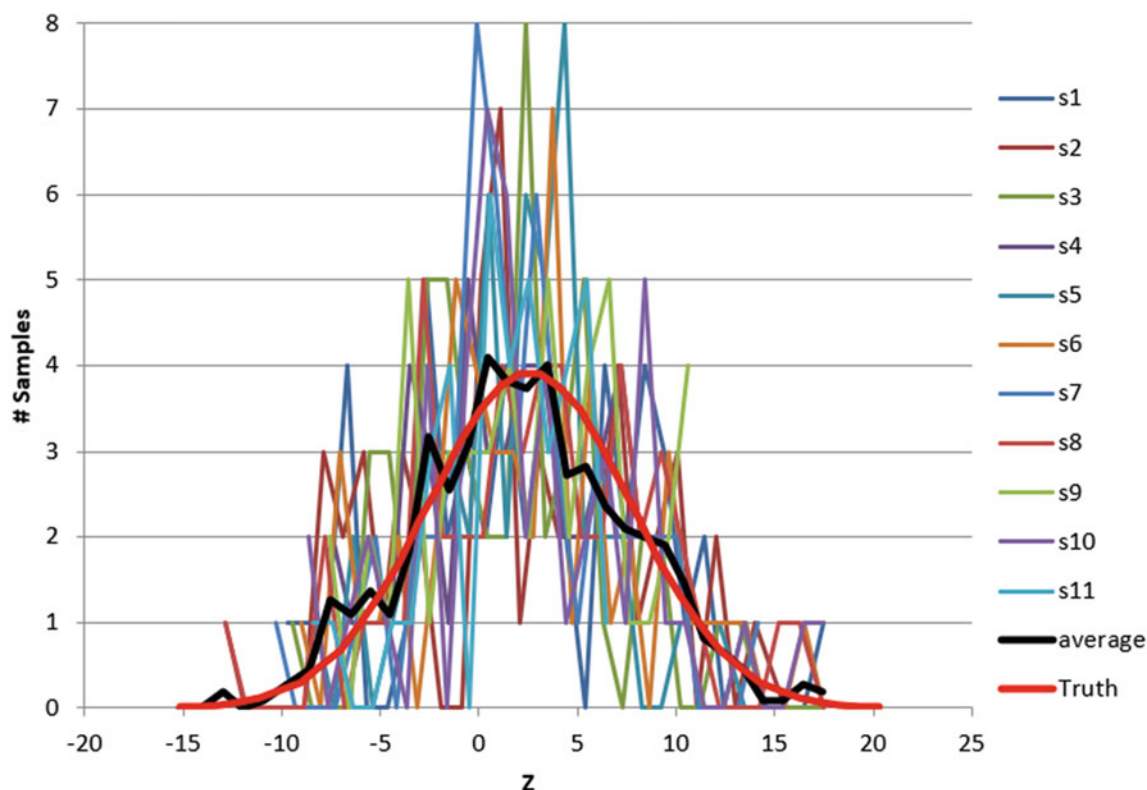
averaged variogram over the 11 samplings and the variogram for the normal population. For each sampling size, the averaged variogram matches very well the true variogram. Contrary to the lognormal population, and similarly to the sample variance of the normal population (Table 2), the fluctuations between the 11 variograms diminish significantly when the sample size increases.

Local Fluctuations for Spatial Estimation

For both populations, ordinary kriging was used for spatial estimation (Eq. 6) for all samplings. Figure 14 shows a cross-plot for the true values (z) versus their kriging estimates (z^*) for the sampling size of 50 in the lognormal population. Although the kriging estimates are not globally biased ($E[Z^*] = E[Z]$), they show a regression line with a slope smaller than 1. This is typical for kriging estimates (or any type of linear spatial averaging). The regression line with a slope smaller than 1 indicates that, on average over the field, the kriging estimates tend to over-estimate the true values when the kriging estimates are greater than their global mean, and they tend to underestimate the true values when

they are smaller than their global mean. In terms of conditional expectations (Eq. 9): $E[Z | Z^* = z^*] < z^*$ when $z^* > E[Z^*]$ and the opposite $E[Z | Z^* = z^*] > z^*$ when $z^* < E[Z^*]$. It is only when the kriging estimates are equal to their global mean that the conditional bias disappears as $E[Z | Z^* = E[Z^*]] = E[Z^*] (= E[Z])$ since kriging is globally unbiased). A spatial estimator such as kriging, achieves global unbiasedness at the price of conditional biases. Note that kriging is still the best spatial estimator. Other linear estimators will usually show larger conditional biases (Isaaks and Srivastava 1989). Bourgault (2021) has shown that the slope of the linear regression is 1 (45° line) when a spatial estimator is both, globally (Eq. 3) and conditionally unbiased (Eq. 9). Therefore, the slope value of the linear regression between the true values versus their kriging estimates is a good measure of the degree of conditional bias. Table 3 gives the average, variance, and range of the regression slopes over the 11 samplings for each of the 3 sampling sizes for the lognormal population. Table 4 gives the average, variance, and range of the regression slopes over the 11 samplings for each of the 3 sampling sizes for the normal population.

Results show significant statistical fluctuations between samplings, especially for the lognormal population. For both



Statistical Bias, Fig. 5 Histograms for 11 samplings (s1–s11) of 50 sample values each for the normal population. The averaged histogram (in black) can be compared to the true histogram (in red)

populations, the averaged regression slope increases closer to 1 with the sampling size. Thus, on average, conditional bias diminishes as the sampling size increases. But for all sampling sizes, the regression slope is on average larger, and its range and variance are smaller, for samplings in the normal population. The normal population is less prone to conditional bias than the lognormal population, even if the variance of the estimates is similar for both.

Spatial Correlation

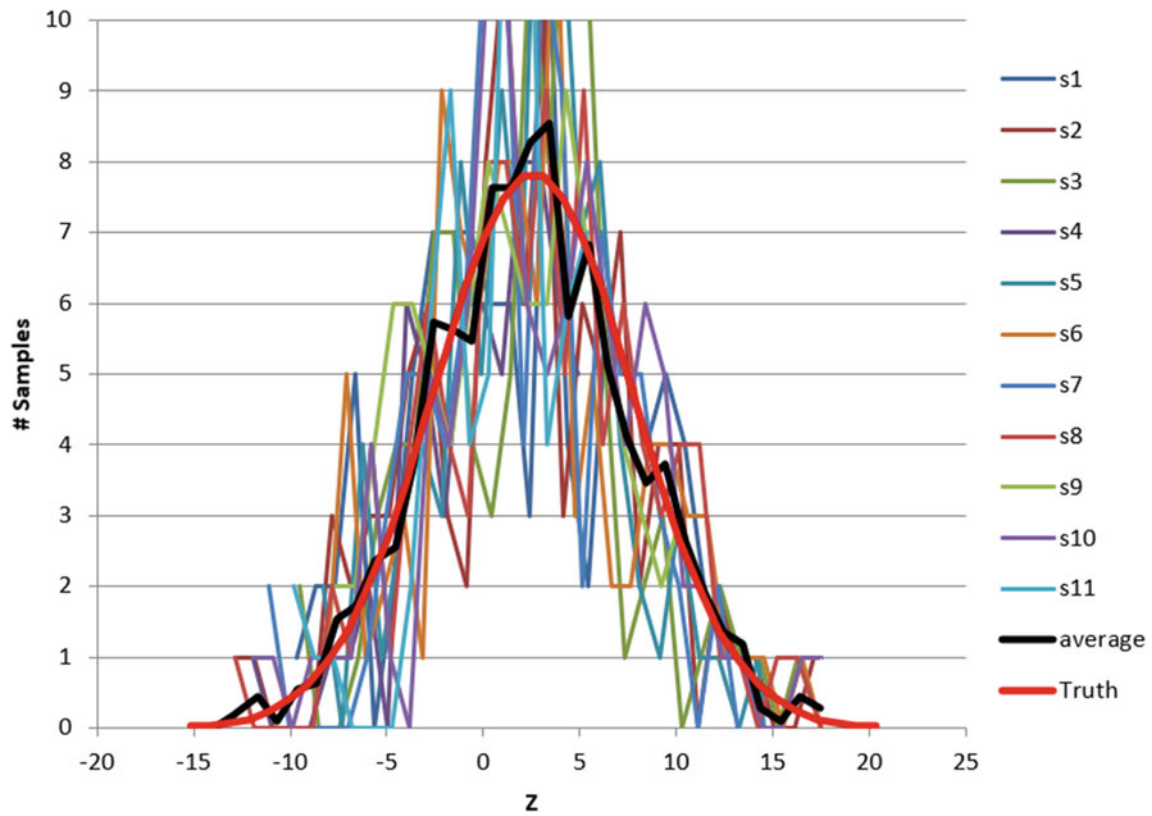
Figure 15 shows the variograms calculated from the 2300 kriging estimates, for each of the 11 samplings of size 200, in the lognormal population. Typically, the curves are lower (less variance), especially near the origin of the plot, when compared to the true variogram of the population. Also, the curves gradient near the origin of the plot becomes smaller than for the true variogram. This is characteristic of a smoother spatial correlation (or greater spatial continuity) (Chiles and Delfiner 2012). Smoothing is a common form of bias in the earth sciences in general and particularly with kriging. Contrary to the averaged variogram from the sample values (Fig. 10), the averaged variogram from the kriging estimates shows a clear bias when compared to the true

variogram. This estimation bias tends to be more severe for kriging estimates from the smaller sampling sizes of 50 and 100 (not shown).

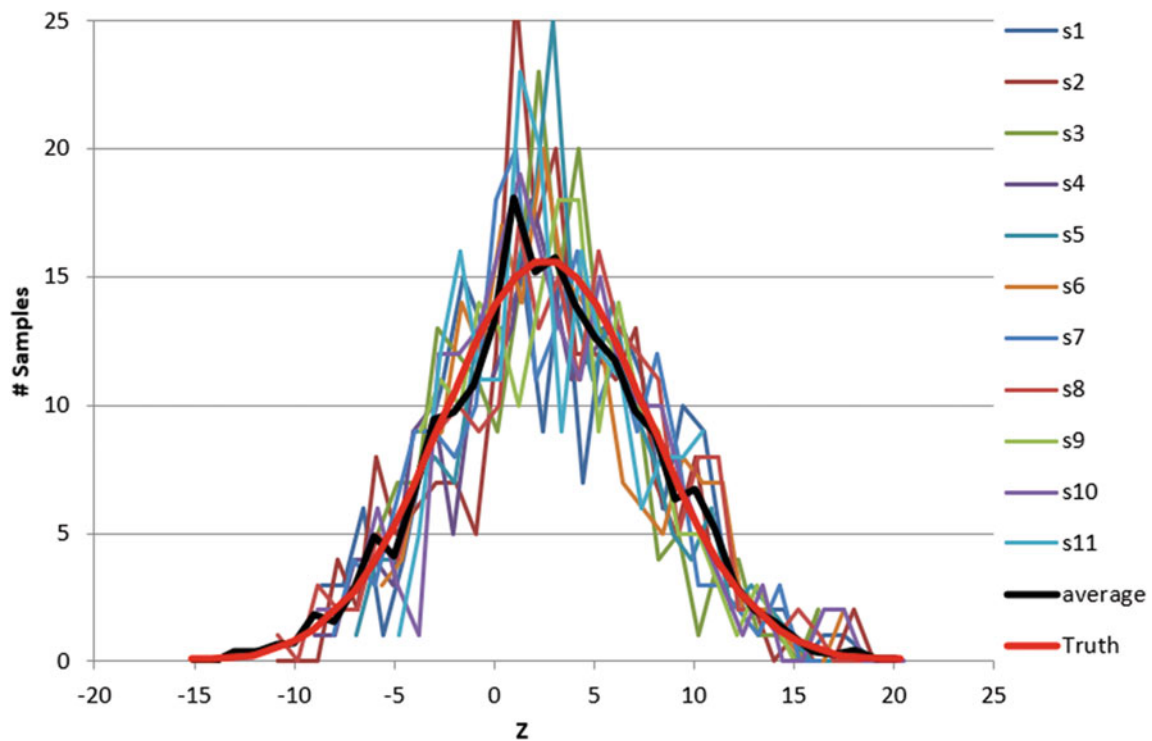
Global Fluctuations for Statistics from Truncated Distributions

Even, if kriging estimates are globally unbiased, and may show minimal conditional bias, another type of statistical bias is yet present. The expected value above a given threshold (t) will be underestimated due to the estimation bias (smoothing) of the kriging estimates (Isaaks 2004; Bourgault 2021). A global bias still exists when truncating the distribution for values above a threshold. This truncation is often required in the practice of mining as not all mineralized spatial locations in the population can be considered economical and should not all be considered for mineral recovery. As an example, Tables 5 and 6 give the expected value for values above the first quartile of the entire population.

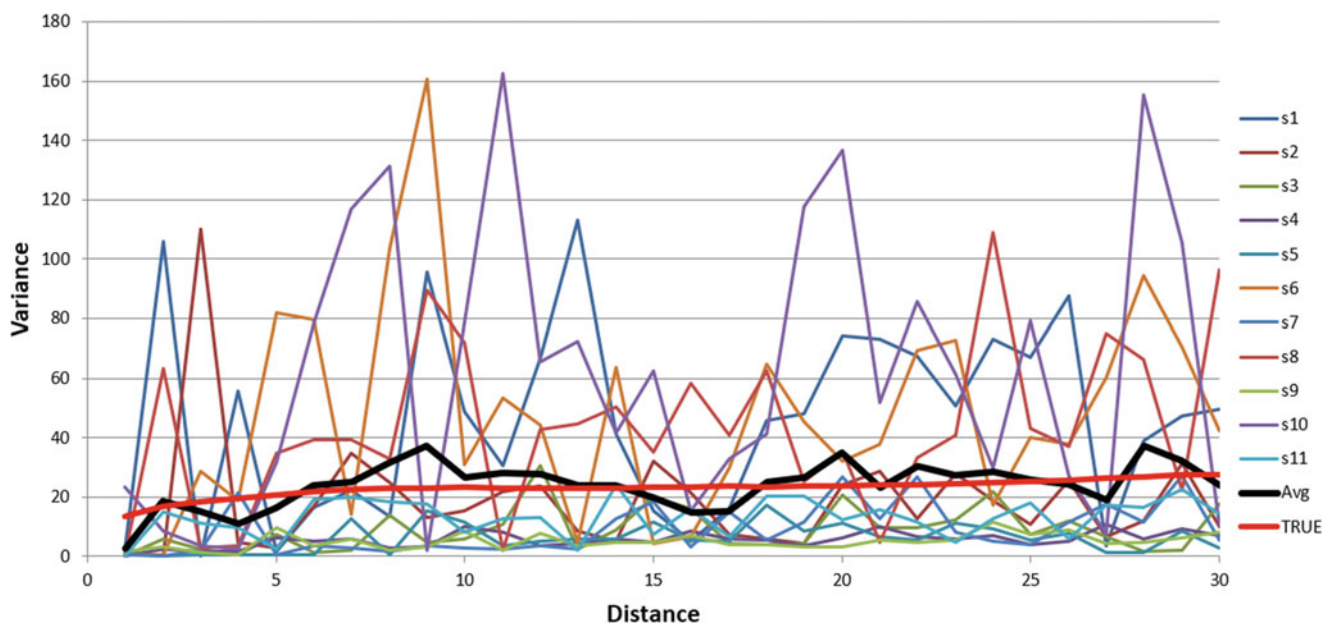
As with other statistics, results show significant statistical fluctuations between samplings for both populations. Increasing the sampling size tends to reduce the fluctuations. It is observed that the average expected value, for values above the first quartile over all samplings, is not biased for the sample



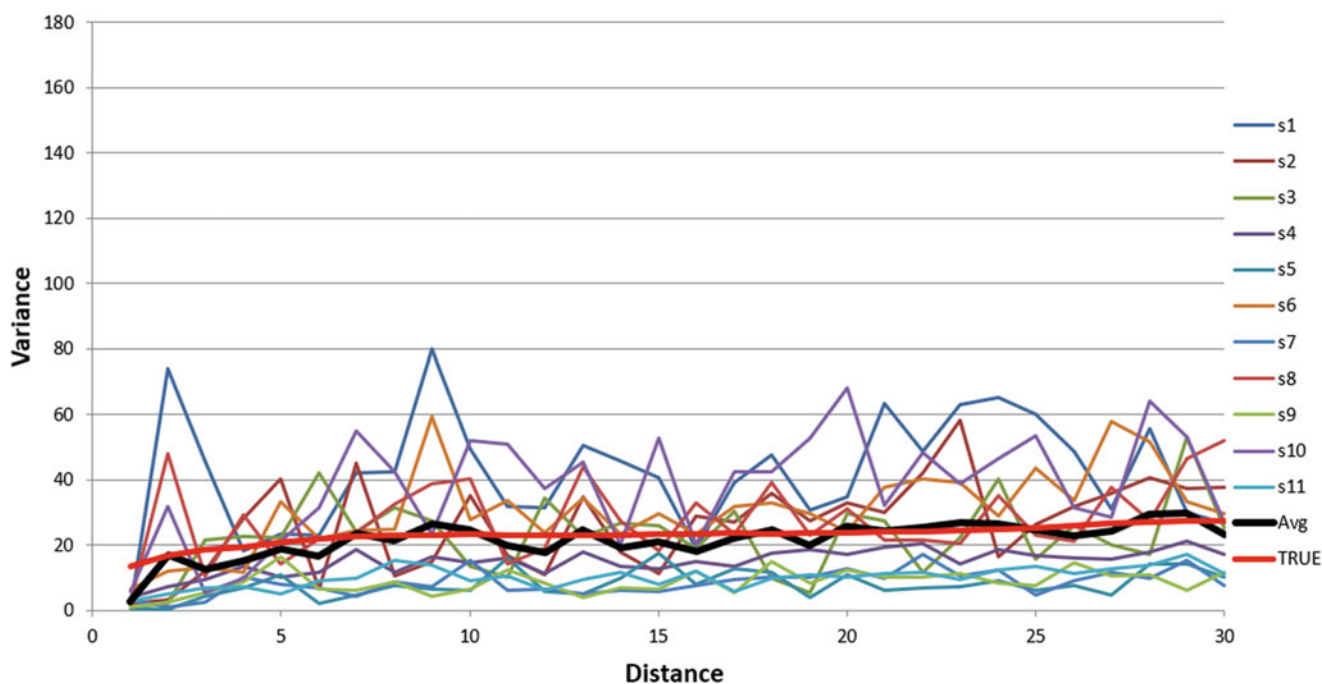
Statistical Bias, Fig. 6 Histograms for 11 samplings (s1–s11) of 100 sample values each for the normal population. The averaged histogram (in black) can be compared to the true histogram (in red)



Statistical Bias, Fig. 7 Histograms for 11 samplings (s1–s11) of 200 sample values each for the normal population. The averaged histogram (in black) can be compared to the true histogram (in red)



Statistical Bias, Fig. 8 Variograms for 11 samplings (s1–s11) of 50 sample values each for the lognormal population. The averaged variogram (in black) can be compared to the true variogram (in red)

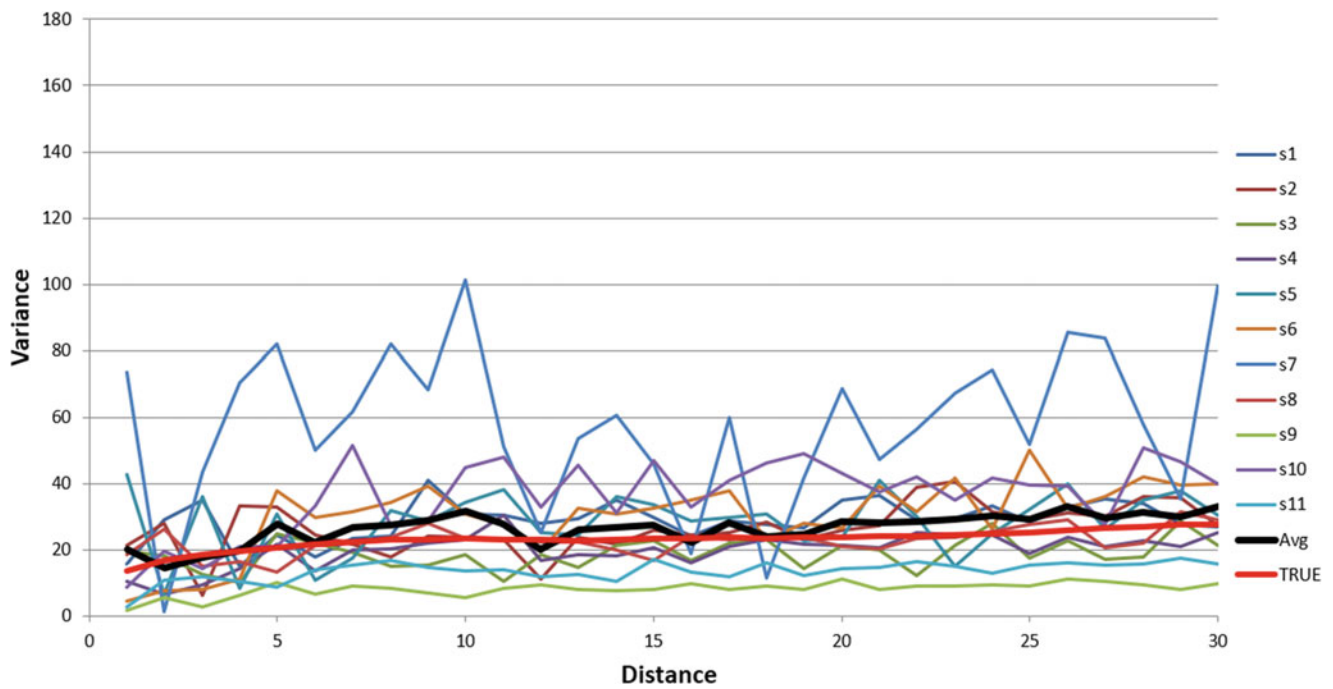


Statistical Bias, Fig. 9 Variograms for 11 samplings (s1–s11) of 100 sample values each for the lognormal population. The averaged variogram (in black) can be compared to the true variogram (in red)

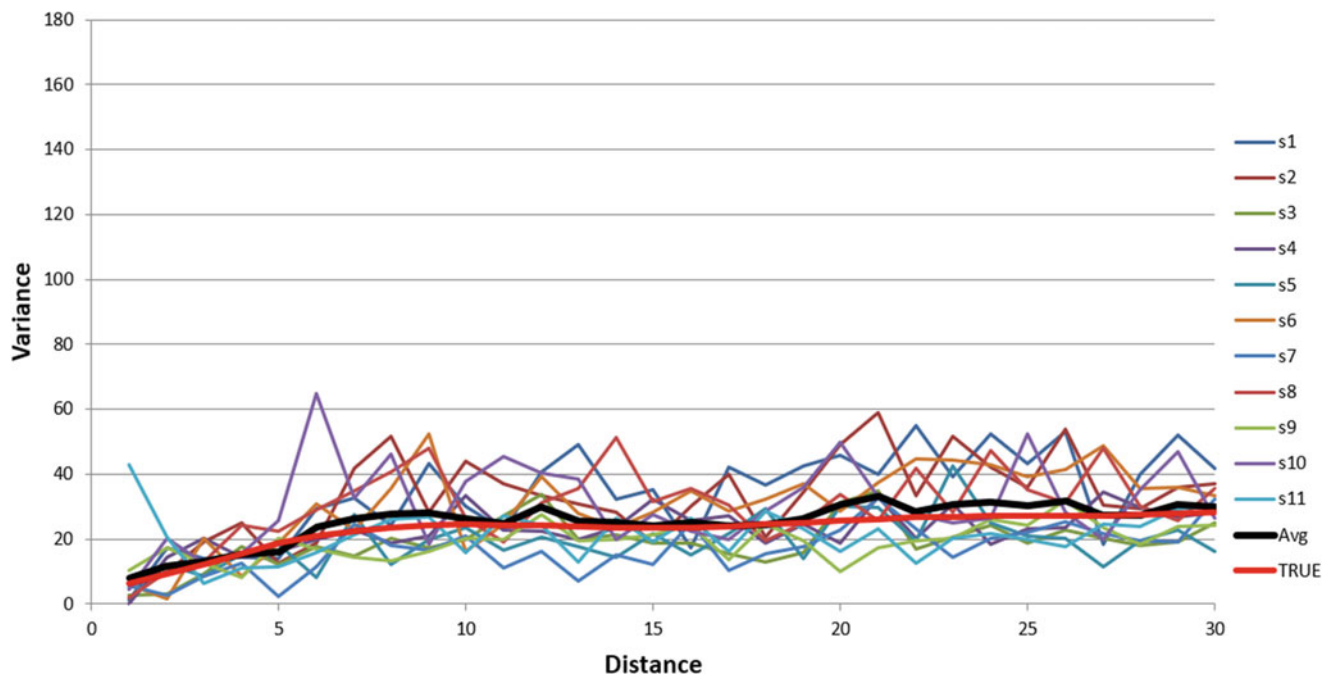
values (z). However, it is biased (smaller) for the estimated values by kriging (z^*), because the kriging estimates are smoother (less variance ranging from about 15 to 18 from Tables 3 and 4) than the population (26.53 from Tables 1 and 2) (see also Fig. 15).

Correcting Conditional Bias

Bourgault (2021) has shown that conditional bias could be corrected by averaging the kriging estimates over all the random samplings. Tables 7 and 8 provide the regression



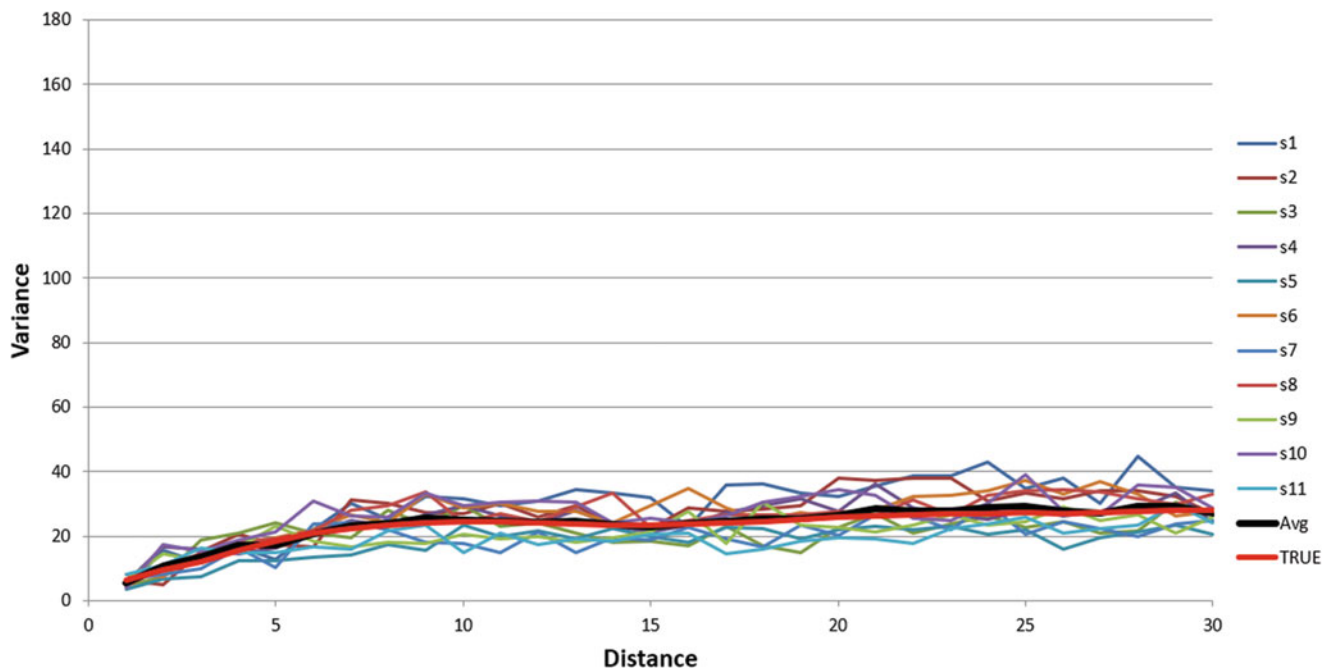
Statistical Bias, Fig. 10 Variograms for 11 samplings (s1–s11) of 200 sample values each for the lognormal population. The averaged variogram (in black) can be compared to the true variogram (in red)



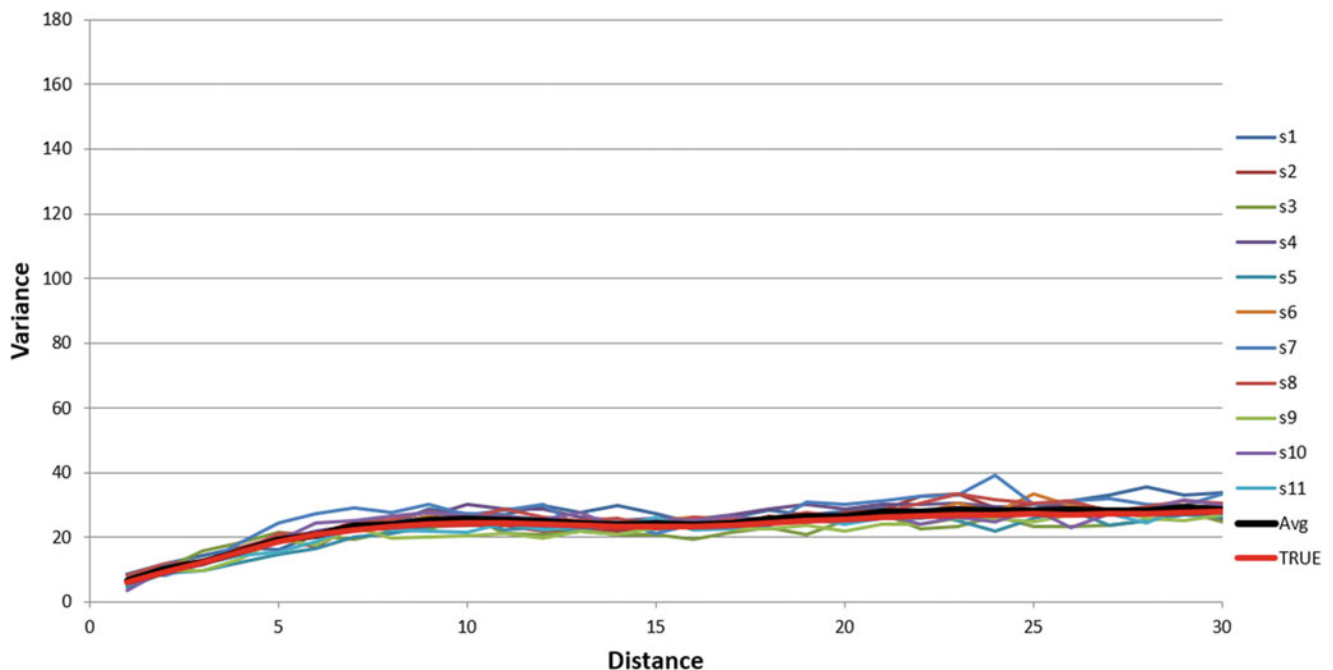
Statistical Bias, Fig. 11 Variograms for 11 samplings (s1–s11) of 50 sample values each for the normal population. The averaged variogram (in black) can be compared to the true variogram (in red)

slopes between the true values and the averaged kriging estimates ($\text{avg}(z^*)$), over the 11 random samplings, for each sample size.

From the results of Tables 7 and 8, it is observed that averaging the kriging estimates corrects well the conditional bias for the lognormal distribution but tends to overcorrect



Statistical Bias, Fig. 12 Variograms for 11 samplings (s1–s11) of 100 sample values each for the normal population. The averaged variogram (in black) can be compared to the true variogram (in red)

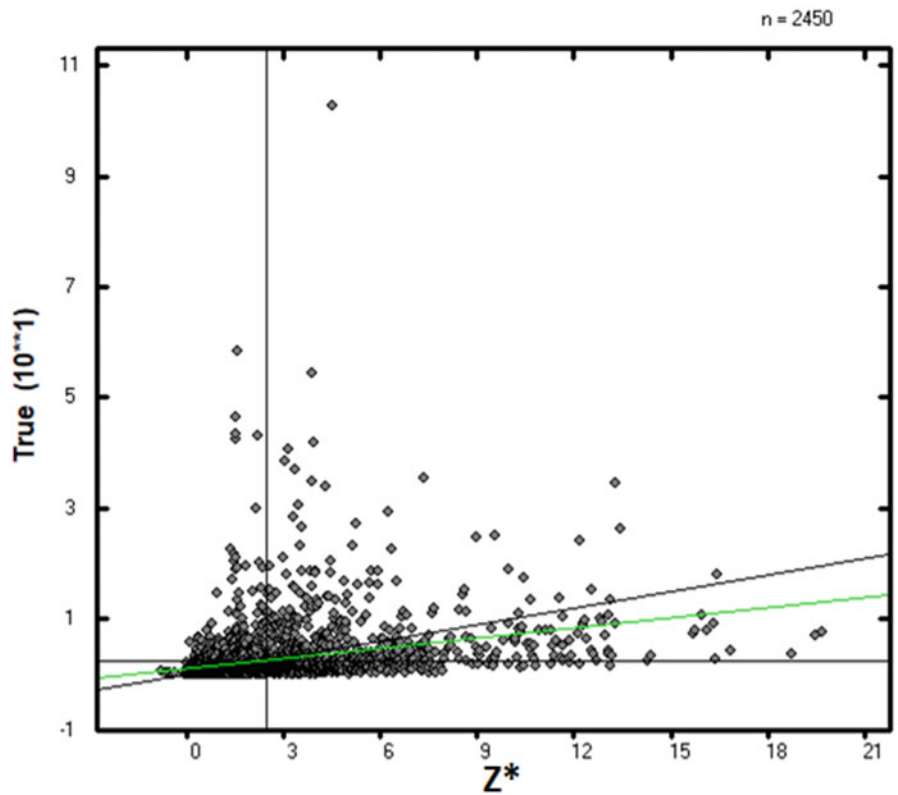


Statistical Bias, Fig. 13 Variograms for 11 samplings (s1–s11) of 200 sample values each for the normal population. The averaged variogram (in black) can be compared to the true variogram (in red)

(slopes larger than 1) for the normal distribution, especially for the smaller sampling sizes. In such a case, instead of averaging the kriging estimates, it is advised to pick the sampling that is associated with the smaller data variance.

For the normal distribution, the sample with the smallest data variance for sampling size 50 is associated with kriging estimates showing a regression slope of 0.84, 0.86 for sampling size of 100, and 0.91 for sampling size 200. Bourgault (2021)

Statistical Bias, Fig. 14 Cross-plot of true values versus kriging estimates (Z^*) using a sample size of 50 for the lognormal population. The regression line is shown in green, the 45° line is shown in black and the horizontal and vertical lines indicate the global means



Statistical Bias, Table 3 Average, variance, and range for the regression slope between the true values and their corresponding kriging estimates for 11 random samplings of different size (50, 100, 200) of the lognormal distribution. Last row provides the averaged variance for the (n^*) kriging estimates

		$n = 50$	$n = 100$	$n = 200$
Regression slope	Average	0.5	0.62	0.68
	Variance	0.06	0.04	0.04
	Range	0.19–0.9	0.35–0.95	0.36–0.97
Variance of estimates	Average	15.61 ($n^* = 2450$)	14.47 ($n^* = 2400$)	16.17 ($n^* = 2300$)

Statistical Bias, Table 4 Average, variance, and range for the regression slope between the true values and their corresponding kriging estimates for 11 random samplings of different size (50, 100, 200) of the normal distribution. Last row provides the averaged variance for the (n^*) kriging estimates

		$n = 50$	$n = 100$	$n = 200$
Regression slope	Average	0.71	0.82	0.90
	Variance	0.01	0.003	0.002
	Range	0.55–0.94	0.76–0.91	0.84–0.96
Variance of estimates	Average	15.28 ($n^* = 2450$)	16.5 ($n^* = 2400$)	18.57 ($n^* = 2300$)

has shown that, in general, samplings with a smaller data variance are associated with a regression slope closer to 1.0 for their kriging estimates. Comparing Table 7 and Table 3 shows that correcting for conditional bias is associated with an important variance reduction for the estimates (ranging from about 14 to 16 from Table 3 versus ranging from about 5 to 9 for the corrected kriging estimates Table 7).

For the lognormal population, the averaged kriging estimates ($\text{avg}(z^*)$) are not conditionally biased (Eq. 9, Table 7). Thus, they are the best estimates to categorize the spatial locations in terms of ore ($\text{avg}(z^*) > t$) or waste ($\text{avg}(z^*) < t$), for example. This brings back the topic of expected value for truncated distribution ($E[Z | Z > t]$). Table 9 shows

Statistical Bias, Table 5 Average, variance, and range for the average above the population first quartile (Q1) for the sample (z) and the (n*) estimated (z*) values of the population for 11 random samplings

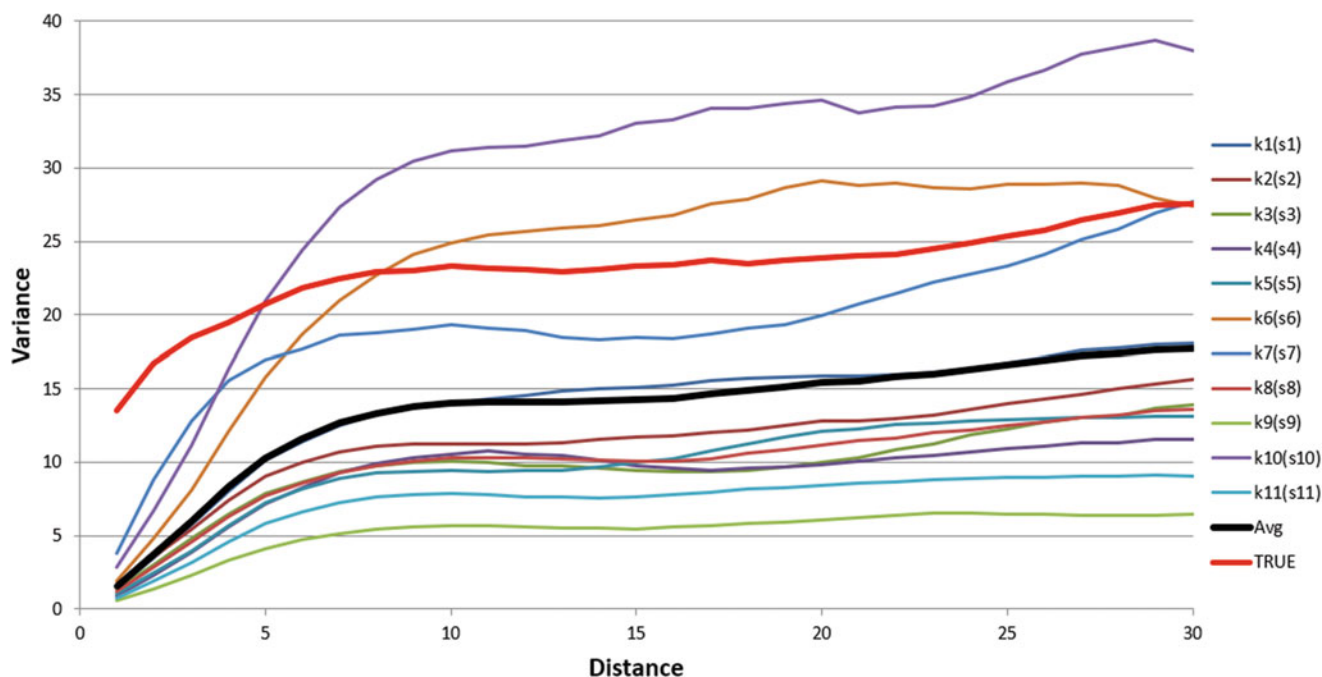
Q1 = 0.34		n = 50	n = 100	n = 200	n = 2500
E[Z Z > Q1]	Average	3.44	3.4	3.51	3.37
	Variance	1.13	0.41	0.13	0
	Range	1.95–5.2	2.39–4.55	2.76–4.07	
E[Z* Z* > Q1]		n* = 2450	n* = 2400	n* = 2300	n* = 0
	Average	3.16	3.0	3.13	
	Variance	0.81	0.28	0.15	
	Range	2.39–5.33	2.22–3.83	2.52–3.83	

of different size (50, 100, 200) of the lognormal distribution. The last column provides the statistic for the entire population

Statistical Bias, Table 6 Average, variance, and range for the average above the population first quartile (Q1) for the sample (z) and the (n*) estimated (z*) values of the population for 11 random samplings

Q1 = -0.89		n = 50	n = 100	n = 200	n = 2500
E[Z Z > Q1]	Average	4.72	4.73	4.81	4.77
	Variance	0.48	0.21	0.21	0
	Range	3.5–6.09	3.76–5.47	4.46–5.29	
E[Z* Z* > Q1]		n* = 2450	n* = 2400	n* = 2300	n* = 0
	Average	3.93	3.99	4.27	
	Variance	0.31	0.1	0.04	
	Range	3.19–5.3	3.29–4.33	3.83–4.57	

of different size (50, 100, 200) of the normal distribution. The last column provides the statistic for the entire population



Statistical Bias, Fig. 15 Variograms using the kriging estimates (k1–k11) from the 11 samplings (s1–s11) of sample size 200 in the lognormal population. The averaged variogram (in black) can be compared to the true variogram (in red)

their expected value for values above the first quartile of the true population.

From the results of Table 9, it is seen that $\text{avg}(z^*)$, over the 11 samplings, is not really conditionally biased for estimating $E[Z | \text{avg}(Z^*) > Q1]$. However, it is underestimating $E[Z | Z > Q1] = 3.37$ from the lognormal population. Even if $E[Z |$

$Z > Q1] = 3.37$ is fairly well estimated by the sample data values (Table 5), it is always underestimated by the kriging estimates, and even more so when they are corrected for conditional bias (Tables 5 and 9) to improve the selectivity for locations above or below the threshold (i.e., between ore and waste). Indeed, on average over the 11 samplings, avg

Statistical Bias, Table 7 Regression slope between the true values and their corresponding averaged kriging estimates for 11 random samplings of different size (50, 100, 200) of the lognormal distribution. Last row provides the variance for the averaged kriging estimates over the 2500 spatial locations

	$n = 50$	$n = 100$	$n = 200$
Regression slope	1.07	1.1	1.07
Variance of estimates	5.08	6.92	8.95

Statistical Bias, Table 8 Regression slope between the true values and their corresponding averaged kriging estimates for 11 random samplings of different size (50, 100, 200) of the normal distribution. Last row provides the variance for the averaged kriging estimates over the 2500 spatial locations

	$n = 50$	$n = 100$	$n = 200$
Regression slope	1.36	1.26	1.14
Variance of estimates	7.76	10.61	14.56

Statistical Bias, Table 9 Average above the population first quartile (Q1) for the averaged kriging estimates for 11 random samplings of different size (50, 100, 200) of the lognormal distribution. The second row provides the conditional expectation for the true values when their corresponding estimates are above the first quartile. The last row provides the averaged conditional expectation, over the 11 random samplings, of the true values when their corresponding kriging estimates are above the first quartile

Q1 = 0.34		$n = 50$	$n = 100$	$n = 200$
$E[\text{avg}(Z^*) \mid \text{avg}(Z^*) > Q1]$		2.87	2.71	2.93
$E[Z \mid \text{avg}(Z^*) > Q1]$		2.59	2.65	2.80
$\text{avg}(E[Z \mid Z^* > Q1])$	Mean	2.81	2.89	2.96
	Range	2.7–2.96	2.8–3.02	2.84–3.17

($E[Z \mid Z^* > Q1]$), for the kriging estimates, is always a bit closer to the true value (3.37) than $E[Z \mid \text{avg}(Z^*) > Q1]$ for the averaged kriging estimates. The underestimation by the averaged kriging estimates has its source in the lower variance for the conditionally unbiased estimates (compare variance of estimates in Tables 3 and 7).

Correcting for Estimation Bias

Some authors (McLennan and Deutsch 2002; Journel and Kyriakidis 2004) have proposed to estimate $E[Z \mid Z > t]$ using conditional simulations (Deutsch and Journel 1998) instead of kriging, since conditional simulations are designed to preserve the data variance in reproducing the sample histogram and variogram. It is a correction of the estimation bias by avoiding smoothing. Bourgault (2021) has shown, that indeed, conditional simulations (Zs) can be used if their resulting statistic ($E[Zs \mid Zs > t]$) is averaged over various samplings. Even if the simulated values do not suffer from the estimation bias, they still suffer from sampling bias just like

the sample histogram and variogram do. But in practice, and unlike the kriging estimates, or averaged kriging estimates, simulations cannot be used to categorize ore and waste on the spatial domain because the simulated values (Zs) cannot be definitively associated with any particular spatial locations as they change for each simulated realization. For recovering the minerals above an economical threshold, $E[Z \mid \text{avg}(Z^*) > t]$ (or $E[Z \mid Z^* > t]$ from the kriging estimates of a particular sampling) can be achieved, but $E[Z \mid Z > t]$ cannot. The statistical bias $E[Z \mid \text{avg}(Z^*) > t] < E[Z \mid Z > t]$ (or $E[Z \mid Z^* > t] < E[Z \mid Z > t]$) cannot be avoided. It can only be estimated, either from the sample values (Tables 5 and 6) or from conditional simulations.

Summary and Conclusions

Statistical estimation fluctuates, from sampling to sampling, leading to statistical bias associated with any given sampling. The fluctuations are more important when the sampling size is smaller, but also if the population has a skewed distribution like it is often the case in earth sciences. Sampling a lognormal population is more sensitive to statistical bias than sampling a normal population. Averaging the statistic of interest over multiple samplings is a proven approach for correcting the statistical bias for basic statistics (average, variance, histogram) and for spatial statistics (variogram). For spatial estimation, estimation bias will be added to the sampling bias. Averaging kriging estimates over multiple samplings ($\text{avg}(z^*)$) can correct for conditional bias, especially when the distribution is highly skewed. Thus, more smoothing, or more estimation bias, helps to reduce conditional bias. This will provide estimated values that will be unbiased for estimating $E[Z \mid \text{avg}(Z^*) > t]$ but will underestimate the true $E[Z \mid Z > t]$. Even if for many statistics their statistical bias can be practically eliminated by averaging them over multiple samplings, the processing of estimated values, or averaged estimated values, by a transfer function will in general provide biased results. For example, the application of a threshold value to mineral estimation will always suffer from some degree of statistical bias due to the smoothing of the estimator ($\text{variance}(\text{avg}(z^*)) \ll \text{variance}(z)$). This is the mining paradox where a variance reduction (smoothing) is required to remove conditional bias, such as to improve selectivity between ore and waste, but also induces a statistical bias in the expected recovery when the ore category is defined by grades above a threshold value.

Cross-References

- [Best Linear Unbiased Estimation](#)
- [Bootstrap](#)

- [Expectation-Maximization Algorithm](#)
- [Simulation](#)
- [Kriging](#)
- [Lognormal Distribution](#)
- [Normal Distribution](#)
- [Random Variable](#)
- [Regression](#)
- [Sequential Gaussian Simulation](#)
- [Smoothing Filter](#)
- [Spatial Statistics](#)
- [Variogram](#)

Bibliography

- Armstrong M (1998) Basic linear geostatistics. Springer, 153 p
- Bourgault G (2021) Clarifications and new insights on conditional bias. *Math Geosci* 53:623–654. <https://doi.org/10.1007/s11004-020-09853-6>
- Chiles JP, Delfiner P (2012) Geostatistics modeling spatial uncertainty. Wiley series in probability and statistics, 2nd edn. Wiley, 726 p
- Deutsch CV, Journel AG (1998) GSLIB Geostatistical software library and user's guide, 2nd edn. Oxford University Press, 369 p
- Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York, 483 p
- Isaaks E (2004) The kriging oxymoron: a conditionally unbiased and accurate predictor. In: Leuangthong O, Deutsch CV (eds) *Geostatistics Banff 2004*, 2nd edn. Springer, pp 363–374
- Isaaks EH, Srivastava RM (1989) An introduction to applied geostatistics. Oxford University Press, Oxford, 561 p
- Journel A, Kyriakidis P (2004) Evaluation of mineral reserves: a simulation approach. Oxford University Press, New York, 216 p
- Krige DG (1951) A statistical approach to some basic mine valuation problems on the Witwatersrand. *J Chem Metall Min Soc S Afr* 52: 119–139
- Matheron G (1965) Principles of geostatistics. *Econ Geol* 58: 1246–1266
- McLennan J, Deutsch C (2002) Conditional bias of geostatistical simulation for estimation of recoverable reserves. In: *CIM proceedings Vancouver 2002*, Vancouver

analysis, resampling methods for the estimation of confidence intervals or hypothesis testing, standard or robust linear regression, compositional data analysis, computational maximum likelihood estimation, nonlinear optimization, and multivariate tools such as principal component analysis. Statistical computing is especially important in the Earth sciences because data tend to be statistically complicated, rarely being described by simple distributions, and because of the proliferation of vast data sets as observing systems expand.

Introduction

It is assumed at the outset that readers of this entry are familiar with elementary probability and statistical concepts, such as set theory, the definition of probability, sample spaces, probability density and cumulative distributions, and measures of location, dispersion, and association, that are covered in the first few chapters of textbooks such as De Groot and Schervish (2011) or Chave (2017a). Some familiarity with standard distributions, such as the Gaussian and chi-square types, is also expected. Statistical computation is implemented in software packages such as Matlab that is widely used in the Earth and ocean sciences, and open source implementations include R and Python with third-party modules.

As noted in the Definition, statistical computing is a vast subject and cannot be covered comprehensively in anything short of a book. The treatment will be limited to exploratory data analysis tools, resampling methods including the bootstrap and permutation tests, linear regression, and nonlinear optimization to implement a maximum likelihood estimator. A more comprehensive treatment of computational data analysis may be found in Chave (2017a), where Matlab implementation is a focus, and at a more advanced level in Efron and Hastie (2016).

Statistical Computing

Alan D. Chave

Department of Applied Ocean Physics and Engineering, Deep Submergence Laboratory, Woods Hole Oceanographic Institution, Woods Hole, MA, USA

Definition

Statistical computing is a broad subject, covering all aspects of the analysis of data or statistical simulation that require the use of some type of computational engine. An incomplete list of statistical computing topics includes exploratory data

Exploratory Data Analysis Tools

Exploratory data analysis comprises the early stages in evaluating a given data set where the goal is their characterization. Let $\{X_i\}$ be a data sample of size N drawn from an unknown probability density function (pdf) $f(x)$. Partition the data range $\{\min(x_i), \max(x_i)\}$ into a set of r bins according to some rule. For example, the rule could partition the data into equal-sized intervals. Count the number of data in each bin. A frequency histogram is a plot of the counts of data in each bin against the midpoint of the bin. Other normalizations can be used, such as transforming counts into probabilities. The outcome is a simple way to characterize a given data set, but it remains

somewhat subjective because the bin sizes and their boundaries are arbitrary.

A better qualitative approach to characterize the distribution of $\{X_i\}$, which are assumed to be independent and identically distributed (iid), is the use of a kernel density estimator, which is a smoothed rather than binned representation of a distribution. The kernel density estimator that represents $f(x)$ is given by

$$\hat{f}(x) = \frac{1}{N\delta} \sum_{i=1}^N K\left(\frac{x - x_i}{\delta}\right) \quad (1)$$

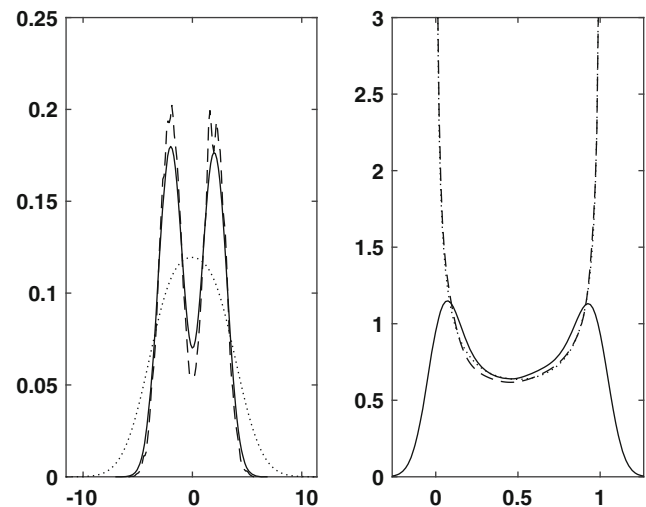
where $K(x)$ is a symmetric kernel function that integrates to one over its support and $\delta > 0$ is the smoothing bandwidth. A number of kernel functions are in use, but the standard Gaussian works well in practice. The smoothing bandwidth δ can be chosen by the user, but for a Gaussian kernel function the optimal value is

$$\delta = \left(\frac{4s_N^5}{3N} \right) \quad (2)$$

where $\hat{s}_N$ is the sample standard deviation. Figure 1a shows a kernel density pdf for 5000 random draws from Gaussian distributions with means of -2 and 2 and a common variance of 1 . The solid line uses (2) to give a smoothing bandwidth of 0.4971 and adequately resolve the peaks from the two Gaussian distributions. It also accurately reproduces the analytical form of the pdf. The dashed line shows the kernel density pdf for $\delta = 0.1$, which overestimates the peak heights and introduces multiple peaks at their top. The dotted line shows the kernel density pdf for $\delta = 2$, which smears the two peaks together.

In many instances, data may be drawn from a distribution whose support is finite, in contrast with the previous example. In this case, the kernel density estimator will give incorrect results unless the kernel density algorithm limits the support. Figure 1b shows a simple example using the arcsine distribution, which is defined only over $(0, 1)$. The arcsine distribution is equivalent to a beta distribution with arguments $(\frac{1}{2}, \frac{1}{2})$. The solid line shows the kernel density pdf when the support is unconstrained. By contrast, the dashed line shows it when the support is limited to $(0, 1)$ and is almost identical to the dotted line that shows the analytical form.

The quantile-quantile or q-q plot enables the assessment of the distribution of a set of data in a qualitative manner and can be made quantitative by adding error bounds based on the Kolmogorov-Smirnov statistic. A q-q plot consists of the quantiles of the target distribution $F^*(x)$ against the order statistics $x_{(i)}$ computed by sorting the sample. The quantiles are those values of the inverse cumulative distribution function at the uniform distribution quantiles:



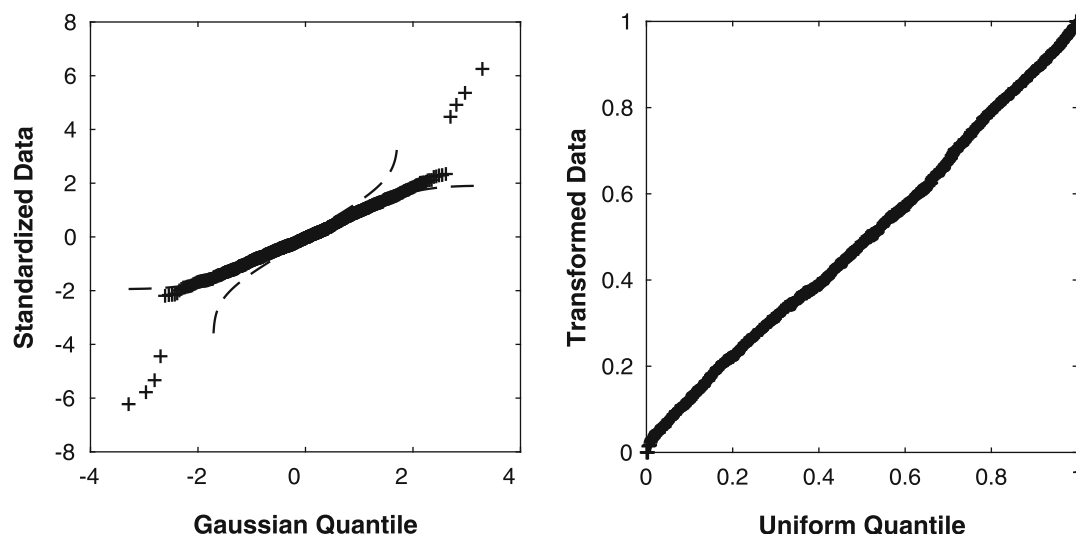
Statistical Computing, Fig. 1 (left) Kernel density pdf for 5000 random draws from Gaussian distributions with means of -2 and 2 and a variance of 1 . The solid line uses (2) for the smoothing bandwidth, while the dashed and dotted lines use values of 0.1 and 1 , respectively. (right) Kernel density pdf for 500 random draws from the arcsine distribution. The solid line assumes unbounded support, while the dashed line has support of $(0, 1)$. The dotted line is the analytic pdf

$$u_i = \frac{i - 0.5}{N} \quad i = 1, \dots, N \quad (3)$$

The order statistics obtain from sorting and optionally scaling the data. If the data sample is drawn from the target distribution, a q-q plot will be a straight line, while departures from a line yield insights into their actual distribution. A q-q plot emphasizes the target distribution tails and hence is useful for detecting outliers but is not sensitive near the distribution mode.

Figure 2a shows a q-q plot for 1000 random samples from a standard normal distribution but with the four smallest and largest values replaced by outlying values. The outliers are readily apparent at the bottom and top of the line. The dashed lines show confidence bounds obtained using the $1 - \alpha$ probability value of the Kolmogorov-Smirnov statistic c_α , plotting $F^{*-1}(u)$ against $F^{*-1}(u \mp c_\alpha)$ with $\alpha = 0.05$; see Chave (2017a, §7.2) for further details. Because the confidence bounds flare outward at the distribution extremes, the outliers are not sufficiently large to be quantitatively detected.

The percent-percent or p-p plot also enables the assessment of the distribution of a set of data in a qualitative manner and can be made quantitative by adding error bounds based on the Kolmogorov-Smirnov statistic. A p-p plot consists of the uniform quantiles (3) plotted against the order statistics $x_{(i)}$ of the data sample transformed using a target cdf, typically after standardizing the data. As with a q-q plot, a p-p plot will be a straight line if the target distribution is correct, and deviations from a straight line provide insight into the actual distribution of the data. However, a p-p plot is most sensitive near the



Statistical Computing, Fig. 2 (left) Q-q plot for 1000 random samples from a standard Gaussian distribution with the four smallest and largest values replaced by outliers, along with dashed lines showing the 95% confidence interval on the result. (right) P-p plot for the same data

distribution mode and hence is not very useful for outlier detection but is quite useful for data that are drawn from long-tailed distributions. Michael (1983) introduced a useful variant on the p-p plot that equalizes its variance at all points.

Figure 2b shows a p-p plot of the contaminated Gaussian sample from Fig. 2a. The outliers are barely visible as a slight turn to the left and right at the bottom and top of the line. Confidence bounds are omitted but can easily be added.

The outcome of a statistical procedure may require the identification and elimination of outlying data. This is called censoring and requires changes to both q-q and p-p plots in its presence, as the target distribution has to be truncated. Let $f_x(x)$ and $F_x(x)$ be the pdf and cumulative distribution function (cdf), respectively, of the data before truncation. The pdf of the truncated data set is

$$f'_x(x') = \frac{f_x(x)}{F_x(d) - F_x(c)} \quad (4)$$

where $c \leq x' \leq d$. Suppose the original number of data is N , and the number of data censored at the bottom and top of the distribution are m_1 and m_2 , respectively. Suitable choices for c and d are the m_1 th and $N - m_2$ th quantiles of the original distribution $f_x(x)$. A q-q or p-p plot computed using (4) as the target distribution against the censored data will yield the correct result.

Exploratory data analysis sometimes requires the computation of combinations of distributions or understanding of the properties of specific distributions that cannot be obtained analytically. A suitable substitute is simulation. An example appears in Chave (2014), where it was postulated that the residuals obtained from magnetotelluric (MT) response function estimation are pervasively distributed as a member of the

stable family (a set of distributions characterized by infinite variance and sometimes mean and which describe phenomena in fields ranging from physics to finance). Stable distributions have four parameters (tail thickness α , skewness, location, and scale), and while there are known relationships for the latter three parameters for ensembles of stable random variables, these do not exist for α . Chave (2014) used simulation to show that ensembles of stable random variables with different tail thickness parameters converge to a stable distribution with a new value for α . He did this by obtaining 10,000 random draws from a uniform distribution over the range $[0.6, 2.0]$ to specify α in random draws from a standardized symmetric stable distribution. The ensemble was fit by a stable distribution with a tail thickness parameter of 1.16.

Resampling Methods

Beginning in the early twentieth century, a variety of parametric (meaning that they are based on some distributional assumptions) estimators for the mean and variance (among other statistics) and for hypothesis testing were established. In many instances, a data analyst will not want to make assumptions about the data distribution required for parametric estimators due to the possibility of mixture distribution, outliers, and non-centrality. This led to the development of estimators based on resampling of the data as computational power grew. The most widely used resampling estimator is the bootstrap introduced by Efron (1979). More recently, permutation hypothesis tests that typically are more powerful than the bootstrap have been developed.

The empirical distribution based on the data sample plays a central role in the bootstrap, taking the place of a parametric

sampling distribution. The bootstrap utilizes sampling with replacement (meaning that a given sample may be obtained more than once) and hence is not exact. It can be used for parameter estimation, confidence interval estimation, and hypothesis testing. Suppose that the statistic of interest is $\hat{\lambda}_N$ along with its standard error. Obtain B bootstrap samples $\mathbf{X}_k^*$ from the data set, each of which contains N samples, by resampling with replacement. Apply the estimator for $\hat{\lambda}_N$ to each bootstrap sample to get a set of bootstrap replicates $\hat{\lambda}_N^*(\mathbf{X}_k^*)$ for $k = 1, \dots, B$. The sample mean of the bootstrap replicates is

$$\bar{\lambda}^* = \frac{1}{B} \sum_{k=1}^B \hat{\lambda}_N^*(\mathbf{X}_k^*) \quad (5)$$

and the standard error on $\hat{\lambda}_N$ is

$$SE_B(\hat{\lambda}_N) = \left\{ \frac{1}{B} \sum_{k=1}^B [\hat{\lambda}_N^*(\mathbf{X}_k^*) - \bar{\lambda}^*]^2 \right\}^{1/2} \quad (6)$$

The same methodology applies to other common statistics such as the skewness or kurtosis. While statistical lore holds that a couple of hundred replicates should suffice, it is better to use many thousands and examine the bootstrap distribution for consistency with a given statistic; see Chave (2017a, §8.2) for the details. Note also that a naïve bootstrap applied to estimators derived from the order statistics (e.g., the interquartile range) will fail to converge.

The bootstrap is most commonly used to obtain confidence intervals when the statistic under study is complicated such that its sampling distribution is either unknown or difficult to work with. The standard bootstrap confidence intervals are computed by simply substituting bootstrap estimates for the statistic and its standard deviation into the Student t confidence interval formula, yielding

$$\Pr \left[\bar{\lambda}^* - t_{N-1} \left(1 - \frac{\alpha}{2} \right) SE_B(\hat{\lambda}_N) \leq \lambda \leq \bar{\lambda}^* + t_{N-1} \left(1 - \frac{\alpha}{2} \right) SE_B(\hat{\lambda}_N) \right] = 1 - \alpha \quad (7)$$

where $\bar{\lambda}^*$ and $SE_B(\hat{\lambda}_N)$ are given by (5) and (6).

A better approach to the naïve formulation (7) is the bootstrap-t confidence interval based on the bootstrap replicates given by the normalized statistic:

$$z_k^* = \frac{\hat{\lambda}_N^* - \bar{\lambda}^*}{SE_B(\hat{\lambda}_N)} \quad (8)$$

The $\{z_k^*\}$ are then sorted, and the $\alpha/2$ and $1 - \alpha/2$ values are extracted for use in (7) in place of the t_{N-1} quantile. Accurate

results require the use of a large (many thousands) number of bootstrap replicates. The bootstrap percentile method directly utilizes the $\alpha/2$ and $1 - \alpha/2$ quantiles of the bootstrap distribution and has better convergence and stability properties than either the standard or bootstrap-t approaches.

The best bootstrap confidence interval estimator is the bias-corrected and accelerated (BCa) approach that adjusts the percentile method to correct for bias and skewness. The formulas for bias correction and acceleration are complicated and will be omitted; see Chave (2017a, §8.2.3) for details. It is the preferred approach for most applications.

The bootstrap is also useful for hypothesis testing, although the results typically are neither exact (meaning that the probability of a type 1 error for a composite hypothesis is α for all of the possibilities which make up the hypothesis) nor conservative (meaning that the type 1 error never exceeds α); an exact test is always conservative. A better alternative for most applications is the permutation test that was introduced by Pitman (1937) but only became practical in recent years due to its computational complexity. Permutation tests are usually applied to the comparison of two populations, although they can be used with a single data set with some additional assumptions. A permutation test is at least as powerful as the alternative for large samples and is nearly distribution-free, requiring only very simple assumptions about the population. Permutation tests are based on sampling without replacement. A sufficient condition for a permutation test to be exact and unbiased is exchangeability of the observations (meaning that any finite permutation of their indices does not change their distribution) in a combined sample.

Hypothesis testing usually utilizes the p-value to make decisions. A p-value is the probability of observing a value of the test statistic as large as or larger than the one that is observed for a given experiment. A small number is taken as evidence for the alternate hypothesis or rejection of the null hypothesis. The definition of small is subjective, but typically a p-value below 0.05 is taken as strong evidence for rejection of the null hypothesis. A p-value is a random variable and should not be confused with the type 1 error α which is a parameter chosen before an experiment is performed and hence applies to an ensemble of experiments. The p-value concept is not without controversy, and the American Statistical Association issued a statement on its concept, context, and purpose in Wasserstein and Lazar (2016) that is certainly pertinent.

The key steps in carrying out a permutation test on two populations $\mathbf{X}$ and $\mathbf{Y}$ are:

1. Choose a test statistic $\hat{\lambda}$ that measures the difference between $\mathbf{X}$ and $\mathbf{Y}$.
2. Compute the value of the test statistic for the original sets of data.

3. Combine $\mathbf{X}$ and $\mathbf{Y}$ into a pooled data set.
4. Compute the permutation distribution of $\hat{\lambda}$ by randomly permuting the pooled data, ensuring that the values are unique to enforce sampling without replacement, and recompute the test statistic $\hat{\lambda}_k$.
5. Repeat step 4 many (typically tens of thousands to millions) of time.

6. Accept or reject the null hypothesis according to the p-value for the original test statistic compared with the permutation test statistics.

Let B be the number of permutations. The two-sided equal tail permutation p-value is given by

$$p_{perm} = 2 \times \min \left\{ \frac{1}{B+1} \left[\sum_{i=1}^B \mathbf{1}(\hat{\lambda}_k^* \geq \hat{\lambda}) + 1 \right], \frac{1}{B+1} \left[\sum_{i=1}^B \mathbf{1}(\hat{\lambda}_k^* < \hat{\lambda}) + 1 \right] \right\} \quad (9)$$

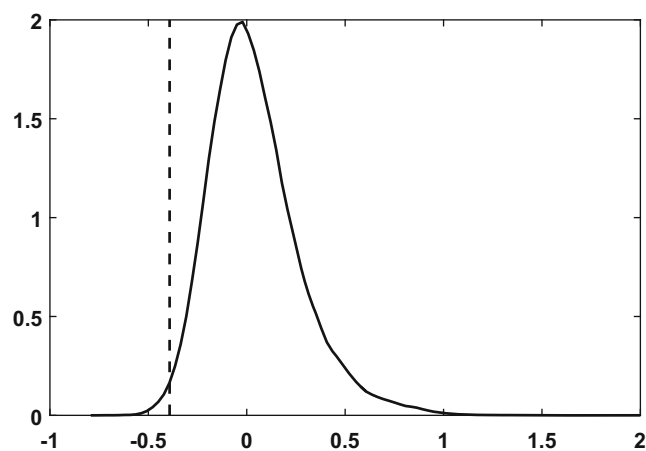
where $\mathbf{1}(x)$ is an indicator function that is either 0 or 1 as x is negative or positive. An example taken from Chave (2017a, §8.3.3) serves to illustrate a permutation test. In the 1950s and 1960s, several countries carried out experiments to see if cloud seeding would increase the amount of rain. Simpson et al. (1975) provide an exemplar data set containing 52 clouds, half of which were unseeded or seeded. Q-q plots of the data show that they are quite non-Gaussian, and a q-q plot of the logarithms of the data is quite linear. A two-sample t test on the data yields a p-value of 0.0511, which constitutes weak acceptance of the null hypothesis that unseeded and seeded clouds yield different amounts of rain. However, a two-sample t test on the logs of the data yields a p-value of 0.0141, strongly rejecting the null hypothesis. A permutation test will be used to test the null hypothesis using 1,000,000 permutations of the merged data set. The test statistic is the difference of the means of the first and last 26 entries in the permuted merged data. A set of 1,000,000 random permutations of the 52 data indices is generated and checked to ensure that none of them constitutes the original data order and that they are unique. The p-value (9) is 0.0441, rejecting the null

hypothesis. Repeating the permutation test using the logs of the data gives a p-value of 0.0143, again strongly rejecting the null hypothesis and showing that a permutation test does not require strong distributional assumptions, in contrast to the parametric t test. Figure 3 shows the permutation distribution compared to a Gaussian distribution using sample mean and variance as its parameters for reference and the value of the test statistic.

Permutation tests can be used for one sample tests if the underlying data distribution is symmetric. It can also be applied to two sample tests on paired data such as come from medical trials, where patients are assigned at random to control and treated groups, and two sample tests for dispersion. Care is required in devising a test statistic for the latter unless the means of the two data sets are known to be the same. The solution is to first center each data set about its median value and then use the difference of the sum of squares or absolute values as the test statistic. Details on these applications may be found in Chave (2017a, §8.3).

A final application of resampling is bias correction in goodness-of-fit tests such as the Kolmogorov-Smirnov or Anderson-Darling types. It is well known that such tests require that the parameters in the target distribution be known a priori, a condition that rarely holds in practice, and they are usually estimated from the data. This results in unknown bias in the test which can be removed using a Monte Carlo approach. The steps to follow are:

1. Obtain the test statistic for the N observations against a target distribution whose parameters are estimated from the data.
2. Compute N random draws with replacement from the target distribution using the same parameters.
3. Obtain the K-S or A-D statistic for the random draw.
4. Repeat steps 2 and 3 a large number of times.
5. Compute the p-value using (9).



Statistical Computing, Fig. 3 Kernel density pdf of the distribution of the test statistic for the cloud seeding data. The vertical dashed line is the test statistic based on all of the data. (Taken from Chave (2017a))

This approach was used by Chave (2017b) to remove bias while assessing the fit of a stable distribution to regression

residuals in the computation of the magnetotelluric response function.

Linear Regression

Linear regression is the most widely used statistical procedure in existence. It encompasses problems that can be cast as a linear model for p parameters given N data, where $p \leq N$ is assumed

$$\mathbf{y} = \leftrightarrow \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (10)$$

where $\mathbf{y}$ is the response N -vector, $\leftrightarrow \mathbf{X}$ is an $N \times p$ matrix of predictors, $\cdot$ denotes the inner product, $\boldsymbol{\beta}$ is a parameter p -vector, and $\boldsymbol{\epsilon}$ is an N -vector of unobservable random errors. Four assumptions underlie linear regression: (a) the model is linear in the parameters, which is implied by (10); (b) the error structure is additive, which is also implied by (10); (c) the random errors have zero mean and equal variance while being mutually uncorrelated; and (d) the rank of the predictor matrix is p .

In elementary statistical texts, the predictor variables are presumed to have no measurement error, which is rarely true in practice. In real applications, $\leftrightarrow \mathbf{X}$ contains random variables, in which case the statistical model underlying (10) is

$$\begin{aligned} E(\mathbf{y} | \leftrightarrow \mathbf{X}) &= \leftrightarrow \mathbf{X} \cdot \boldsymbol{\beta} \\ \text{cov}(\mathbf{y} | \leftrightarrow \mathbf{X}) &= \sigma^2 \leftrightarrow \mathbf{I}_N \end{aligned} \quad (11)$$

where the leading elements are the conditional expected value and covariance, σ^2 is the population variance, and $\leftrightarrow \mathbf{I}_N$ is the $N \times N$ identity matrix. The least squares estimator for (10) is

$$\hat{\boldsymbol{\beta}} = (\leftrightarrow \mathbf{X}^H \cdot \leftrightarrow \mathbf{X})^{-1} \cdot \leftrightarrow \mathbf{X}^H \cdot \mathbf{y} \quad (12)$$

where the superscript H denotes the Hermitian transpose to accommodate complex response and predictor variables.

A list of the properties of the linear regression estimator is given in Chave (2017a, §9.2). A summary is:

1. The unconditional expected value of (12) is

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \quad (13)$$

so that linear regression is unbiased presuming assumption (c) holds.

2. The unconditional covariance matrix for $\hat{\boldsymbol{\beta}}$ is

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 E((\leftrightarrow \mathbf{X}^H \cdot \leftrightarrow \mathbf{X})^{-1}) \quad (14)$$

and in practice the expected value is replaced with an estimate.

3. The predicted value N -vector is

$$\hat{\mathbf{y}} = \leftrightarrow \mathbf{X} \cdot (\leftrightarrow \mathbf{X}^H \cdot \leftrightarrow \mathbf{X})^{-1} \cdot \leftrightarrow \mathbf{X}^H \cdot \mathbf{y} \equiv \leftrightarrow \mathbf{H} \cdot \mathbf{y} \quad (15)$$

where $\leftrightarrow \mathbf{H}$ is the $N \times N$ hat matrix. The hat matrix is a projection matrix and hence is symmetric and idempotent. Its diagonal elements $\hat{h}_{ii}$ are real and satisfy $0 \leq \hat{h}_{ii} \leq 1$. Chave and Thomson (2003) showed that when the rows of $\leftrightarrow \mathbf{X}$ are complex multivariate Gaussian, the distribution of $\hat{h}_{ii}$ is beta with parameters p and $N - p$.

4. The estimated regression residuals are given by

$$\hat{\mathbf{r}} = \mathbf{y} - \leftrightarrow \mathbf{X} \cdot \hat{\boldsymbol{\beta}} = (\leftrightarrow \mathbf{I}_N - \leftrightarrow \mathbf{H}) \cdot \mathbf{y} \quad (16)$$

5. The sum of squared residuals is

$$\hat{\mathbf{r}}^H \cdot \hat{\mathbf{r}} = \mathbf{y}^H \cdot (\leftrightarrow \mathbf{I}_N - \leftrightarrow \mathbf{H}) \cdot \mathbf{y} \quad (17)$$

6. An unbiased estimate for the variance is

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{r}}^H \cdot \hat{\mathbf{r}}}{N - p} \quad (18)$$

7. The regression residuals $\hat{\mathbf{r}}$ are uncorrelated with the predicted values $\hat{\mathbf{y}}$.
8. The parameter vector $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (Gauss-Markov theorem).
9. Statements 1–8 do not require any distributional assumptions. If the random errors are real and N -variate Gaussian distributed with mean μ and common variance σ^2 , then $\hat{\boldsymbol{\beta}}$ is also the maximum likelihood estimator and hence $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) / \hat{\sigma}$ converges in distribution to a standardized Gaussian distribution. Further,

$$\begin{aligned} \hat{\boldsymbol{\beta}} &\sim N_p(\boldsymbol{\beta}, \sigma^2 (\leftrightarrow \mathbf{X}^H \cdot \leftrightarrow \mathbf{X})^{-1}) \\ \frac{(N - p) \hat{\sigma}^2}{\sigma^2} &\sim \chi_{N-p}^2 \end{aligned} \quad (19)$$

where $\sim$ means “is distributed as” and N_p is a p -variate Gaussian. In addition, the regression estimator is consistent, and $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent and serve as sufficient statistics for estimating $\boldsymbol{\beta}$ and σ^2 . Complex random errors require that propriety also be considered; see Schreier and Scharf (2010) for the details.

Equation (12) is called the normal equation solution and possesses poor numerical properties, particularly when the

predictor matrix has entries that cover a wide range. The problem comes from inverting $\leftrightarrow \mathbf{X}^H \cdot \leftrightarrow \mathbf{X}$, and the issue can be avoided if that step is removed. The predictor matrix can always be factored into the product of two matrices $\leftrightarrow \mathbf{Q} \cdot \leftrightarrow \mathbf{R}$, where $\leftrightarrow \mathbf{Q}$ is $N \times p$ orthonormal, so that $\leftrightarrow \mathbf{Q}^H \cdot \leftrightarrow \mathbf{Q} = \leftrightarrow \mathbf{I}_p$. The second matrix is $p \times p$ upper triangular, so that all of the entries to the left of the main diagonal are zero. Combining the QR decomposition with the least squares problem yields

$$\mathbf{Q}^H \cdot \mathbf{y} = \leftrightarrow \mathbf{R} \cdot \hat{\boldsymbol{\beta}} \quad (20)$$

The left side of (19) is a p -vector. The p -th row of $\hat{\boldsymbol{\beta}}$ multiplied by the (p, p) element of $\leftrightarrow \mathbf{R}$ is equal to the p th element of the left side of (20). Stepping up one row, there is only one unknown because the p -th element of $\hat{\boldsymbol{\beta}}$ is already determined. This process of stepping back row by row is called back substitution, and the least squares problem has been solved without ever computing $\leftrightarrow \mathbf{X}^H \cdot \leftrightarrow \mathbf{X}$ or its inverse. QR decomposition is the best approach for numerically solving least squares problems.

Statistical inference for linear regression comprises a set of tools for assessing the fit of a given set of data to a linear model. Analysis of variance (ANOVA) provides a gross tool to assess the need for linear regression and, along with the widely used residual sum of squares, is not very informative. A more useful approach is a test of the null hypothesis $H_0 : \hat{\beta}_j = \beta_j^*$ against the alternate hypothesis $H_1 : \hat{\beta}_j \neq \beta_j^*$, where β_j^* is a particular value for the j th element in $\boldsymbol{\beta}$. The most widely used case sets $\beta_j^* = 0$ to test whether a given parameter is required. Let $\hat{\boldsymbol{\zeta}} = \leftrightarrow \mathbf{X}^H \cdot \leftrightarrow \mathbf{X}$. The test statistic is

$$\hat{t}_j = \frac{\sqrt{\hat{\zeta}_{jj}}(\hat{\beta}_j - \beta_j^*)}{\hat{\sigma}} \quad (21)$$

and is distributed as Student's t distribution with $N - p$ degrees of freedom if the null hypothesis is true. It is standard practice to produce a table that lists each element in $\boldsymbol{\beta}$, its standard error, the test statistic (21), and the corresponding p -value. If a given element of $\boldsymbol{\beta}$ has a p -value that lies below the type 1 error level α , then it should be removed from the linear model, although this must be done carefully and in conjunction with assessing the regression for outlying data.

The well-known equivalence between hypothesis testing and confidence interval estimation yields that the $1 - \alpha$ confidence interval on $\hat{\beta}_j$ is the set of all values for β_j^* for which the null hypothesis is acceptable with a type 1 error of α . Consequently, it follows that the $1 - \alpha$ confidence interval on β_j is

$$\begin{aligned} \hat{\beta}_j - t_{N-p} \left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{\hat{\mathbf{r}}^H \cdot \hat{\mathbf{r}}}{[(N-p)\hat{\zeta}_{jj}]} } &\leq \beta_j \\ &\leq \hat{\beta}_j + t_{N-p} \left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{\hat{\mathbf{r}}^H \cdot \hat{\mathbf{r}}}{[(N-p)\hat{\zeta}_{jj}]} } \end{aligned} \quad (22)$$

When there is more than one parameter in $\boldsymbol{\zeta}$, the Bonferroni method of replacing α with α/p should be utilized to avoid underestimating the confidence interval.

The bootstrap may also be used to estimate both the regression parameters and confidence intervals on them and is preferred because it does not depend as strongly on distributional assumptions compared to the parametric form (22). First, B bootstrap replicates are obtained by sampling with replacement from the response and predictor variables, and the linear regression estimator is applied to each to yield the $B \times p$ matrix $\{\hat{\boldsymbol{\beta}}_k^*\}$ whose column averages are the bootstrap estimate of the regression coefficients $\bar{\boldsymbol{\beta}}^*$. Bootstrap confidence intervals using the percentile method follow by sorting the columns of $\hat{\boldsymbol{\beta}}_k^*$ and taking the $[Ba/2p]$ and $[B(1 - \alpha)/2p]$ entries as the confidence interval for each. The BCa method may also be utilized. A double bootstrap estimator may be used to assess the significance of the regression coefficients; see Chave (2017), §9.4.3, and §9.5.2 for the details.

It is important to evaluate the regression residuals for lack of correlation, or randomness, to see that assumption (c) above is in effect. This is most easily achieved using the nonparametric Wald-Wolfowitz runs test that is described in Kvam and Vidakovic (2007) §6.5. The regression residuals should also be evaluated for the presence of autocorrelation using the Durbin-Watson test. Positive autocorrelation in the residuals will downward bias the parameter standard errors, hence increasing the false rejection rate on the parameters themselves. The details may be found in Chave (2017a) §9.4.4.

Influential data are those that are statistically unusual or that exert undue control on the outcome of a linear regression. They are classified as outliers or leverage points when they appear in the residuals and predictors, for which the hat matrix diagonal is a useful statistic. It is certainly possible to hand cull these from small data sets, but as data sets get larger this becomes impractical. What is needed is estimators that are robust to outlying data but that approach the efficiency of standard linear regression when outliers are absent. Such estimators have been devised over the past 50 years. The most widely used robust estimators are maximum-likelihood-like, or M-estimators. Chave et al. (1987) review the topic in a geophysical context.

M-estimators turn maximum likelihood estimation backward by a priori specifying a loss function that ensures robustness while maintaining efficiency rather than

specifying their distribution. The loss function $\rho(r)$ for a Gaussian distribution is $r^2 + c$, where c is a constant and r is a residual (e.g., difference between a datum x_i and the mean λ), and the estimation problem is

$$\min_{\lambda} \sum_{i=1}^N \rho(r_i) \quad (23)$$

A typical M-estimator will utilize a loss function that is Gaussian at the center but with longer tails at the extremes.

Taking the derivative in (23) with respect to λ yields the M-estimation analog to the normal equations:

$$\sum_{i=1}^N \psi(r_i) = 0 \quad (24)$$

where $\psi(r) = \partial_r \rho(r)$, and $\psi(r)$ is called the influence function.

Huber (1964) introduced a simple robust estimator based conceptually on a model that is Gaussian at the center and double exponential in the tails. The double exponential is the natural distribution for the L1 (least absolute values) norm, just as the Gaussian is the natural estimator for the L2 (least squares) norm. The loss and influence functions for the Huber estimator are

$$\begin{aligned} \rho(r) &= r^2 \quad |r| \leq \alpha \\ &= \alpha|r| - \frac{\alpha^2}{2} \quad |r| > \alpha. \\ \psi(r) &= r \quad |r| \leq \alpha \\ &= \alpha \operatorname{sgn}(r) \quad |r| > \alpha \end{aligned} \quad (25)$$

When $\alpha = 1.5$, the Huber estimator has 95% efficiency with purely Gaussian data, in contrast with an L1 estimator that has a ~60% variance penalty. From (25), the Huber estimator clearly bounds the influence of extreme data where the L2 estimator does not. More extreme loss and influence functions may be devised; for example, trimming the data at $|r| = \alpha$ means replacing the second terms in (25) with $\alpha^2/2$ and 0, respectively.

The numerical implementation of a robust estimator requires two additional steps because (24) is nonlinear, as the residuals are not known until the equation is solved. The argument in (23) or (24) must be independent of the scale of the data; in the Huber estimator (25), the L2 to L1 transition occurs at $\alpha = 1.5$ standard deviations, so the argument needs to be expressed in standard deviation rather than data units. This can be achieved by replacing r_i in (24) with r_i/d , where d is the sample scale estimate divided by the population scale estimate for the target distribution (which is usually Gaussian). The parameter d should not be sensitive to extreme data;

a suitable choice is s_{IQ}/σ_{IQ} , where IQ means the interquartile range.

The second step in implementing a robust estimator is linearization of the equations. This is accomplished by rewriting (24) as a weighted least squares problem:

$$\begin{aligned} \sum_{i=1}^N \psi(r_i) &\approx \sum_{i=1}^N w_i r_i = 0 \\ w_i &= \frac{\psi(r_i/d)}{r_i/d} \end{aligned} \quad (26)$$

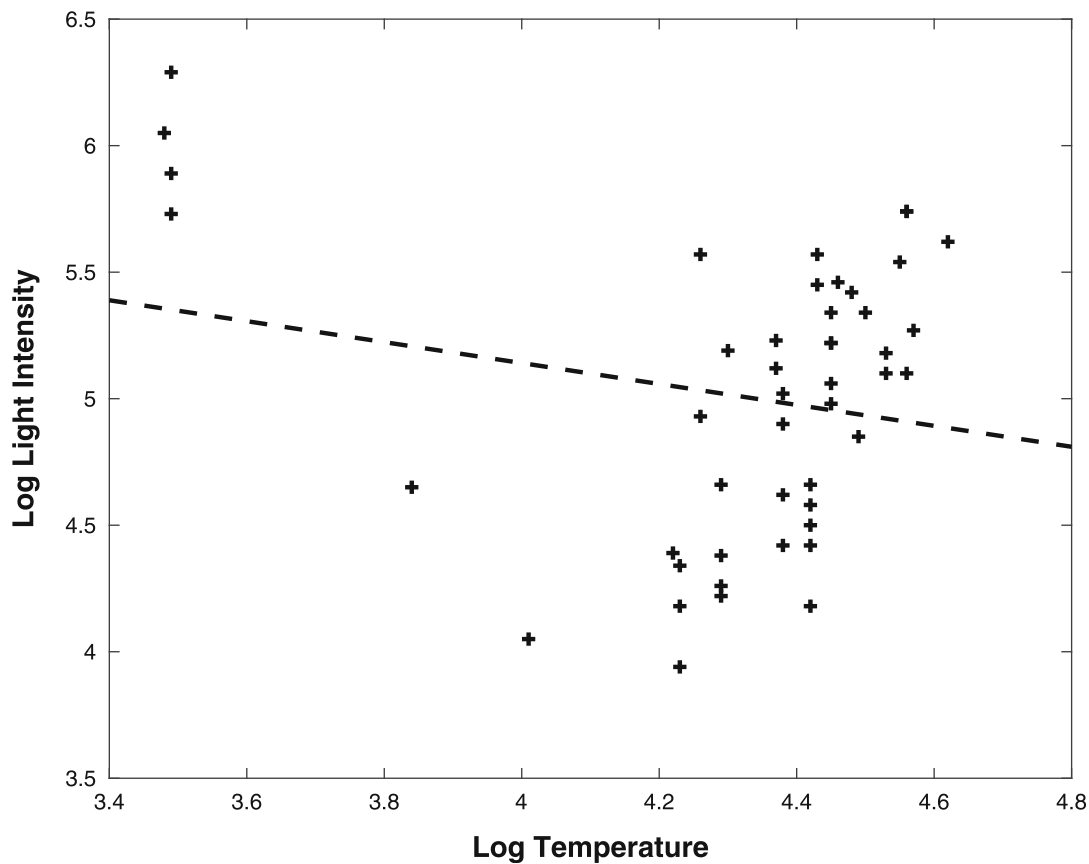
Linearization is achieved by iteratively solving (26) and computing the weights w_i and scale d using the residuals r_i from the prior iteration. Equation (26) is initialized using standard least squares (i.e., $w_i = 1$), and the iterative solution proceeds until the sum of squared residuals does not change above a threshold amount (such as 1%).

The robust estimator extends to linear regression if the residuals in (24) are identified as the regression residuals (15). Performing the minimization that yielded (24) gives

$$\sum_{i=1}^N \psi(r_i/d) x_{ij} = 0 \quad (27)$$

whose solution using weighted least squares is straightforward. The result should be assessed using the runs test, Durbin-Watson test, and t-statistic p-value, along with a q-q plot to check for outliers, just as for ordinary linear regression.

Rousseeuw and Leroy (1987, p. 27) presented what is now a classic data set comprising the logarithms of the effective surface temperature and light intensity of stars from the cluster Cygnus OB1. Chave (2017a) §9.5.2 presented their analysis including an intercept using tedious manual culling that showed that four outlying values due to red giant stars were “pulling” the fit and yielding a result that is orthogonal to the visual best fit. In §9.6.1, Chave (2017a) repeated the analysis using a Huber robust estimator. Only one iteration ensued, and the normalized sum of squared residuals was reduced from 0.3052 to 0.3049. The residuals are random and uncorrelated, but the regression is not fitting anything very well. Figure 4 shows the data and the fitted line. It is clear that the four red giant stars at the upper left are “pulling” the fit and forcing it to be rotated about 90° from a visual best fit through the main cluster of stars at the right side of the plot. In fact, the robust fit is not very different from what is achieved with ordinary least squares. The problem is that the red giant stars produce leverage manifest in the predictors rather than outliers manifest in the residuals, and robust estimators are insensitive to leverage points. In fact, M-estimators often make leverage worse or create leverage points that were not present at the start.



Statistical Computing, Fig. 4 The star data (crosses) along with a robust linear regression fit including an intercept (dashed)

This limitation has led to the introduction of bounded influence estimators that protect against both outliers and leverage points. Chave and Thomson (2003) replaced (26)–(27) with

$$\sum_{i=1}^N w_i^{[k]} v_i^{[k]} r_i^{[k-1]} x_{ij} = 0 \quad j = 1, \dots, p \quad (28)$$

where the superscript in brackets denotes the iteration number. The first set of weights w_i are standard M-estimator weights such as the Huber set. The second set of weights v_i utilize a hat matrix measure of leverage that is applied in an iteratively multiplicative fashion to prevent instability as leverage points appear and disappear during the solution of (28). The form used is

$$v_i^{[k]} = v_i^{[k-1]} \exp\left(-e^{\chi(x_i - \chi)}\right) \quad (29)$$

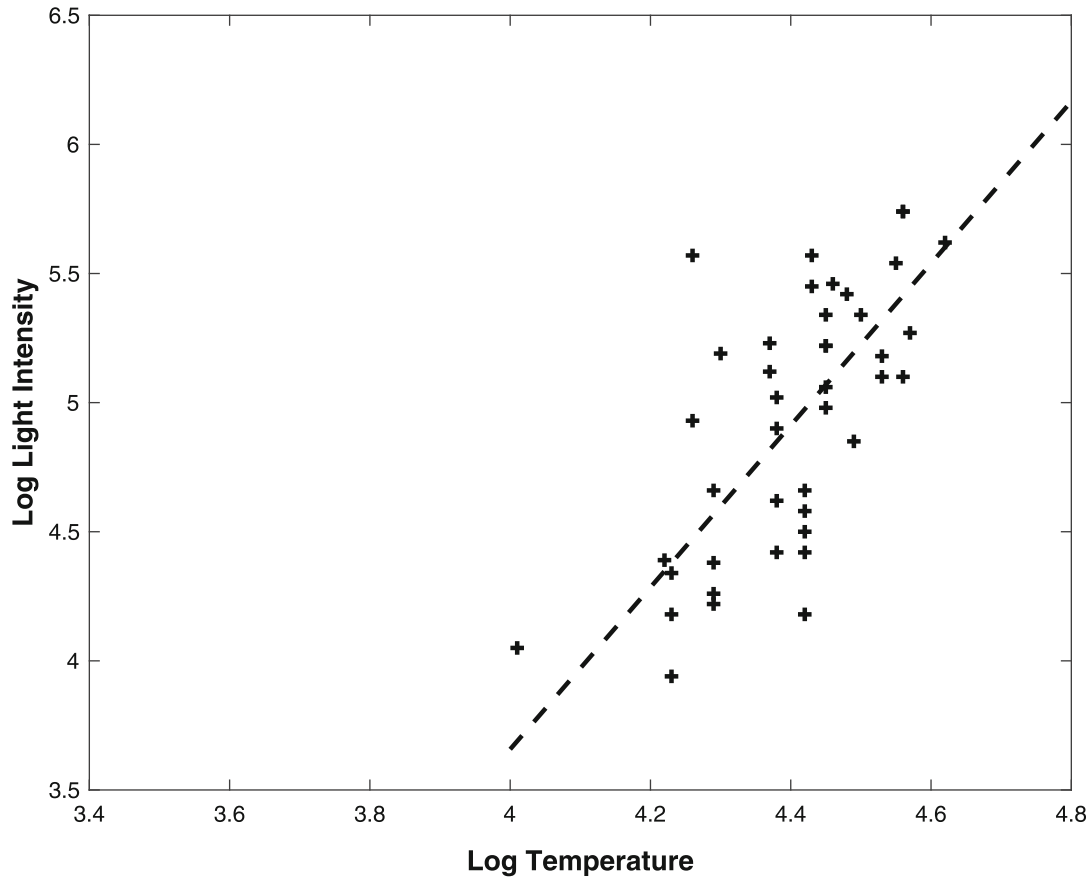
The statistic x_i is chosen to be $Nh_{ii}^{[k]}/p$, where $h_{ii}^{[k]}$ is the diagonal elements of the hat matrix that includes the product of the M-estimator and leverage weights. The free parameter χ sets the value where down-weighting begins, and a suitable

value is the Nth quantile of the beta distribution with parameters $(p, N - p)$ at some chosen probability level, such as 0.99.

Returning to the star data, the bounded influence estimator (28) was applied, resulting in four iterations during which the normalized residual sum of squares was reduced from 0.3052 to 0.1140. The residuals are random and uncorrelated. Weighted residual and hat matrix diagonal qq-plots reveal no outliers or leverage points remaining, and five data have been removed by the estimator. Figure 5 shows the result, which compares favorably with the hand-culled result from Chave (2017a) §9.5.2.

Nonlinear Multivariable Maximum Likelihood Estimation

As a final example of statistical computing, the implementation of a maximum likelihood estimator (MLE) for the analysis of magnetotelluric (MT) data will be described. MT utilizes measurements of the electric and magnetic fields on Earth's surface to infer the electrical structure beneath it. The fundamental datum in MT is a location-specific, frequency-



Statistical Computing, Fig. 5 The star data (crosses) and a bounded influence linear regression fit to them including an intercept (dashed).

The axis limits are the same as for Fig. 4, and data that have been eliminated by bounded influence weighting have been omitted

dependent tensor $\leftrightarrow \mathbf{Z}$ linearly connecting the horizontal electric and magnetic fields $\mathbf{E} = \leftrightarrow \mathbf{Z} \cdot \mathbf{B}$. This relation does not hold exactly when $\mathbf{E}$ and $\mathbf{B}$ are measurements, and instead the row-by-row linear regression is employed:

$$\mathbf{e} = \leftrightarrow \mathbf{b} \cdot \mathbf{z} + \boldsymbol{\epsilon} \quad (30)$$

where $\mathbf{e}$ is the electric field response N -vector, $\leftrightarrow \mathbf{b}$ is the $N \times 2$ magnetic field predictor matrix, $\mathbf{z}$ is the MT response 2-vector, and $\boldsymbol{\epsilon}$ is an N -vector of random errors, with all of these entities being complex. The time series measurements are Fourier transformed into the frequency domain over N segments that are of order one over the frequency of interest in length; see Chave (2017b) for details.

MT was plagued by problems with bias and erratic responses until the late 1980s, when robust and then bounded influence estimators were devised that dramatically improved results. All of these approaches are based on the robust model comprising a Gaussian core of data plus outlying non-Gaussian data that cause analysis problems. However, Chave (2014, 2017b) showed that the residuals from a robust estimator were systematically long tailed, so that the Gaussian core does not exist, and then showed that the residuals were

pervasively distributed according to a stable model. This immediately suggested that an MLE based on the stable model could be implemented.

For stable MT data, the pdf of a single residual is $S(\hat{r}_i|\boldsymbol{s})$, where $\hat{r}_i = e_i - \mathbf{b}_i \cdot \hat{\mathbf{z}}$ is the i th estimated residual, $\mathbf{b}_i$ is the i th row of $\leftrightarrow \mathbf{b}$ in (30), and $\boldsymbol{s} = (\alpha, \beta, \gamma, \delta)$ are the tail thickness, skewness, scale, and location parameters for a stable distribution. For independent data, the sampling distribution is

$$S_N(\hat{\mathbf{r}}|\boldsymbol{s}) = \prod_{i=1}^N S(\hat{r}_i|\boldsymbol{s}) \quad (31)$$

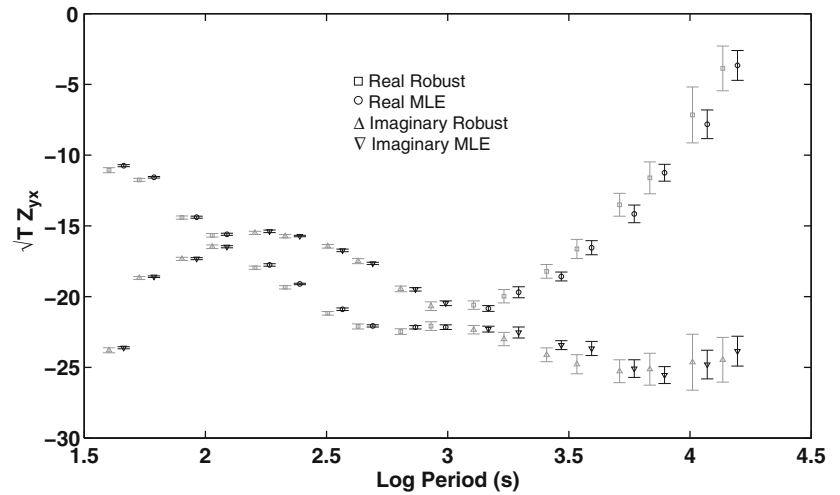
The likelihood function is the sampling distribution regarded as a function of the parameters for a given set of residuals. The MLE is obtained by maximizing the likelihood function, or equivalently its logarithm:

$$L(\boldsymbol{s}, \mathbf{z}|\hat{\mathbf{r}}) = \sum_{i=1}^N \log S(\hat{r}_i|\boldsymbol{s}) \quad (32)$$

The first-order conditions for the MLE solution follow by setting the derivatives of (32) with respect to the parameters to zero:

Statistical Computing,

Fig. 6 The real and imaginary parts of the Z_{xy} component of the MT response scaled by the square root of period for a site located in South Africa. The confidence limits for the robust (gray) and stable MLE (black) are obtained using the jackknife and diagonal elements of the improper Fisher information matrix, respectively, assuming that the tail probability is apportioned equally among the eight real and imaginary elements of the response tensor. The stable MLE estimates have been offset slightly for clarity. (Taken from Chave (2014))



$$\partial_{\hat{r}_j} L(\mathbf{s}, \mathbf{z} | \hat{\mathbf{r}}) = \sum_{i=1}^N \frac{\partial_{\hat{r}_j} S(\hat{\mathbf{r}}_i | \mathbf{s})}{S(\hat{\mathbf{r}}_i | \mathbf{s})} = 0 \quad j = 1, \dots, 4$$

$$\partial_{z_k} L(\mathbf{s}, \mathbf{z} | \hat{\mathbf{r}}) = \sum_{i=1}^N \frac{\partial_{z_k} S(\hat{\mathbf{r}}_i | \mathbf{s})}{S(\hat{\mathbf{r}}_i | \mathbf{s})} b_{ik} = 0 \quad k = 1, \dots, 4 \quad (33)$$

The sufficient condition for the solution to (33) to be a maximum is that the Hessian matrix of the log likelihood function (32) be negative definite.

Chave (2014) solved (33) using a two-stage approach that decouples the stable distribution parameter vector $\mathbf{s}$ and the MT response function $\mathbf{z}$. In the first stage, the stable distribution was estimated using the stable MLE algorithm of Nolan (2001). In the second stage, the MT response function was obtained using the second part of (33) while fixing the stable distribution parameters. This proceeds iteratively until convergence is obtained. The starting solution is the ordinary least squares result with 5% trimming.

Equations (31) can be solved using a weighted least squares approach, but convergence slows as the tail thickness parameter α decreases toward 1 and fails for smaller values. Experience showed that MT data exhibit tail thicknesses as low as 0.8, so this limitation was a real problem. An alternative approach utilized nonlinear multivariable minimization of an objective function based on a trust region algorithm for the second stage, including explicit provision of the gradient and Hessian matrix to speed convergence. The objective function that was minimized is the negative log likelihood given by (32) preceded by a minus sign and with the stable parameter vector fixed. This method worked well for all values of the tail thickness parameter at the cost of one to two orders of magnitude more computer time than weighted least squares.

Figure 6 shows a typical result from the stable MLE compared to a standard robust estimator. The stable MLE produces a smoother result as a function of frequency, as

expected from diffusion physics, and smaller error estimates as compared to a bounded influence estimator. The latter is because the MLE is asymptotically efficient, reaching the lower bound of the Cramér-Rao inequality which a bounded influence estimator cannot achieve. In fact, there are substantial differences (up to 10 standard deviations) between the stable MLE and bounded influence results, although these are subtle in Fig. 6. The tail thickness parameter hovers around 1 at periods under 30 s. The p-values for a test of the null hypothesis that the data are stably distributed based on the Kolmogorov-Smirnov statistic does not reject at any frequency once the test is corrected for bias due to estimation of the distribution parameters from the data using the resampling approach described above.

Summary

This chapter has covered four topics in statistical computing: exploratory data analysis, resampling methods, linear regression, and nonlinear optimization. The first topic includes tools to statistically characterize a data set using a kernel density estimator, quantile-quantile plot, percentile-percentile plot, and simulation. Resampling methods include the use of the bootstrap to compute estimators and confidence limits and permutation methods for hypothesis testing that are nearly exact. Linear regression was described both theoretically and numerically and then extended to robust and bounded influence estimators that automatically handle unusual data (e.g., outliers). Nonlinear optimization was described based on the solution for the magnetotelluric response function after recognizing that such data in the frequency domain pervasively follow a stable distribution. This set of topics is not comprehensive, and many additional ones exist that comprise statistical computing.

Cross-References

- [Compositional Data](#)
- [Computational Geoscience](#)
- [Constrained Optimization](#)
- [Exploratory Data Analysis](#)
- [Frequency Distribution](#)
- [Geostatistics](#)
- [Hypothesis Testing](#)
- [Iterative Weighted Least Squares](#)
- [Multivariate Analysis](#)
- [Normal Distribution](#)
- [Optimization in Geosciences](#)
- [Ordinary Least Squares](#)
- [Regression](#)

Bibliography

- Chave AD (2014) Magnetotelluric data, stable distributions and impropriety: an existential combination. *Geophys J Int* 198: 622–636
- Chave AD (2017a) Computational statistics in the earth sciences. Cambridge University Press, Cambridge
- Chave AD (2017b) Estimation of the magnetotelluric response function: the path from robust estimation to a stable mle. *Surv Geophys* 38: 837–867
- Chave AD, Thomson DJ (2003) A bounded influence regression estimator based on the statistics of the hat matrix. *J R Stat Soc Ser C* 52: 307–322
- Chave AD, Thomson DJ, Ander ME (1987) On the robust estimation of power spectra, coherences, and transfer functions. *J Geophys Res* 92: 633–648
- De Groot MH, Schervish MJ (2011) Probability and statistics, 4th edn. Pearson, London
- Efron B (1979) Bootstrap methods: another look at the jackknife. *Ann Stat* 7:1–26
- Efron B, Hastie T (2016) Computer age statistical inference. Cambridge University Press, Cambridge
- Huber P (1964) Robust estimation of a scale parameter. *Ann Math Stat* 35:73–101
- Kvam PH, Vidakovic B (2007) Nonparametric statistics with applications to science and engineering. Wiley, Hoboken
- Michael JR (1983) The stabilized probability plot. *Biometrika* 70: 11–17
- Nolan JP (2001) Maximum likelihood estimation and diagnostics for stable distributions. In: Lévy processes: theory and applications. Birkhäuser, Cham
- Pitman EJG (1937) Significance tests which may be applied to samples from any distribution. *J R Stat Soc Suppl* 4:119–130
- Rousseeuw PJW, Leroy AM (1987) Robust regression and outlier detection. Wiley, New York
- Schreier PJ, Scharf LL (2010) Statistical signal processing of complex-valued data. Cambridge University Press, Cambridge
- Simpson J, Olsen A, Eden JC (1975) A Bayesian analysis of a multiplicative treatment effect in weather modifications. *Technometrics* 17: 161–166
- Wasserstein RL, Lazar NA (2016) The ASA's statement on p-values: context, process and purpose. *Am Stat* 70:129–133

Statistical Inferential Testing

R. Webster

Rothamsted Research, Harpenden, UK

Definition

Information on the constituents of rock, sediment, and soil necessarily derive from samples; it is never complete. Inferences drawn from data are uncertain, and the main aim of statistical testing is to attach probabilities to those inferences. The tests are prefaced by hypotheses or models, such as no differences among population means or relations between variables, the so-called null hypotheses. They are designed to refute those hypotheses by determining the probabilities that the hypotheses or models are true, given the data. A result is declared statistically significant when the probability, P , is less than some small value, conventionally $P < 0.05$. Statistical significance is not the same as scientific significance, and calculated values of P should always be presented along with standard errors in reports and papers so that clients and readers can draw their own inferences from results.

Introduction

Statistical inference is a broad subject. In the geosciences it can be narrowed to inferences about *populations* that may be drawn from *sample* data. Rock formations and sediments are continuous bodies of variable material. Soil mantles the land surface except where it is broken by rock, and it too varies both from place to place and from time to time. Knowledge about many of the properties of these materials is incomplete because it derives from observations on samples; one cannot measure them everywhere. Further, measurements themselves are more or less in error. Knowledge of the true states of the materials is therefore uncertain, and one of the main aims of statistics in this context is to put that uncertainty on a quantitative footing; it is to attach probabilities to the inferences drawn from data. Following from it is the notion of statistical significance, of separating meaningful information, i.e., signal, from “noise,” i.e. from unaccountable variation in data or from sources of no interest. The information might be change of some property of the soil arising from change of use, the difference between analyses of the metal content of an ore from two laboratories, and the response of some treatment to combat pollution in a sediment. Given the uncertainties in data, the question then becomes: are observed differences or response so large in relation to the unaccountable variation

that their occurrence by chance is highly improbable? If the answer is “yes,” then the result is said to be “significant”; it will be accompanied by a value P , a measure of this improbability – see below.

A significance test is prefaced by a hypothesis. In the above examples this would be that there has been no effect on the soil caused by change of land use, that results from different laboratories do not differ, that the pollution in the sediment has not abated in response to the treatment. That is the “null hypothesis,” often designated H_0 , and the test is designed to reject it, *not to confirm it*. The alternative that there is a difference is denoted either H_1 or H_A .

A valid test requires (a) a plausible underlying model of the phenomenon being investigated and (b) that the sampling has been suitably randomized to provide the data. Embodied in the model is a statistical distribution within which to judge the data.

The Normal Distribution

Statistical tests of significance are set against a background of variation with assumed probability distributions, by far the commonest of which is the normal, or Gaussian, distribution. Its probability density is given by

$$y = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \quad (1)$$

where μ is the mean and σ^2 is the variance of the distribution. Its graph of y against x is a familiar bell-shaped curve. A variable with mean $\mu = 0$, variance $\sigma^2 = 1$, and this distribution is a standard normal deviate, often denoted z . Many variables describing the properties of geological bodies and materials are distributed approximately in this way or can be made so by transformation of their scales of measurement. Further, the distributions of sample means increasingly approach normal as the sample sizes increase whatever the distributions of individual measurements, a result known as the central limit theorem.

Tests With z

One use of z is to place confidence limits about a population mean, μ , estimated by the mean, $\bar{x}$, of a sample. For a sample of size n with variance s^2 , the variance of its mean is s^2/n and its square root is the standard error. In these circumstances the lower and upper limits on μ for some given confidence are

$$\bar{x} - z\sqrt{s^2/n} \text{ and } \bar{x} + z\sqrt{s^2/n}.$$

Frequently used values are

Confidence(%)	80	90	95	99
z	1.28	1.64	1.96	2.58

The z test is often used to decide whether a sample belongs to some population or that it deviates from some norm with mean μ and variance σ^2 . The null hypothesis is no deviation: $\bar{x} = \mu$. The test is given by

$$z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}}, \quad (2)$$

for which the corresponding probability, P , can be calculated or found in tables of the standard normal deviate for the number of degrees of freedom.

In practice the population variance is rarely known and is estimated by the sample variance, s^2 . In these circumstances, unless the sample is large ($n > \approx 60$), z must be replaced by student's t :

$$t = \frac{\bar{x} - \mu}{\sqrt{s^2/n}}. \quad (3)$$

Again, the values of P corresponding to those of t for the number of degrees are tabulated in published tables and nowadays can be calculated in most statistical packages. For sample sizes exceeding, the distributions of z and t are so similar that the tests give almost identical results.

Equations (2) and (3) lead to probabilities that the differences between $\bar{x}$ and μ lie in either tail of the distributions; the differences may be either negative or positive. The tests are two-tailed or two-sided in the jargon. In many instances investigators are interested in only one tail. For example, a public health authority may want to know only whether the lead content of the soil exceeds some threshold; it would calculate

$$c = \bar{x} - z\sqrt{\sigma^2/n} \text{ or } c = \bar{x} + t\sqrt{\sigma^2/n}, \quad (4)$$

and obtain the corresponding probability, P , which would be half that for the two-sided test.

The F Test

The quantity F , named in honor of R.A. Fisher, is the quotient resulting from the division of one variance, say s_1^2 , by a smaller one s_2^2 : s_1^2/s_2^2 . Fisher worked out its distribution if s_1^2 and s_2^2 are variances of two independent random samples from two normally distributed populations with the same variance σ^2 . The calculated value of F is compared with that distribution to obtain the probability, P , given the evidence, that the null hypothesis, i.e., no difference between the

variances of the populations from which the samples were drawn, is true. The larger is the calculated F , the smaller is P , other things being equal.

More often F is invoked to test for differences among means of classes in experiments and surveys after analyses of variance (ANOVAs). Table 1 is an example; it summarizes the ANOVA of lead (Pb) in the soil in the Swiss Jura classified by geology (from Atteia et al. 1994) with measurements in mg kg⁻¹ transformed to their common logarithms. The probability of $F = 5.95$ under the null hypothesis of no differences among the means is $P < 0.001$; the null hypothesis is clearly untenable. Table 2 reveals why; the soil on the Argovian rocks is poorer in Pb than on the others.

Table 2 lists the means of both the measurements in mg kg⁻¹ and of their transforms to common logarithms to bring the distributions close to normal. The last column lists the standard errors derived from the residual mean square from the ANOVA set out in Table 1. The differences are not huge but with so many data the test is very sensitive.

The natural scientist might seek reasons for the soil on the Argovian's being poorer in Pb. The environmental protection agency might be more concerned because the concentrations of Pb on four of the five formations exceed the statutory limit of 50 mg kg⁻¹.

Individual means may be compared. The standard error of a difference is

$$\text{SED} = \sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}, \quad (5)$$

where s^2 is the residual mean square from the ANOVA and n_1 and n_2 are the sizes of the two samples. Student's t for the comparison of the means, say $\bar{x}_1$ and $\bar{x}_2$, is then

$$t = (\bar{x}_1 - \bar{x}_2) / \text{SED} \quad (6)$$

In this example, the comparison of the Argovian with the four others would be the main interest. It would require a vector of weighting coefficients $\mathbf{w} = \{1, -\frac{1}{4}, -\frac{1}{4}, -\frac{1}{4}, -\frac{1}{4}\}$ by which to multiply both the means and standard errors to obtain the appropriate value of t . Details are set out in Sokal and Rohlf (2012), which is among the most up-to-date and nearly comprehensive didactic texts.

The practice of comparing every pair or sets of pairs in this way is to be deprecated. In particular, one cannot assign a probability to a difference only when it becomes apparent from an ANOVA. Hypotheses should be set out first, and experiments and surveys designed to test them. Ideally, they should comprise orthogonal contrasts, the number of which is equal to the degrees of freedom between classes; Sokal and Rohlf (2012) and Webster and Lark (2018) explain.

If there are only two classes, then the F test is equivalent to the t test: $F = t^2$.

Tests for Normality

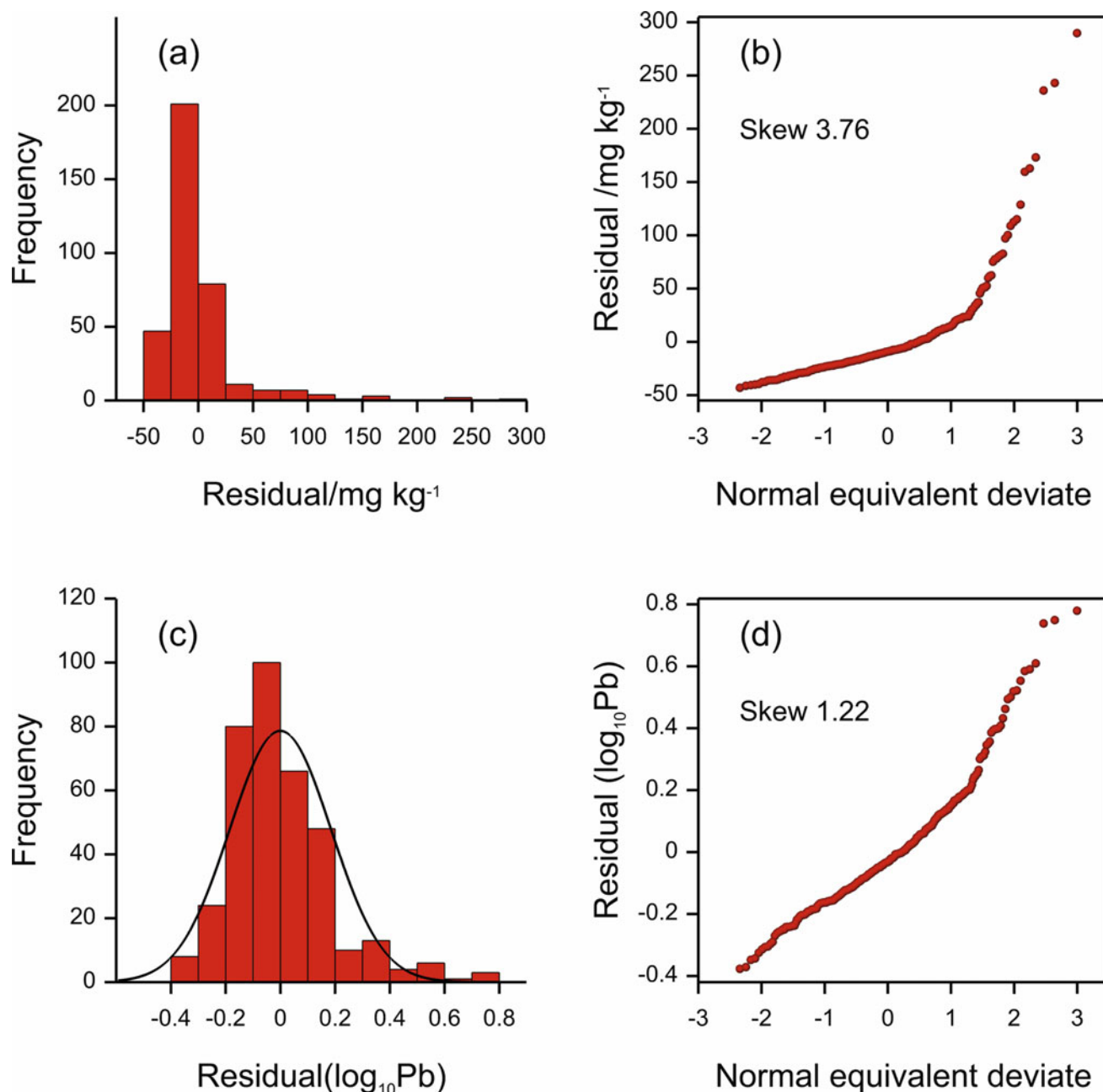
Geoscientists often worry that the validity of ANOVAs and regression analyses are compromised if data are not normally distributed. The tests such as the Kolmogorov–Smirnov test, the Shapiro–Wilk test, or a χ^2 test on frequencies of a histogram are of little help. With many data almost any departure from normal will be judged significant, whereas with few data the tests are inconclusive. The tests do not reveal in what way the distributions deviate from normal. Further, what matter for ANOVA and regression are the distributions of the residuals, not those of the data. Far better guides are histograms and Q–Q graphs of the residuals (see Welham et al. 2015). Figure 1 shows the distributions of the residuals from the ANOVAS of the lead data both on the original measurements, Fig. 1a and b, and on their logarithmic transforms, Figure 1c and d, summarized in Table 2. The histogram of the residuals from the

Statistical Inferential Testing, Table 1 Analysis of variance for log₁₀ Pb in the soil of the Swiss Jura on five geological formations

Source	Degrees of freedom	Sum of squares	Mean square	F ratio	P
Between classes	4	0.83017	0.20754	5.95	<0.001
Within classes (residual)	358	12.08266	0.03847		
Total	362	13.31261			

Statistical Inferential Testing, Table 2 Mean concentrations of lead (Pb) in the soil on five geological formations in the Swiss Jura and the means of their common logarithms and their standard errors (from the ANOVA of Table 2)

Formation	Number of observations	Mean/mg kg ⁻¹	Mean of logarithms	Standard error
Argovian	76	43.9	1.641	0.02142
Kimmeridgian	126	57.0	1.744	0.01664
Sequanian	90	66.7	1.777	0.01968
Portland	7	56.2	1.740	0.07058
Quaternary	67	56.6	1.722	0.02281



Statistical Inferential Testing, Fig. 1 Histograms and Q-Q graphs of residuals from the analysis of variance of the lead in the soil of the Swiss Jura; (a) and (b) of the original measurements in mg kg^{-1} and (c) and (d)

on their logarithmic transforms. The curve in (c) is that of a normal distribution fitted to the data

ANOVA of the original measurements shows that they are far from normal, and the points on the Q-Q graph form a strong curve. Those of the transformed data are closer to normal, though still somewhat skewed (skewness coefficient 1.22 and points on the Q-Q graph not quite lying on straight line) (Fig. 1c and d). The ANOVA for comparing means is robust against small departures from normality, and it can be used with confidence in such circumstances.

Significance and *P* Value

A *P* value from a statistical test is the probability that the specified statistical model or hypothesis can be held as true in the light of the data. The smaller is this value of *P*, the stronger is the evidence against the model or hypothesis. It is this evidence that constitutes statistical significance, and *P* is its measure. R.A. Fisher originally suggested that a $P \leq 0.05$ was reasonable grounds for rejecting a null hypothesis, for

example, no differences among means, and declaring significance. It has subsequently become a convention. Such has been the misunderstanding and abuse of this threshold that the American Statistical Association felt bound to issue a statement on the matter (Wasserstein and Lazar 2016). The statement comprised six principles summarized below:

1. The first principle reads “*P*-values can indicate how incompatible are the data with a specified statistical model.” Usually the model is a null hypothesis, for example, no differences among means or responses to treatments, and *P* is a measure of the probability that the hypothesis is true.
2. A *P* value *does not* measure the probability that a proposed null hypothesis is true or that the data arose purely at random.
3. Scientific conclusions, business decisions, or policy for action should not be based on *P* values alone.
4. Proper inference requires full reporting and transparency. The concern here is the selection of results with small *P* values that will lead to publications in the scientific literature, a practice known as “*P* hacking.”
5. A *P* value or its equivalent statistical significance does not measure the size of an effect or the importance of a result. Statistical significance is not the same as scientific or practical significance.
6. By itself, a *P* value is not a good measure of evidence regarding a model or hypothesis. A result, small value of *P*, might be statistically significant because with many data the test is very sensitive but of no scientific significance. A nonsignificant result, large *P*, might mean that the null hypothesis is correct, but it does not measure the probability that it is so; a wise scientist takes the view that there is too little evidence to conclude that observed differences apply to the populations from which the samples are drawn.

One can still draw mistaken conclusions as the result of significance test.

Mistakes can be of two kinds, denoted Type I and Type II. The first occurs when one rejects the null hypothesis on the basis of sample evidence, i.e., declares differences significant, when the populations do not differ. The second is when one accepts the null hypothesis, i.e., states that there is insufficient evidence for differences, when the populations do differ. One must realize that $P = 0.05$ is not a dividing line between falsehood and truth. Rather a *P* value lies on a continuous scale from more to less probable. Probabilities can now be readily computed in modern statistical packages; they should be included in all reports of statistical analyses as recommended by Wasserstein and Lazar (2016). If they are published along with the means and their standard errors, then readers can reach their own inferences.

Summary

Geoscientists who want to draw conclusions about populations of rock, sediments, or soil from sample data must contend with uncertainty. Statistical inference requires plausible models of the distributions of the variables of interest and suitably randomized designs by which the sample data are obtained. Inferential test are prefaced by hypotheses or models, and the tests compute the probabilities, *P*, that the hypotheses are true or that the data do comply with the proposed models. Two common tests, the *z* test for comparing individual values with population statistics and the *F* test in the analysis of variance for comparing mean values, are based on underlying normal distributions. They are robust against small departures from that distribution, and tests for normality have little merit; histograms and Q–Q graphs of residuals are better guides. The American Statistical Association recently set out six principles for reporting probabilities from inferential tests. It recommended complete transparency, with *P* values placed on a continuous scale instead of more or less than the conventional significance at $P \leq 0.05$. Computed values of *P* should be reported along with estimated means and their standard errors. Readers and clients can then judge for themselves whether statistically significant results are scientifically significant or practically important.

Bibliography

- Atteia O, Dubois J-P, Webster R (1994) Geostatistical analysis of soil contamination in the Swiss Jura. *Environ Pollut* 86:315–327
- Sokal RR, Rohlf FJ (2012) *Biometry: the principles and practice of statistical research*, 4th edn. W.H. Freeman, San Francisco
- Wasserstein RL, Lazar NA (2016) The ASA’s statement on *p*-values: context, process and purpose. *Am Stat* 70:129–133
- Webster R, Lark RM (2018) Analysis of variance in soil research: let the analysis fit the design. *Eur J Soil Sci* 69:126–139
- Welham SJ, Gezan SA, Clark SJ, Mead A (2015) *Statistical methods in biology: design and analysis of experiments and regression*. CRC Press, Taylor and Francis Group, Boca Raton

Statistical Outliers

Mehala Balamurali and Raymond Leung
Rio Tinto Centre for Mine Automation, Australian Centre for Field Robotics, Faculty of Engineering, The University of Sydney, Sydney, NSW, Australia

Definitions

Outliers refer to data points that differ significantly from other observations (Grubbs 1969). Although extreme values can

occur by chance from any distribution, the outliers encountered during mining are mainly the result of collecting assay samples from adjacent domains with contrasting composition to the targeted domain; these outliers are included unintentionally due to inaccurate geological domain boundary definitions. Henceforth, the term *sample* (or assay sample) is used synonymously with *data point*. For clarity, each sample is represented by a p -dimensional vector which measures the concentration of p geochemical components at one location, as opposed to repeated measurements over a set of random variables in the context of a dynamical system. Statistically, outliers may be considered as values that derive from another process or source (Hampel et al. 2011; Barnett and Lewis 1994). In this exposition, outliers are assumed to belong to a separate population and their attributes are expected to be vastly different to the data points sampled from the main population.

Impact of Outliers in Geological Domaining

In open pit mining, the accuracy of grade estimation is impacted by the existence of outliers within a geological domain. The presence of outliers contributes directly to interpolation errors of the grade in the form of overestimation or underestimation. This bias not only undermines confidence and grade control efforts, it also has a flow-on effect on the execution of subsequent mining processes and ability to adhere to production targets. Significantly, mine planning and excavation activities are informed by the grade prediction model. Hence, artifacts and inaccuracies in the model caused by outliers can ultimately affect the efficiency and profitability of the overall mining operation.

An integral part of standard mine practice is interpreting the deposits into geological domains. Geological domains are used to group correlated geochemical data together for sectional interpretation and 3D modeling purposes, particularly where there is lack of continuity between different parts of the region. These domains are further used to ensure that all the grades spatially located in a modeling unit are grouped together for grade estimation of that unit. Geological domain definitions are derived from exploration and blast-hole data. These observations often involve a dozen or so variables. They are also subject to estimation errors and measurement noise. The true extent depends on the quality of the mining program and various controls which affect how ore samples are collected and assayed. Principally, outliers occur because samples are mistakenly included from outside the target domain. Techniques for rectifying such errors have been proposed in (Leung et al. 2022), but these lie outside the scope of the current discussion. Acknowledging that incorrect grouping of samples inevitably occurs, attention is turned to identifying the outliers.

Outliers may be mingled with the target domain data for a variety of reasons. Some outliers are just bad data that arise due to logging, sampling, or calibration errors. Irrespective of their origin, outliers can be handled in a passive or proactive way. Robust estimation techniques offer resilience and constitute the former, they implicitly ignore the influence of outliers and minimize their impact on statistical estimates and domain interpretations. Proactive approaches may additionally reclassify the removed outliers and consider whether the outliers collectively form meaningful subpopulations. This is useful when latent domains are present within the dominant (target) domain, it allows existing knowledge about an orebody to be updated and offers new insight. It presents an opportunity for new localized geological features to be discovered. Regardless of the motivations, identifying the outliers is an important first step.

In this summary of outlier detection work, we will discuss (1) multivariate outlier detection methods, (2) t-SNE-based subdomain detection, and (3) performance comparison of t-SNE and MCD with different sample truncation strategies. This study draws upon material from previously published works that focus on iron ore mining in the Pilbara region in Western Australia.

Techniques: Multivariate Outlier Detection in Geochemical Data

Traditional methods for identifying univariate and bivariate outliers are well known to geologists (Rousseeuw and Leroy 1987; Barnett and Lewis 1994). However, detection of multivariate outliers is still an active research area. Although outlier detection can be applied to individual variables in multivariate geochemical data (Reimann et al. 2005; Filzmoser et al. 2005), this is not generally advisable as it ignores the correlations (covariances) that exist between different components.

For univariate outlier detection, a simple approach that can be used is the standard deviation method. In geochemical analysis, values that fall outside mean $\pm$ stdev are considered as extreme data points in (Hawkes and Webb 1962). Reimann et al. (2005) explained that geochemistry outliers are not often extreme values and instead originate from different, often superimposed, distributions associated with processes that are rare in the environment. Moments describing the data [mean $\pm$ stdev] work well when the data is really an extreme value, but the rule fails when the presence of outlier populations are relatively large in comparison to the target population.

Tukey (1977) proposed a test for screening outliers. A graphical tool is used to display the distribution of the univariate data by displaying its median, 25th percentile (lower quantile Q_1), 75th percentile (upper quantile Q_3). The

interquartile range is defined as $Q_3 - Q_1$ as in a boxplot. Inner fences and outer fences are defined by the interval $[Q_1 - k \times (Q_3 - Q_1), Q_3 + k \times (Q_3 - Q_1)]$ with $k = 1.5$ and $k = 3$, respectively. Any values outside the outer fences are shown as outliers in the graph.

Tukey's method is considerably effective compared with the standard deviation method. They both work well when the data is normally distributed. However, geochemical data often do not follow normality and instead are skewed in one direction. Thus, their distribution is nearly a lognormal distribution (Iglewicz and Hoaglin 1993). The outliers can be identified from transformed distribution. Tukey's method identifies outliers when the distribution is skewed because it makes no assumption about the distribution. However, in Tukey's approach, more points are detected as outliers when the data is highly skewed; the method may fail when the sample size is small.

Many others have studied the problem of univariate outlier detection. A recurring theme is that univariate extreme points are identified based on a significantly increased distance from the center of the data compared to the other points (inliers). Unlike univariate outliers, multivariate outliers are identified as distant points from a robust ellipse or multidimensional cloud that retain samples – the majority – which are in agreement (Gnanadesikan and Kettenring 1972; Rousseeuw and Leroy 1987; Barnett and Lewis 1994). In the literature, there is a general recognition that clusters of similar compositions can be identified more effectively in subspaces rather than in a full dimensional space as outliers (Strehl and Ghosh 2002).

Univariate outlier detection is generally inadequate for identifying multivariate outliers in geochemical data. Concretely, a data point may look extreme and deviate enough with respect to one or more variables for it to satisfy the independent requirements of univariate outlier, yet the criteria of a multivariate outlier may not be automatically satisfied when all variables (geochemical components) are jointly considered. In univariate outlier detection, distance to the center

is used; but in multivariate outlier detection, both distance and structure of the data are needed (Rousseeuw 1985). Several methods for measuring dispersion in multivariate data are presented next.

Classical Mahalanobis Distance

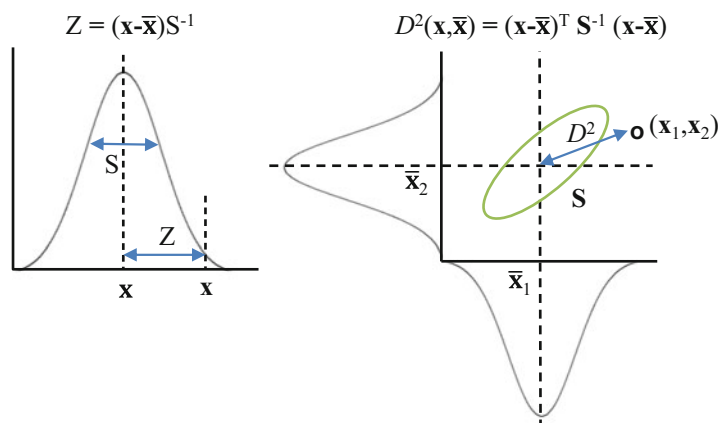
The Mahalanobis distance measures how far a point $\mathbf{x}$ is from the mean of a dataset (Mahalanobis 1936; De Maesschalck et al. 2000). For multivariate outlier detection, the squared Mahalanobis distance is given by

$$D^2(\mathbf{x}, \bar{\mathbf{x}}) = (\mathbf{x} - \bar{\mathbf{x}})^T S^{-1} (\mathbf{x} - \bar{\mathbf{x}}) \quad (1)$$

In this equation, $\mathbf{x}$ denotes a single observation and $\bar{\mathbf{x}}$ denotes the centroid; both are p -dimensional vectors. A dataset comprises n observations (assay samples) and p features (variables). Each element in $\bar{\mathbf{x}}$ represents the mean concentration of a chemical component. If matrix X holds the original data with columns centered by their means, then the $p \times p$ covariance matrix may be computed as $S = X^T X / (n - 1)$.

An observation with a Mahalanobis distance $D^2(\mathbf{x}, \bar{\mathbf{x}}) > k$ is regarded as an outlier. Figure 1 provides a visualization of the univariate and bivariate distance measures. In the univariate case, distance is measured by $z = (x - \bar{x}) s^{-1}$, where s is the sample standard deviation. Points with high z -scores are treated as outliers. In the multivariate case, outlier detection considers patterns between all variables. Since the Mahalanobis distance is sensitive to outliers, Rousseeuw (1985) described a method for removing outliers (i.e., retaining representative samples) in a given domain to minimize the impact of outliers on the S estimate. This estimator is robust as it uses only a subset of observations (the h points) with the smallest covariance determinant. The z -score-bivariate analogy generalizes to higher dimensions via the Mahalanobis distance. Assuming the data distribution is multivariate Gaussian [for compositional data such as chemical assays, this is only approximately true after log-ratio

Statistical Outliers, Fig. 1 Left: univariate (standard normal deviate). Right: bivariate (generalized distance) distance measures



transformation is applied], it has been shown that the distribution of the Mahalanobis distance behaves as a chi-square distribution given a large number of samples (Garrett 1989). Therefore, a popular choice for the proposed cutoff point is $k = \chi^2_{(p, 1-\alpha)}$, where χ^2 stands for the chi-square distribution and α is the significance level – usually taken as 0.025 (Rousseeuw 1985).

Minimum Volume Ellipsoid

The minimum volume ellipsoid (MVE) estimator finds the mean ($\bar{\mathbf{x}}_J^*$) and covariance (S_J^*) from $h(\leq n)$ points that minimize the volume of the enclosing ellipsoid. This ellipsoid is computed from the covariance matrix associated with the subsample. Formally, $MVE(X) \rightarrow (\bar{\mathbf{x}}_J^*, S_J^*)$ where the optimal set $\{X_i | X_i \in J\}$ satisfies

$$\text{volume}(\text{ellipsoid}(S_J^*)) \leq \text{volume}(\text{ellipsoid}(S_K^*)) \quad (2)$$

for all observation subsets $K \neq J$ with $\#(J) = \#(K) = h$.

Concretely, X_i denotes the i^{th} row (or assay sample) in the observation matrix X . We interpret J as a collection of h row vectors from X such that the ellipsoid fitted to the data (i) passes through $p + 1$ data points, (ii) contains the h observations, and (iii) has the smallest volume among all sets K with cardinality h that also satisfy conditions (i) and (ii). By construction, outliers must be excluded from the set J . Typically, $h = [(n + p + 1)/2]$.

Minimum Covariance Determinant

The minimum covariance determinant (MCD) estimator finds the robust mean and covariance from $h(\leq n)$ points that minimize the determinant of the covariance matrix associated with the subsample. It looks for an ellipsoid with the smallest volume that encompasses the h data points. Formally, $MCD(X) \rightarrow (\bar{\mathbf{x}}_J^*, S_J^*)$ where the optimal set $\{X_i | X_i \in J\}$ satisfies

$$|S_J^*| \leq |S_K^*| \quad \forall \text{ subsets } K \neq J \text{ with } \#(J) = \#(K) = h. \quad (3)$$

In other words, J minimizes the volume of the ellipsoid among all subsets K that contain h observations from X . Note: $|S_J^*|$ is used as a proxy measure since the actual volume is proportional to $\det(S_J^*)$.

Obtaining the center and covariance estimates in the Mahalanobis distance using the MCD estimator produces a robust distance measure that is not skewed by outliers. Meanwhile, outliers will continue to possess large values based on this robust distance measure. The most common cutoff value k selected for the $D^2(\mathbf{x}, \bar{\mathbf{x}}) > k$ test is again based on chi-square critical values. The chi-square Q-Q plot provides a useful way to visually assess whether distances are χ_p^2

distributed. To generate a plot of robust distance versus estimated percentiles, data points are ordered according to their Mahalanobis distances, and paired up with an ordered sequence of χ_p^2 percentiles which may be obtained by drawing n samples from a chi-square distribution with p degrees of freedom. If the data is from a multivariate normal distribution, then the data points should be on a straight line. However, points with a higher Mahalanobis distance than the chi-square quantiles do not follow a straight line trend, thus they are treated as outliers.

Application: Multivariate Outlier Detection in the Context of Iron Ore Mining in Western Australia

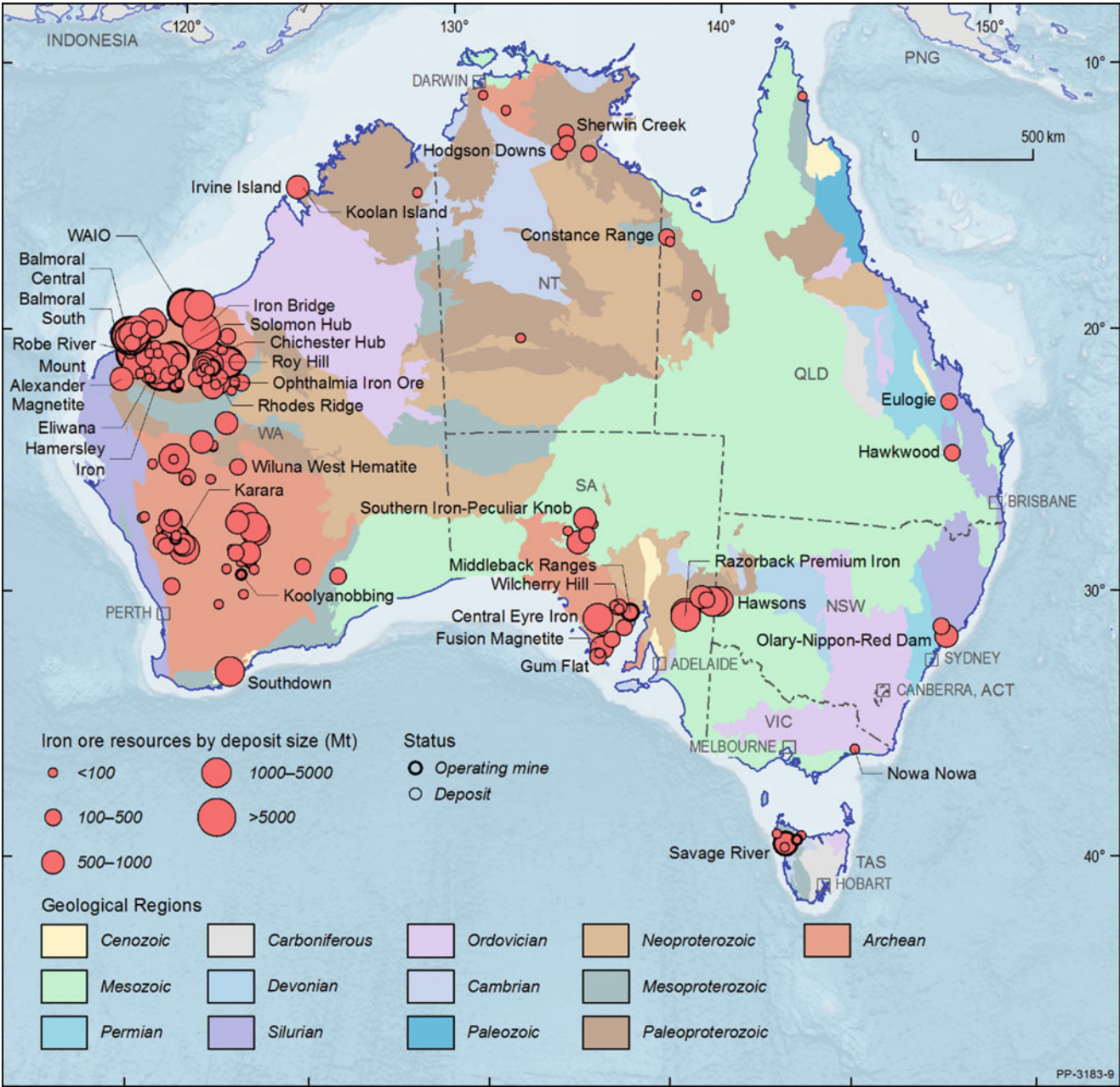
In mining, assay analysis is performed routinely to determine the geochemical composition of subterranean ore bodies. This is achieved by sampling material extracted from drilled holes. In the context of iron ore mining in the Pilbara region in Western Australia (Fig. 2), the chemical components of interest include Fe, SiO₂, Al₂O₃, P, LOI, S, MgO, Mn, TiO₂, and CaO. These measurements constitute a point cloud in a high-dimensional space which makes it more difficult to articulate meaningful differences between samples that belong to different geological domains.

Balamurali and Melkumyan (2015) evaluated the standard classical Mahalanobis distance method and the robust distance methods: *minimum volume ellipsoid* (MVE) and *minimum covariance determinant* (MCD), to find the presence of outliers in the given geological domain. The study further evaluated the methods with different contamination levels and concluded that the robust methods outperform the classical Mahalanobis distance-based approach. However, the performance of robust methods tends to decrease with increasing contamination levels. The authors introduced ratios between the weight percentage (wt%) of chemical grades: SiO₂/Al₂O₃, LOI/Al₂O₃, and TiO₂/Fe as additional variables, and observed a significant increase in the outlier detection rate even with a high contamination level (Fig. 3).

In addition, Gaussian process classification and the chi-square Q-Q plot were tested to gain insight with respect to finding abnormal data points. Not surprisingly, some of the multivariate outliers identified were not detected as univariate outliers. On the other hand, univariate outliers pertaining to trace elements (nondominant chemical components) may trigger false positives when the samples themselves are not considered multivariate outliers.

Ensembles of Subsets of Variables

In order to further understand the influence of raw chemical variables and the ratios between the variables in outlier detection, a novel method of outlier detection was proposed in (Balamurali and Melkumyan 2018) using ensembles of



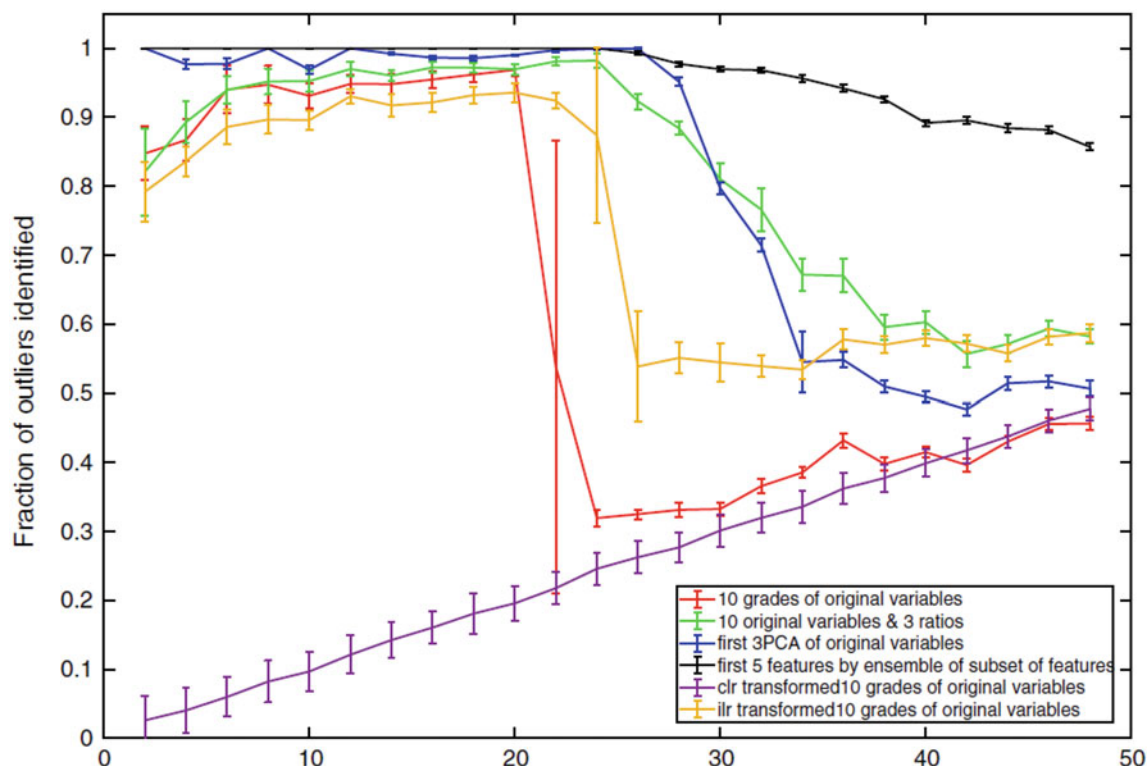
Statistical Outliers, Fig. 2 Geological overview of the Pilbara region in Western Australia. The economically valuable banded-iron formations are hosted mainly by the Hamersley Group of the Mount Bruce Supergroup © Commonwealth of Australia (Geoscience Australia) 2021

subsets of variables (EoSV). This study extended the previous work by incorporating a variable selection method. Similar to the work that incorporated three additional variables, in this work, all possible combinations of ratios between the raw chemical variables were used as additional variables. The Tree Bagger (Breiman 1998) method was used to combine features that have a large discrimination power at different contamination levels. Previous studies recognized that feature engineering can significantly improve training and model prediction outcomes (Mitra et al. 2002; Saeys et al. 2007).

The EoSV method demonstrated that the relevance of features varies with contamination level in the outlier classification problem. Hence, the stability of the influence of features was assessed and a method was proposed to keep only the features with stable influences in outlier detection.

t-Distributed Stochastic Neighbor Embedding (t-SNE) for Outlier Detection

Although multivariate outlier identification is important, it is a means rather than an end. Very often, there is value in



Statistical Outliers, Fig. 3 Comparing robust outlier detection methods using MCD with different feature combinations (Reference: Balamurali and Melkumyan (2018))

retaining the outliers for further analysis. When the outliers are viewed in the context of the raw and filtered data, spatial clustering may emerge. These clusters may suggest the existence of latent domains and highlight locations where the existing domain boundary is possibly incorrect.

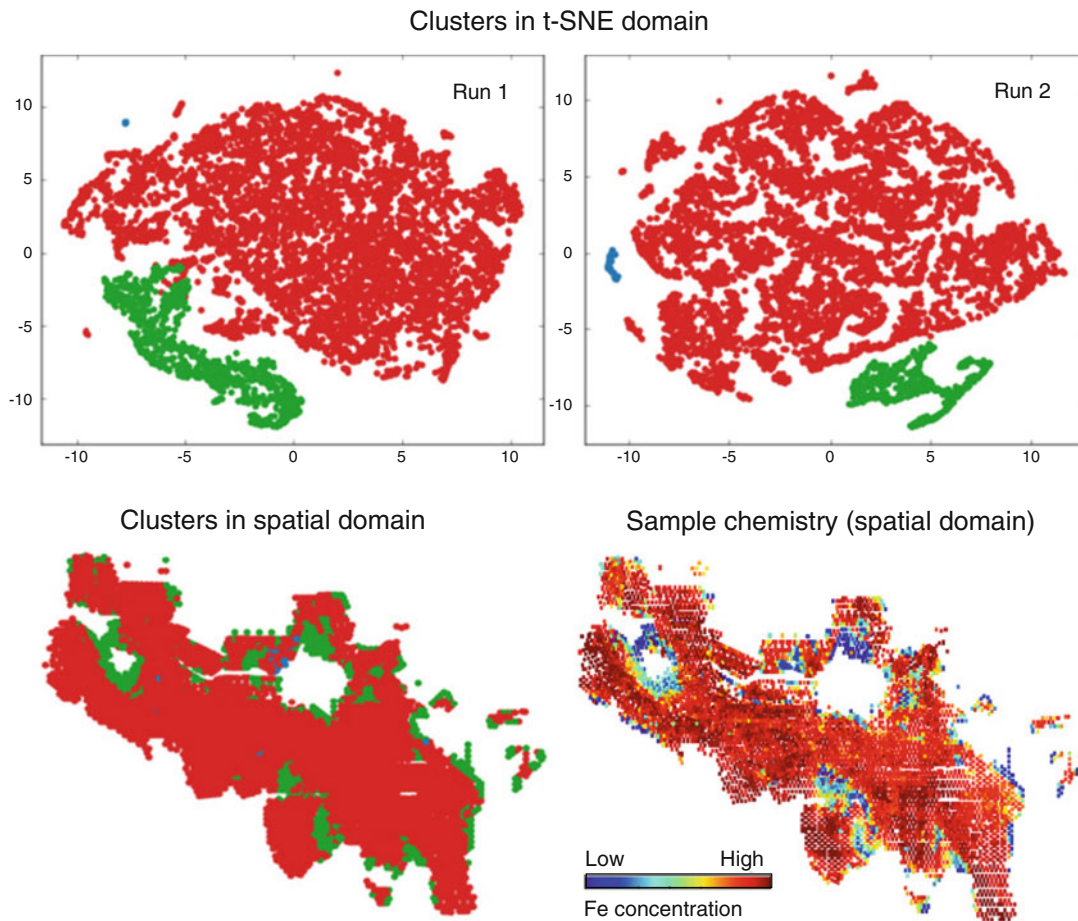
Figure 4 demonstrates a successful application of t-SNE (Van der Maaten and Hinton 2008) to outlier identification. The proposed method combines the t-SNE dimensionality reduction technique with spectral clustering (Von Luxburg 2007) to uncover persistent outliers. Figure 4 (top) shows two random realizations of the t-SNE and the chemical signatures of two clusters in the t-SNE domain. The smaller cluster corresponds to outliers sampled from a spatially adjacent geological domain (see Fig. 4 bottom). These green points are chemically distinct from the main population (target domain) and they are spatially correlated. By projecting the assay data to two or three dimensions, t-SNE allows geochemical differences between the main population and outliers to be visualized in a scatter plot and meaningful geological domain clusters to be discerned (Balamurali and Melkumyan 2016).

Although t-SNE was successfully used in identifying sub-populations, because of its stochastic nature, t-SNE provides different results in different runs (Fig. 4). In order to produce stable results, the authors proposed a novel methodology for

ensemble clustering based on t-distributed stochastic neighbor embedding (EC-tSNE) to extract persistent patterns that more reliably correlate with geologically meaningful clusters (Balamurali and Melkumyan 2020). The proposed method uses silhouette scores to select the optimal number of clusters to cancel out inconsistency. EC-tSNE uses a consensus matrix (Monti et al. 2003) and agglomerative hierarchical clustering to achieve this. A map of multivariate outliers (see Fig. 4 bottom-left) can clearly identify the “contaminated” sites where samples are collected outside the target domain and potential contamination (or dilution) of high-grade ore material may occur.

MCD Outlier Detection with Sample Truncation Strategy

Following the experiments on EC-tSNE, Leung et al. (2019) proposed two sample truncation methods: *maximum hull distance* and *maximum silhouette*, to reject outliers in multivariate geochemical data. This builds on the foundation where two essential points remain unchanged: (i) the distance between samples is computed using robust estimators and (ii) outliers manifest as points that are a long distance away from the consensus values such as the robust mean obtained from the sampled population. The key observation is that outliers are relegated to the upper tail region when points are sorted by their robust distances in ascending order and



Statistical Outliers, Fig. 4 Top row: Results from t-SNE spectral clustering are displayed in t-SNE coordinates for two random runs. Bottom row (left): corresponding results in spatial coordinates; (right)

the normalized Fe grade of samples determined by geochemical assays (Reference: Leung et al. (2019))

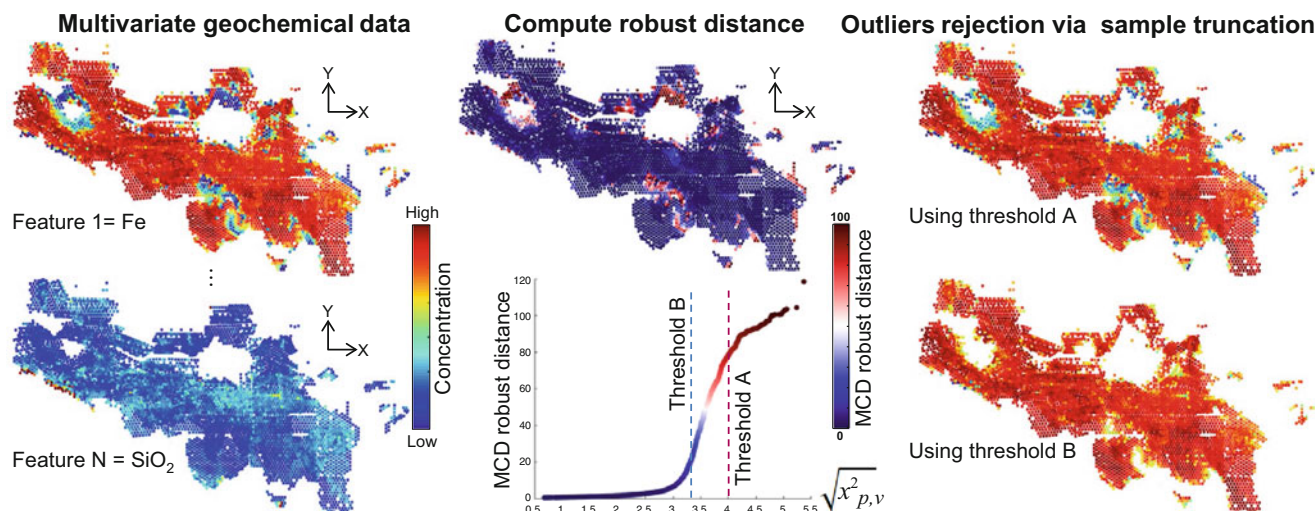
plotted in a chi-square quantile plot. Sample truncation simply refers to the action of removing samples from this ordered sequence beyond a certain cutoff point (see Fig. 5). This approach allows the outlier removal decision to be made after the points are sorted by their robust distances; it also provides flexibility as the rejection threshold may be chosen by a domain expert or computed automatically to reduce false negatives, for instance. Traditionally, a fixed threshold based on the chi-square critical value, $\chi^2_{p,1-\alpha}$, is used for MCD. For compositional data analysis, this choice may result in sub-optimal outlier rejection. Figure 5 illustrates the improvement that may be obtained using one of the proposed MCD sample truncation strategies, where threshold A and B correspond to choosing $D \leq \sqrt{\chi^2_{p,1-\alpha}}$ and the proposed cut-off value, respectively. For the former, the extent of outlier removal is clearly inadequate. The outliers identified using MCD were compared to the EC-tSNE results in (Leung et al. 2019) and the MCD sample truncation strategies were recommended as

an alternative to EC-tSNE since t-SNE is computationally costly even for datasets of a moderate size ($n > 10$ k).

Summary

This chapter described the fundamental characteristics of multivariate outliers, focusing specifically on a scenario most often encountered during mining whereby outliers result from collecting assay samples from an adjacent rather than the intended domain due to inaccurate domain boundary definitions. The presence of outliers can impact the performance of grade prediction models and this has far-reaching consequences for grade control and mine planning as it affects target conformance, process efficiency, and profitability of a mining operation.

In the technical section, univariate and multivariate outlier identification methods were reviewed. This covered Tukey's method, the use of Mahalanobis distance, and robust



Statistical Outliers, Fig. 5 Outlier rejection based on MCD robust distance and threshold selection (Reference: Leung et al. (2019))

alternatives based on minimum volume ellipsoid (MVE). Additionally, more recent approaches based on ensemble clustering t-distributed stochastic neighbor embedding (EC-tSNE) and MCD robust distances were introduced. The importance of feature selection was highlighted. Using real geochemical assay data collected from a Pilbara iron ore mine; the case study demonstrated successful extraction of outliers using both EC t-SNE and MCD robust distances. When spatially coherent outlier clusters emerge, latent domains and the areas where domain boundary delineation is poor can both be identified. This can give potentially new insight and local knowledge into the structural geology.

While there is no universal outlier detection method that works in all circumstances, the MCD robust distance approach combined with appropriate threshold selection strategies seems to offer promising results, in terms of efficiency and efficacy. At the very least, it allows outliers to be segregated and, if necessary, modeled as a separate process (or latent domain) during grade estimation.

Cross-References

- [Geologic Time Scale Estimation](#)
- [Mine Planning](#)
- [Multiple Point Statistics](#)
- [Object Boundary Analysis](#)
- [t-Distributed Stochastic Neighbor Embedding](#)
- [Tukey, John Wilder](#)

Acknowledgments This work has been supported by the Australian Centre for Field Robotics and the Rio Tinto Centre for Mine Automation, the University of Sydney.

Bibliography

- Balamurali M, Melkumyan A (2015) Multivariate outlier detection in geochemical data. In: Proceedings of the IAMG conference, vol 17. International Association of Mathematical Geosciences, Freiberg, pp 602–610
- Balamurali M, Melkumyan A (2016) t-SNE based visualisation and clustering of geological domain. In: International conference on neural information processing. Springer, Berlin, pp 565–572
- Balamurali M, Melkumyan A (2018) Detection of outliers in geochemical data using ensembles of subsets of variables. *Math Geosci* 50(4): 369–380. <https://doi.org/10.1007/s11004-017-9716-8>
- Balamurali M, Melkumyan A (2020) Computer aided subdomain detection using t-sne incorporating cluster ensemble for improved mine modelling. *International Association for Mathematical Geosciences* (under review)
- Barnett V, Lewis T (1994) Outliers in statistical data. Wiley series in probability and mathematical statistics applied probability and statistics. Wiley, New York
- Breiman L (1998) Arcing classifier (with discussion and a rejoinder by the author). *Ann Stat* 26(3):801–849. <https://doi.org/10.1214/aos/1024691079>
- Daisy Summerfield “;Australian Resource Reviews: Iron Ore 2019”, Geoscience Australia. Available at: <https://www.ga.gov.au/scientific-topics/minerals/mineral-resources-and-advice/australian-resource-reviews/iron-ore>
- De Maesschalck R, Jouan-Rimbaud D, Massart DL (2000) The Mahalanobis distance. *Chemom Intell Lab Syst* 50(1):1–18
- Filzmoser P, Garrett RG, Reimann C (2005) Multivariate outlier detection in exploration geochemistry. *Comput Geosci* 31(5):579–587
- Garrett RG (1989) The chi-square plot: a tool for multivariate outlier recognition. *J Geochem Explor* 32(1–3):319–341
- Gnanadesikan R, Kettenring JR (1972) Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* 28:81–124
- Grubbs FE (1969) Procedures for detecting outlying observations in samples. *Technometrics* 11(1):1–21
- Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (2011) Robust statistics: the approach based on influence functions, vol 196. Wiley, New York
- Hawkes HE, Webb JS (1962) Geochemistry in mineral exploration. Harper & Row, New York

- Iglewicz B, Hoaglin D (1993) How to detect and handle outliers (the ASQC basic reference in quality control). In: Mykytka EF (ed) Statistical techniques. American Society for Quality, Statistics Division, Milwaukee
- Leung R, Balamurali M, Melkumyan A (2019) Sample truncation strategies for outlier removal in geochemical data: the MCD robust distance approach versus t-SNE ensemble clustering. *Math Geosci*:1–26. <https://doi.org/10.1007/s11004-019-09839-z>
- Leung R, Lowe A, Chlingaryan A, Melkumyan A, Zigman J (2022) Bayesian surface warping approach for rectifying geological boundaries using displacement likelihood and evidence from geochemical assays. *ACM Transactions on Spatial Algorithms and Systems* 18(1). <https://doi.org/10.1145/3476979>. Preprint available at <https://arxiv.org/abs/2005.14427>
- Mahalanobis PC (1936) On the generalized distance in statistics. *National Institute of Science of India*, vol 2, pp 49–55. http://bayes.acs.unf.edu:8083/BayesContent/class/Jon/MiscDocs/1936_Mahalanobis.pdf
- Mitra P, Murthy C, Pal SK (2002) Unsupervised feature selection using feature similarity. *IEEE Trans Pattern Anal Mach Intell* 24(3): 301–312
- Monti S, Tamayo P, Mesirov J, Golub T (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach Learn* 52(1):91–118
- Reimann C, Filzmoser P, Garrett RG (2005) Background and threshold: critical comparison of methods of determination. *Sci Total Environ* 346(1–3):1–16
- Rousseeuw PJ (1985) Multivariate estimation with high breakdown point. *Math Stat Appl* 8(283–297):37
- Rousseeuw PJ, Leroy AM (1987) Robust regression and outlier detection. *Wiley series in probability and mathematical statistics*. <https://doi.org/10.1002/0471725382>
- Saeyns Y, Inza I, Larranaga P (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics* 23(19):2507–2517
- Strehl A, Ghosh J (2002) Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *J Mach Learn Res* 3(Dec):583–617
- Tukey JW (1977) *Exploratory data analysis*, vol 2. Pearson, Reading
- Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9(11):2579–2605
- Von Luxburg U (2007) A tutorial on spectral clustering. *Stat Comput* 17(4):395–416

Statistical Quality Control

Jörg Benndorf
TU Bergakademie Freiberg, Freiberg, Germany

Definition

Statistical quality control (SQC) is the application of statistical methods and optimization to monitor and maintain the quality of products and services. Two branches of SQC are distinguished. The first one is concerned about the quality of particular product items and refers to the name acceptance sampling. Here, a decision is made to either accept or reject the items based on their quality. The second branch is

concerned about controlling the process in such a way to guarantee the on spec quality of the products and is referred to as statistical process control. It typically includes the steps (1) establishment of expected performance indicators, (2) targeted process monitoring, (3) comparison between observed and expected performance, and (4) corrective adjustment of process control parameters and decisions.

Although, initially developed for manufacturing industry in the 1920s and then intensively applied during World War II, the principles of SQC found its way in many non-manufacturing applications, also in earth sciences (e.g., Montgomery 2020). As a part of a quality management system to fulfill quality requirements, today, SQC can be a key constituent to implement ISO 9000/ISO 9001 – quality management system standards.

Acceptance Sampling

The idea of acceptance sampling is to choose a representative sample of a product lot and test for defects. If the number of defects exceeds a certain acceptance criterion, the product lot is rejected. The underlying sampling plan is based on a hypothesis test and involves the following steps:

- Step 1: Definition of the size of the product lot to be samples
- Step 2: Definition of a representative random sample size for the lot
- Step 3: Definition of an acceptance criterion for not rejecting the lot (number of defects)

The null-hypothesis H_0 is that the lot is of acceptable quality. The alternative hypothesis H_A is that the lot is not of acceptable quality. There are two types of risks associated with acceptance sampling:

- The lot is rejected although it has the acceptable quality. This is referred to the type I error in statistical testing and is quantified by the error probability α .
- The lot is accepted although it is not of acceptable quality. This is referred to the type II error in statistical testing and is quantified by the error probability β .

A main decision in acceptance sampling is the number of required samples necessary to guarantee acceptable probabilities α and β while still reacting on a desired deviation limit. This number is defined based on the understanding of the product variability, the uncertainty in monitoring using the test statistical value. An example in a later section demonstrates this process.

Statistical Process Control

Statistical Process Control (SPC) is a methodology for measuring and controlling quality during the manufacturing process. It requires that quality data are obtained in the form of product or process observations in regular intervals or in real-time during manufacturing. This data is then plotted on a graph with predetermined control limits, the so-called control chart. Control limits are determined by the capability of the process, whereas specification limits are determined by the client's needs. A typical example of such a control chart is given in Fig. 1 for quality control of run-of-mine (ROM) ore. Control limits correspond to the 2σ and 3σ confidence intervals.

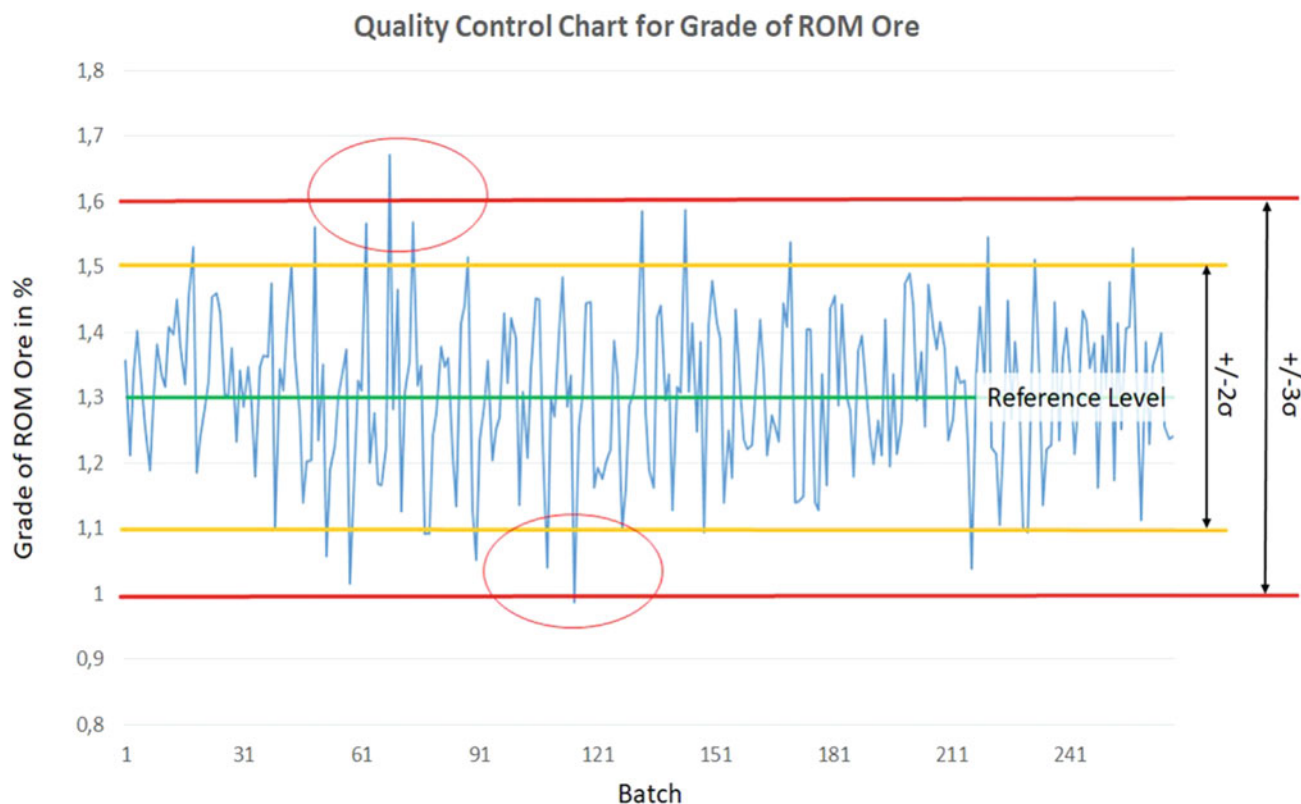
According to the founder of this technique, W.A. Shewhart, the variation observed within the process results from two sources (Shewhart 1931):

- Superposition of a number of sources of small variation and background noise (unavoidable)
- Large variability resulting from inaccuracies in process control caused by process defects such as imprecise working equipment or human error (avoidable)

A reactive process control is managed by the means of the control chart and the defined upper and lower limits. Typically and as illustrated in Fig. 1, limits are derived by statistical inspection of the sample population. The process is deemed "in control" if sample measurements are between the limits. Observations, which are close to the limits or even exceed these indicate the likely occurrence of avoidable causes within the process and call for a root cause analysis and corresponding corrective action. In this way, a so-called iterative Shewhart-cycle can be implemented, which consists of the following steps:

- Plan: Plan and design the process and establish expectations in form of process indicators.
- Do: Implement the design.
- Check: Monitor the process and identify differences between expectations and outcomes or statistical anomalies.
- Act: Adjust the process if deviations are significant.

This iterative cycle is often used to continuously monitor and adjust the process to guarantee the possible improvement of product quality.



Statistical Quality Control, Fig. 1 Example of a control chart

Application of Acceptance Sampling in Environmental Monitoring

One goal of environmental monitoring for any industrial activity is to check the compliance of emission or contamination levels with respect to restrictions imposed by permitting authorities or legislation. Restrictions can be defined in a way that the background contamination existing before the industrial activity starts should not increase significantly. In a sense of statistical quality control, environmental monitoring can be interpreted as a process to take a set of representative samples of a contamination or emission variable within a possible impacted area or domain. This is done to test if a contamination level at a given point in time is statistically and significantly increased compared to the background contamination. Sampling is repeated regularly as imposed by the application dynamics or the permitting requirements and also to understand possible timely trends.

An essential task of this type of environmental monitoring is to define the appropriate number of samples in space and in time to meet a desired quality of the applied hypothesis test (e.g., De Gruijter et al. 2006). The quality of the hypothesis test is defined by the acceptable levels of risks associated with the type I and type II error in acceptance sampling. Thus, the required number of samples is a direct function of the frequency distribution of the variable considered within the domain, the threshold of interest, and the error probabilities α and β are related to the statistical test.

Example

A simple example shall illustrate the reasoning behind this type of application. As part of an environmental impact assessment, the average Cd content of soil within a possible impacted area by mining activities is determined during baseline monitoring by a set of 50 randomly selected independent samples. The mean value of the data and the variance of the data are calculated according to Eqs. (1) and (2).

$$\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i \quad (1)$$

$$\sigma^2(z) = \frac{1}{(n-1)} \sum_{i=1}^n (z_i - \bar{z})^2 \quad (2)$$

The standard deviation of the mean value is

$$\sigma(\bar{z}) = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^n (z_i - \bar{z})^2} \quad (3)$$

Applied to the sample values, a mean Cd content of $\bar{z}_0=0.20$ mg/kg with a standard deviation (mean) of $\sigma(\bar{z}_0) = \pm 0.02$ mg/kg has been obtained. The latter has been derived

based on the standard deviation of the samples of $\sigma(z)=\pm 0.14$ mg/kg.

After some months of operation, during a first monitoring epoch, the mean value has been determined again. This time, the mean Cd content is $\bar{z}_1 = 0.23$ mg/kg with a comparable standard deviation of the mean value $\sigma(\bar{z}_1) = \pm 0.02$.

The first obvious question is whether the Cd content has statistically significantly increased in comparison to the baseline monitoring or in other words: What is the difference between both mean values of systematic or random nature? To answer this question, we can start with the following thoughts:

The difference between the two mean values is distributed with an expectation of

$$\Delta = \mu_1 - \mu_0, \quad (4)$$

where μ_0 and μ_1 are the true mean values of the Cd content for the two epochs, and a standard deviation of

$$\sigma_{\bar{z}_1 - \bar{z}_0} = \sqrt{\sigma^2(\bar{z}_1) + \sigma^2(\bar{z}_0)}. \quad (5)$$

The resulting normalized standard variable is

$$z = \frac{\bar{z}_1 - \bar{z}_0 - (\mu_1 - \mu_0)}{\sqrt{\sigma^2(\bar{z}_1) + \sigma^2(\bar{z}_0)}} \quad (6)$$

To test, if the difference is of systematic or random nature, a null hypothesis

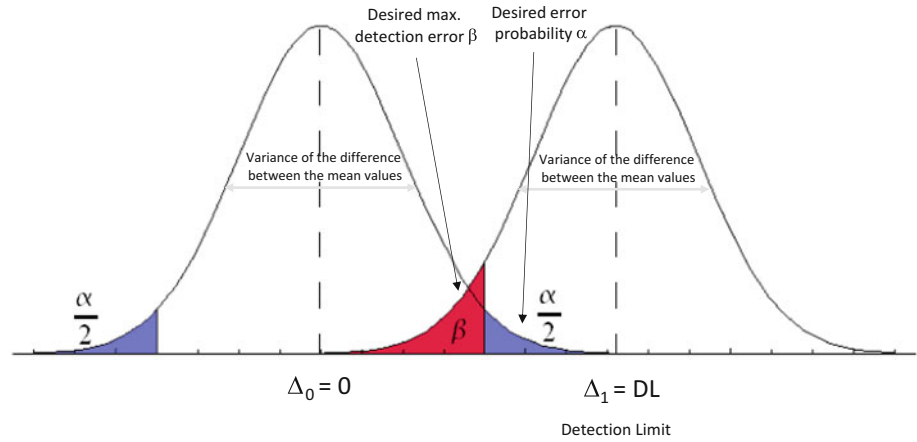
$$H_0 : \Delta_0 = \mu_1 - \mu_0 = 0 \quad (7)$$

can be formulated that assumes a zero difference between the mean Cd content of the baseline monitoring and monitoring epoch 1. The corresponding test statistical value is

$$z = \frac{\bar{z}_1 - \bar{z}_0 - (\Delta_0)}{\sqrt{\sigma^2(\bar{z}_1) + \sigma^2(\bar{z}_0)}} \quad (8)$$

This statistical test value is to be compared to the quantile of the standard normal distribution u that corresponds to the desired error probability α related to the type I error in statistical testing. Depending on whether the focus is on a change in general or a directed change (increase or decrease), a quantile of the standard normal distribution is chosen. This corresponds to either the $100-\alpha/2$ confidence interval or the $100-\alpha$ interval.

For the numeric example, we are interested in a change in general of the mean Cd content. With a desired error probability $\alpha = 5\%$, the corresponding quantile of the standard

Statistical Quality Control,**Fig. 2** Hypothesis test for change detection

normal distribution is $u_{100-\alpha/2}=1.96$. This value is compared to the test statistical value, which for the numerical example is $z = 1.06$. H_0 is accepted, if $z < u_{100-\alpha/2}$. Thus, in a sense of acceptance sampling, the quality of the product (the impact of the mining activity on the Cd content) is within the limits.

The second question is now related to the required number of samples to detect a change within a predefined detection limit with related error probabilities α and β . Figure 2 illustrates the case for a two-sided problem, that means a change in any direction, positive or negative, is to be detected.

The detection limit Δ_1 is equal to the difference between the true mean values of the Cd content $\mu_1 - \mu_0$. According to Fig. 2, the detection limit DL is obtained by summing up the quantiles of the confidence interval $100 - \alpha/2$ related to H_0 and β related to the alternative hypothesis H_A . Thus the detection limit is

$$DL = u_{100-\alpha/2} \cdot \sqrt{\sigma^2(\bar{z}_1) + \sigma^2(\bar{z}_0)} - u_\beta \cdot \sqrt{\sigma^2(\bar{z}_1) + \sigma^2(\bar{z}_0)}. \quad (9)$$

$u_{100-\alpha/2}$ quantile of the standard normal distribution related to a two-sided confidence level of $100 - \alpha$

u_β quantile of the standard normal distribution related to an error probability β (detection error)

Assuming $\alpha = 5\%$ and $\beta = 5\%$ and the standard deviation of the mean value for each epoch of being equals with $\sigma(\bar{z}) = \pm 0.02$ mg/kg, the detection limit is $DL = 0.10$ mg/kg.

What is now the minimum number of data n required to guarantee a certain detection limit, say $DL = 0.05$ mg/kg?

First, we assume that standard deviations of the mean value remain the same for two epochs

$$\sigma^2(\bar{z}_1) = \sigma^2(\bar{z}_2) = \sigma^2(\bar{z})$$

Solving Eq. (9) with respect to $\sigma^2(\bar{z})$ results in

$$\sigma(\bar{z}) = \frac{1}{\sqrt{2}} \left(\frac{DL}{u_{100-\alpha/2} - u_\beta} \right). \quad (10)$$

For the numerical example, a necessary standard deviation of the mean of $\sigma(\bar{z}) = \pm 0.01$ mg/kg is required. According to the relationship

$$\sigma(\bar{z})^2 = \frac{1}{n} \sigma^2(z) \quad (11)$$

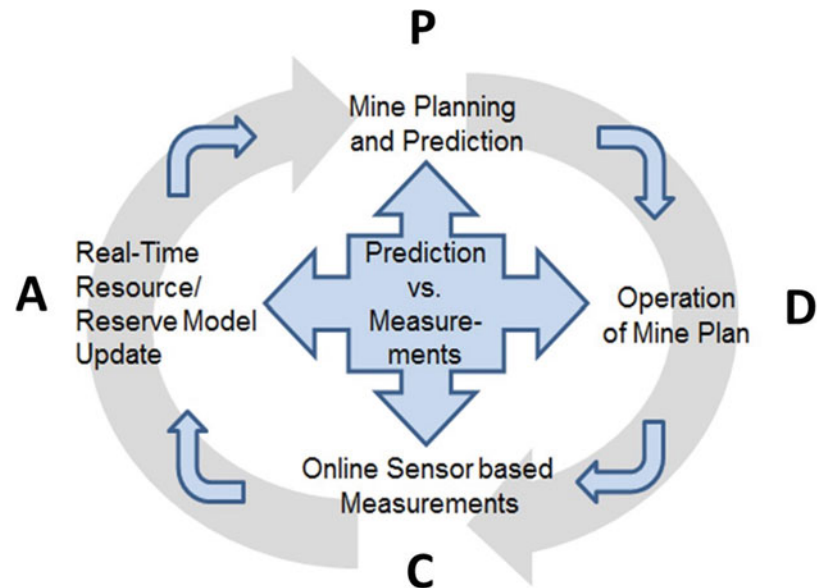
and taking into account the standard deviation of the data of $\sigma(z) = \pm 0.14$ mg/kg, the required number of samples is approximately 200.

Application of SPC in Grade Control in Mining

A typical application of statistical quality control in the field of geosciences and geoenvironment is ore quality or grade control in mineral resource extraction. Benndorf (2020) documents a closed-loop framework to implement a real-time process monitoring and control framework for grade control in mineral resource exploitation. This is based on the plan-do-check-act (PDCA) iterative management cycle as suggested by Shewhart. It is general and applicable for surface mining and underground operations and can be interpreted as follows:

P – Plan and predict: Based on the mineral resource and grade control model, strategic long-term mine planning, short-term scheduling, and production control decisions are made. Performance indicators such as expected ore tonnage extracted per day, expected ore quality attributes, and process efficiency are predicted.

Statistical Quality Control,
Fig. 3 The real-time mining
 concept. (After Benndorf and
 Jansen 2017)



D – Do: The mine plan is executed.

C – Check: Production monitoring systems continuously deliver data about process indicators using modern sensor technology. For example, the grade attributes of the ore extracted are monitored using cross-belt scanners. Differences between model-based predictions from the planning stage and actual measured sensor data are detected.

A – Act: Differences between prediction and production monitoring are analyzed and root causes investigated. One root cause may be the uncertainty associated with the resource or grade control model to predict the expected performance. Another root cause may be the precision and accuracy of sensor measurements. Using innovative data assimilation methods (e.g., Wambeke et al. 2018), differences are then used to update the resource model and mine planning assumptions, such as ore losses and dilution. With the updated resource and planning model, decisions made in the planning stage may have to be reviewed and adjusted in order to maximize the process performance and meet production targets through the use of sophisticated optimization methods (e.g., Paduraru and Dimitrakopoulos 2018) (Fig. 3).

The two key requirements in this application of statistical process control to real-time grade control are the ability to update the resource/grade control model quickly and to derive fast new decisions based on the new model. Key enablers are techniques of data assimilation for resource model updating and mine planning optimization.

Conclusions

Natural variability is a key characteristic of any spatial or temporal process or variable considered in earth sciences.

Due to the interaction of humans with the geosphere, variables may change significantly over time. The ability to monitor impacts of anthropogenic activities with respect to environmental or economic indicators requires distinguishing between natural variability and systematic change. Techniques of SQC allow for a statistically sound judgment of this change by designing appropriate sampling strategies, not only to detect change but also to take corrective action in case some indicators drift away. The availability of online sensors and sensor networks in earth observation allows today for the implementation near real-time closed-loop concepts for transparent impact monitoring and process control. In this context, results of SQC provide the means of an objective and transparent communication of project impact.

Bibliography

- Benndorf J (2020) Closed loop management in mineral resource extraction. Springer, Cham. <https://doi.org/10.1007/978-3-030-40900-5>
- Benndorf J, Jansen JD (2017) Recent developments in closed-loop approaches for real-time mining and petroleum extraction. *Math Geosci* 49(3):277–306
- De Gruijter J, Brus DJ, Bierkens MF, Knotters M (2006) Sampling for natural resource monitoring. Springer
- Montgomery DC (2020) Introduction to statistical quality control. Wiley
- Paduraru C, Dimitrakopoulos R (2018) Adaptive policies for short-term material flow optimization in a mining complex. *Mining Technol* 127(1):56–63
- Shewhart WA (1931) Economic control of quality of manufactured product. Van Nostrand, New York
- Wambeke T, Elder D, Miller A, Benndorf J, Peattie R (2018) Real-time reconciliation of a geometallurgical model based on ball mill performance measurements – a pilot study at the Tropicana gold mine. *Mining Technol* 127(3):115–130

Statistical Rock Physics

Gabor Korvin

Earth Science Department, King Fahd University of
Petroleum and Minerals, Dhahran, Kingdom of Saudi Arabia

Definition

Statistical Rock Physics is a part of Rock Physics (Petrophysics) where we use concepts, methods, and computational techniques borrowed from Stochastic Geometry and Statistical Physics to describe the interior geometry of rocks; to derive their effective physical properties based on their random composition and the random arrangement of their constituents; and to simulate geologic processes that had led to the present state of the rock.

Introduction

This chapter is about the novel computational tools used in Rock Physics, based on concepts, methods, and computational techniques borrowed from Stochastic Geometry and Statistical Physics. Most techniques that will be discussed belong to the toolbox of the “Digital Rock Physics” (DRP) technology of Core Analysis or the “Statistical Rock Physics” (SRP) procedure of Seismic Reservoir Characterization. My aim has been to help the reader to understand the claims and published results of these powerful new technologies, and – most importantly – to creatively apply stochastic geometry and statistical physics in their own research to other fields of rock physics. For this purpose, the theory discussed will be illustrated on diverse applications: hydraulic permeability, rock failure, NMR, contact forces in random sphere packings, compaction of shale, HC maturation. There are five sections, each with a few applications. Section “[Stochastic Geometry](#)”, introduces Gaussian Random Fields and the Kozeny-Carman equation (1.1), and then gives three applications: specific surface from 2D microscopic image, estimating tortuosity, and reconstructing 3D pore space from a 2D image (sections “[ACF \(Autocorrelation Function\) Based Estimates of the Specific Surface](#)”, “[The Random Geometry of Tortuous Flow Paths](#)”, and “[Reconstructing 3D Pore Space from 2D Sections](#)”). Section “[Maximum Entropy Methods](#)” is devoted to Maximum Entropy (ME). ME is explained in section “[The Maximum Entropy Method](#)”; section “[Maximum Entropy \(ME\) Pore Shape Inversion](#)” applies ME to pore-shape determination; section “Shale Compaction and Maximal Entropy” derives Athy’s empirical law of shale compaction from the ME principle; section “Contact Force Distribution in Granular Media” is about contact-force distribution in loose

granular packs. Section “[Self-Consistency and Effective Media](#)” treats the Effective Medium and Differential Effective Medium methods. EM is introduced on the example of random resistor networks in section “[Effective Medium \(EM\) Approximation](#)”, and DEM is illustrated by the iterative computation of the effective elastic moduli of porous or cracked rock in section “[Differential Effective Medium \(DEM\) Approximation](#)”. Section “[Thermodynamic Algorithms](#)” summarizes three important thermodynamic algorithms: random walk in section “[Random Walk](#)” (applications: spin magnetic relaxation, and electric conductivity in porous rocks); Lattice Boltzmann Method in section “[Lattice Boltzmann Methods for Flow in Porous Media](#)” (application: hydraulic flow); and Simulated Annealing (already introduced in section “[Reconstructing 3D Pore Space from 2D Sections](#)” as a technique for pore space reconstruction) is applied in section “[Simulated Annealing](#)” to find the thermal conductivity distribution of soil constituents. The last part, section “[Computational Phase Transitions](#)”, is on Computational Phase Transitions: section “[Percolation Theory](#)” deals with Percolation Theory (application: permeability of kaolinite-bearing sandstones); section “[Renormalization Group \(RNG\) Models of Rock Failure](#)” with the Renormalization Group (RNG) approach to rock failure; section “[Discrete Scale Invariance](#)” with determining the critical point of phase transition (based on Discrete Scale Invariance); the final section “[Arrhenius Law and Source-Rock Maturity](#)” is about how Arrhenius Law of Physical Chemistry is applied to estimate the maturation of source rocks.

Stochastic Geometry

Random Functions Along a Line, Random Fields, Kozeny-Carman Equation

A *random function*, $\{f(x)\}_\alpha$, is a family of functions depending on a random parameter α , where the independent variable x varies along some line. A given $f(x)$ picked at random from among all possible $\{f(x)\}_\alpha$ s is a *realization*. At some fixed point on the line x_1 , $f(x_1)$ is a *random value*, and it attains different values y with the probabilities $\text{Prob}[f(x_1) < y] = F(x_1, y)$. The following expected values (taken with respect to α , i.e., over all realizations of $\{f(x)\}_\alpha$) are often enough to characterize a random function: $\langle f(x_1) \rangle$, $\langle f^2(x_1) \rangle$, $\langle f(x_1) \cdot f(x_2) \rangle$, termed *mean value*, *mean square value*, and *autocorrelation function*. A random function is *translation invariant* if its statistical properties do not change with respect to a shift along the line; in particular, if $\langle f(x) \rangle = \langle f(0) \rangle$; $\langle f^2(x) \rangle = \langle f^2(0) \rangle$; $\langle f(x_1) \cdot f(x_2) \rangle = \langle f(0) \cdot f(x_2 - x_1) \rangle := R_{ff}[x_1 - x_2]$ where the function R_{ff} is the *autocorrelation function (ACF)* of $f(x)$. It is an *even function*, $R(\xi) = R(-\xi)$, assuming its maximum at $\xi = 0$, $R_{ff}(0) = \langle f^2(x) \rangle$. The *normalized autocorrelation function* is $\rho_{ff}(\xi) = R_{ff}(\xi)/\langle f^2(x) \rangle$.

If $\mathbf{x}(x,y)$ or $\mathbf{x}(x,y,z)$ is a point in two- resp. three-dimensional Euclidean space, then $\{f(x,y)\}_a$ resp. $\{f(x,y,z)\}_a$ are called *random fields* over the plane, or space, a given field $f(\mathbf{x})$ picked out at random is a *realization* of the field.

A random field is *homogeneous* if its statistical properties are invariant with respect to a shift, that is if $\langle f(\mathbf{x}) \rangle$ and $\langle f^2(\mathbf{x}) \rangle$ are constant, and the autocorrelation function only depends on the difference of $\mathbf{x}$ and $\mathbf{y}$ $\langle f(\mathbf{x})f(\mathbf{y}) \rangle = R_{ff}(\mathbf{x}, \mathbf{y}) \equiv R_{ff}(\mathbf{x} - \mathbf{y})$. If the statistical properties of the field are also invariant with respect to rotations and reflections, we speak about a *homogeneous and isotropic random field*. In such a field the autocorrelation is a function of the magnitude of $\mathbf{x} - \mathbf{y}$: $\langle f(\mathbf{x})f(\mathbf{y}) \rangle = R_{ff}(|\mathbf{x} - \mathbf{y}|)$. The autocorrelation function $R(r) = a^2 \exp(-r/r_0)$, where $r =$

$\sqrt{[(x_1 - y_1)^2 + (x_2 - y_2)^2 + (z_1 - z_2)^2]}$, belongs to an isotropic field, and r_0 is the *correlation distance*, that is the value of r for which the autocorrelation function decreases to $1/e$ times its value at $r = 0$.

Most rock-physical applications of random geometry are related to the *Kozeny-Carman (KC) Equation* which expresses the permeability of a porous sedimentary rock as

$$k = \frac{1}{b} \Phi^3 \frac{1}{S_{\text{spec}}^2} \frac{1}{\tau^2} \quad (1)$$

where b is a shape factor (of the pore throats) of order one, $\Phi \in [0, 1]$ is porosity, S_{spec} is *specific surface*, the dimensionless parameter τ is called *hydraulic tortuosity*,

$$\tau = \frac{\langle L_{\text{hydr}} \rangle}{L} \geq 1$$

where $\langle L_{\text{hydr}} \rangle$ is the average hydraulic path length between two points a distance L . Combining $\frac{1}{b\tau^2}$ to a single constant C , the KC equation is expressed as

$$k = C \frac{\Phi^3}{S_{\text{spec}}^2}. \quad (2)$$

There are three concepts of specific surface: S_{spec} = surface area per unit bulk volume of the rock; S_0 = surface area per unit volume of solid material; S_p = surface area per unit volume of pore space. For any arrangement consisting of spherical grains of the same radius r one has

$$S_{\text{spec}} = (1 - \Phi)S_0; S_{\text{spec}} = \Phi S_p; S_p = \frac{1 - \Phi}{\Phi} S_0. \quad (3)$$

(Φ is porosity) and the KC equation can be written in three forms

$$k = C \frac{\Phi^3}{S_{\text{spec}}^2} \quad (4a)$$

$$k = C \frac{\Phi^3}{(1 - \Phi)^2 S_0^2} \quad (4b)$$

$$k = C \frac{\Phi}{S_p^2} \quad (4c)$$

where

$$C = \frac{1}{b\tau^2}. \quad (4d)$$

An important geometric theorem (Debye's theorem, see Korvin 2016) about specific surface area S_p (surface area per unit volume of pore space) states that *Select in a 2D image of an isotropic porous rock two randomly placed points A and B a distance r apart, where r is small, $r \ll 1$. Then the probability that A and B are in different media (i.e., one is in a pore, the other in a grain) is*

$$S_p \cdot \frac{r}{2} \quad (5)$$

ACF (Autocorrelation Function) Based Estimates of the Specific Surface

A plane section of the rock can be represented as a map of binary values

$$P(x,y) = \begin{cases} 1 & \text{if } (x,y) \in \Pi \\ 0 & \text{if } (x,y) \notin \Pi \end{cases}$$

where Π is the total set of pores in the microscopic image of the rock. $P(x,y)$ is the *characteristic function* of the pores. Assume that $P(x,y)$ is a translation invariant and isotropic random field. If Φ is the overall porosity, a randomly selected point (x,y) is with probability Φ in pore, with probability $1 - \Phi$ in the rock matrix. Taking expected values over different realizations of $P(x,y)$, the mean and variance of $P(x,y)$ are

$$\langle P(x,y) \rangle = \Phi \cdot 1 + (1 - \Phi) \cdot 0 = \Phi$$

$$\begin{aligned} \langle [P(x,y) - \Phi]^2 \rangle &= \langle P^2(x,y) - 2\Phi P(x,y) + \Phi^2 \rangle = \Phi - \Phi^2 \\ &= \Phi(1 - \Phi). \end{aligned}$$

In case of translation- and rotation invariance of $P(x,y)$ its normalized *Autocorrelation Function*

$\rho_{PP}(\xi, \eta)$ only depends on $r = \sqrt{\xi^2 + \eta^2}$:

$$\rho_{PP}(\xi, \eta) = \frac{\langle [P(x, y) - \Phi] \times [P(x + \xi, y + \eta) - \Phi] \rangle}{\langle [P(x, y) - \Phi]^2 \rangle} = \rho_{PP}(r).$$

Using Debye's theorem (Eq. 5), the specific surface of the pore boundaries can be estimated from the slope of the ACF $\rho_{PP}(r)$ at $r = 0$ as

$$S_p = -4 \frac{d}{dr} \rho_{PP}(r) \Big|_{r=0}. \quad (6)$$

Using the Kozeny-Carman Eq. (4) and assuming reasonable default values for shape factor b and tortuosity τ , Eq. (6) can be used to predict permeability from the ACF of the microscopic image. If the characteristic function $P(x, y)$ of the pore space has an exponential ACF with correlation distance r_0 , that is, if $\rho_{PP}(r) = \exp\left[-\frac{r}{r_0}\right]$, we get $S_{\text{spec}} = \frac{4}{\rho_0}$ and using this in the KC equation gives

$$k = \text{const} \cdot \Phi^3 r_0^2. \quad (7)$$

The exponential ACF $\rho(r) = \exp\left[-\frac{r}{r_0}\right]$ works well for ordinary black-and-white photographs, but for microscopic images of sedimentary rock a *stretched exponential* function

$$\rho(r) = \exp\left[-\left(\frac{r}{r_0}\right)^n\right] \quad (8)$$

has been found (Ioannidis et al. 1996) to better describe the measured ACF. Here r_0 is the correlation distance, and n a positive real exponent. The derivative of this ACF at $r = 0$ is $\frac{d}{dr} \rho(r) \Big|_{r=0} = -\frac{n}{r_0} r^{n-1} \Big|_{r=0}$ which, except for $n = 1$, cannot express the specific surface by means of Eq. (6) because for $n < 1$ it is divergent, while for $n > 1$ it is zero. To avoid this difficulty, one can use an *average correlation distance*

$$I_S = \int_0^\infty \rho(r) dr = \int_0^\infty \exp\left[-\left(\frac{r}{r_0}\right)^n\right] dr = \frac{r_0}{n} \Gamma\left(\frac{1}{n}\right) \quad (9)$$

and, instead of using (Eq. 7), which was valid for the ACF $R(r) = \exp\left[-\frac{r}{r_0}\right]$, expresses permeability as $k = \text{const} \cdot \Phi^B I_S^C$, that is,

$$\log k = A + B \cdot \log \Phi + C \cdot \log I_S \quad (10)$$

To find the value of $I_S = \frac{r_0}{n} \Gamma\left(\frac{1}{n}\right)$, one needs the correlation distance r_0 and the stretching exponent n , for which precise

estimates of Φ and of the ACF are needed. A practical problem that arises here is the *thresholding* of the images, that is, distinguishing pores from grains.

The Random Geometry of Tortuous Flow Paths

A simple scaling argument (Korvin 2016) shows that the hydraulic tortuosity in 2D sections of granular porous sedimentary rocks is an increasing function of grain size, while it decreases with porosity. Let L be the vertical size of the section considered (assuming that flow goes from top to bottom); Φ porosity (in fraction); τ tortuosity; r_0 , P_0 , A_0 characteristic size, characteristic perimeter, characteristic area of the grains in the 2D cross section; Z average number of pores adjacent to a grain (in the 2D section). We shall denote by $D_{P/A}$ the exponent in the celebrated Mandelbrot's *perimeter-area scaling law* (Mandelbrot 1982) applied to the grains in the microscopic 2D section:

$$P = P_0 \left(\frac{\sqrt{A}}{\sqrt{\pi} r_0} \right)^{D_{P/A}} \quad (11)$$

We proceed to show that *the average tortuosity of a flow path from top to bottom is given by*

$$\left\langle \frac{L_{\text{hydr}}}{L} \right\rangle = \tau = \Phi + \frac{(1 - \Phi)}{Z} \left(\frac{P_0}{r_0} \right) \left(\frac{\sqrt{A}}{r_0 \sqrt{\pi}} \right)^{D_{P/A}} \quad (12)$$

Note the special cases of Eq. (12). For $\Phi = 1$ we have $\tau = 1$; for $\Phi = 0$ there are no pores at all, the grain/pore coordination number is $Z = 0$ and consequently $\tau = \infty$ as it should be (see section “[Percolation Theory](#)”; Korvin 1992b).

Along a randomly selected vertical line of length L a part ΦL of the line goes through pore space. In these parts the flow goes along straight line segments. The remaining $(1 - \Phi)L$ length of the vertical line is filled by grains, and the fluid would cross $\frac{(1 - \Phi)L}{r_0}$ grains if it could go straight. But it cannot, because every time it reaches a grain it changes direction and continues in a “throat” following the curvature of the grain's periphery. By the definition of *coordination number* Z , the periphery P of a grain is adjacent to Z other grains, so that every individual “detour” adds a length $\left(\frac{P}{Z}\right)$ to the hydraulic path. This length of the detour is, by Mandelbrot's Eq. (11), equal to $\left(\frac{P}{Z}\right) = \frac{P_0}{Z} \left(\frac{\sqrt{A}}{r_0 \sqrt{\pi}} \right)^{D_{P/A}}$. As there are $\frac{(1 - \Phi)L}{r_0}$ such detours, the average tortuosity down to the bottom is indeed $\left\langle \frac{L_{\text{hydr}}}{L} \right\rangle = \tau = \Phi + \frac{(1 - \Phi)}{Z} \left(\frac{P_0}{r_0} \right) \left(\frac{\sqrt{A}}{r_0 \sqrt{\pi}} \right)^{D_{P/A}}$. Equation (12) compares fairly well with other theoretical predictions (Matyka et al. 2008) of the dependence of tortuosity on porosity. In a published *Lattice Gas (LG)* model $\tau = 0.8(1 - \Phi) + 1$; in a *percolation model* $\tau = 1 + a \frac{(1 - \Phi)}{(\Phi - \Phi_c)^m}$ (a and m are fitting parameters, Φ_c is the percolation threshold). The rule $\tau = \Phi^{-p}$ for Archie's *electric tortuosity*, or $\tau \approx 1 - p \log \Phi$ (p is

a fitting parameter; “log” will denote *natural logarithm* from now on) was obtained for cube-shaped grains and in theoretical studies on diffusive transport in porous systems composed of freely overlapping spheres, the empirical relation $\tau = 1 + p(1 - \Phi)$ was found for sandy or clay-silt sediments, and $\tau = [1 + p(1 - \Phi)]^2$ was reported for marine muds.

Reconstructing 3D Pore Space from 2D Sections

In this technique, statistical information, obtained from binary images of thin-sections, is used to create a 3D representation of the porous structure. The resulting 3D model honors the statistics of the thin sections, and is used to determine the topological attributes and effective physical properties of the rock. A powerful reconstruction algorithm of this kind employs an optimization technique imitating a thermodynamic process (*Simulated Annealing*, Yeong and Torquato 1998; see section “*Simulated Annealing*”).

The statistical properties needed are the first two moments of the characteristic function $\chi(\vec{r})$ (which is unity if $\vec{r}$ lies in the pore space, and zero otherwise): the porosity, $\Phi = \langle \chi(\vec{r}) \rangle$, and the normalized autocorrelation function, $\rho(\vec{u}) = \frac{\langle [\chi(\vec{r}) - \Phi][\chi(\vec{r} + \vec{u}) - \Phi] \rangle}{\Phi(1 - \Phi)}$. The ACF is fitted by the model $\rho(u) = \exp[-(u/r_0)^n]$, where $u = |\vec{u}|$, r_0 is a characteristic length scale (correlation length), and n is an adjustable parameter. In each iteration step i we compute the “energy” function $E_i(\rho) = \sum_k [\rho_i(u_k) - \rho_{ref}(u_i)]^2$ which is to be minimized, where $\{u_i\} = \{0, \Delta u, 2\Delta u, \dots\}$ is the digitized argument of the ACF, $\rho_i(u)$ is the actually observed ACF in this iteration step, $\rho_{ref}(u)$ is the ACF of the field to be reconstructed. Starting from a random configuration with porosity Φ , two voxels of different phases are exchanged at each iteration step. Thus Φ remains constant. The new configuration is accepted with the probability P given by the Metropolis rule

$$P = \begin{cases} 1 & \text{if } E_i \leq E_{i-1} \\ \exp[(E_{i-1} - E_i)/T] & \text{if } E_i > E_{i-1} \end{cases}$$

where T plays the role of temperature. In case of rejection, the old configuration is restored. By decreasing T , fields f_i with minimal energy $E_i(\rho)$ will be generated, that is their correlation properties will be similar to that of the reference field f_{ref} .

A 3D pore space can be transformed to a network of nodes (pore bodies) connected by bonds (pore channels, throats). In the most commonly used method one applies a 3D *morphological thinning algorithm* to extract the pores’ medial axes for the skeleton of the pore space. The network model is useful to determine the rock’s local and overall connectivity properties. Local connectivity is expressed by the pore’s

coordination number Z_c , defined as the number of throats connected to the given pore. The connectivity of a volume V of the rock is measured by its *specific genus* G_V , $G_V = G/V = (b - n + 1)/V$ where b and n are the number of branches and nodes in the pore network (see Korvin 1992b: 302–303). The parameter $G = b - n + 1$ is the maximal number of independent paths between two points in pore space, that is, the maximum number of independent cuts that can be made in the branches of the network without disconnecting it into separate networks. (Only pores with $Z \geq 3$ are considered “nodes” in the topological context.) For large networks, the genus per node G_0 is related to average coordination number by $G_0 = (\langle Z_{\geq 3} \rangle / 2) - 1$.

Maximum Entropy Methods

The Maximum Entropy Method

Suppose a measurable rock property λ can assume values belonging to L distinct ranges $\Lambda_1, \dots, \Lambda_L$. If we measure λ on a large number N of samples, we will find N_1 values in range $\Lambda_1, \dots, N_L$ values in range Λ_L . Letting $p_i = N_i/N$, the set of numbers

$$\{p_1, p_2, \dots, p_L\}, p_i \geq 0, \sum_{i=1}^L p_i = 1 \quad (13)$$

constitute a *discrete probability distribution*. It can represent different degrees of randomness: the distribution $p_1 = 1, p_2 = p_3 = \dots = p_L = 0$ is *not random*; the distribution $p_1 = \frac{1}{2}, p_2 = \frac{1}{2}, p_3 = \dots = p_L = 0$ is *not too random*, the distribution where all lithologies are equally possible, $p_1 = p_2 = p_3 = \dots = p_L = \frac{1}{L}$ is the *most random*. To characterize quantitatively the “randomness” of the distribution $\{p_1, p_2, \dots, p_L\}$, count how many ways one can classify the N samples such that $N_1 = p_1 N$ belongs to $\Lambda_1, N_2 = p_2 N$ to $\Lambda_2, \dots,$

$N_L = p_L N$ to Λ_L . The number of such classifications is given by

$$\Pi = \frac{N!}{N_1! N_2! \dots N_L!} \quad (14)$$

The larger is Π , the more random the distribution. Instead of Π it is easier to estimate $\log \Pi$ (“log” always means natural logarithm in this chapter),

$$\log \Pi = \log N! - \log N_1! - \dots - \log N_L! \quad (15)$$

If $n \gg 1$ we have the approximate Stirling’s formula

$$\begin{aligned}
\log(n!) &= \log(1 \cdot 2 \cdot 3 \cdots n) \\
&= \log 1 + \log 2 + \log 3 + \cdots + \log n \\
&\approx \int_1^n \log x dx = n \log n - n \sim n \log n
\end{aligned} \quad (16)$$

Using this approximation in Eq. (15):

$$\begin{aligned}
\log \Pi &\sim N \log N - \sum_{i=1}^L N_i \log N_i = - \sum_{i=1}^L N_i \log \frac{N_i}{N} \\
&= -N \sum_{i=1}^L \frac{N_i}{N} \log \frac{N_i}{N} \\
&= -N \sum_{i=1}^L p_i \log p_i = N \cdot S(p_1, p_2, \cdots, p_L) \quad (17)
\end{aligned}$$

where $S(p_1, p_2, \cdots, p_L) = - \sum_{i=1}^L p_i \log p_i$ is the *Shannon entropy* of the probability distribution $(p_1, p_2, \cdots, p_L)$.

In Rock Physics we frequently have to solve an *overdetermined* system of equations

$$\left. \begin{aligned} F_1(\xi_1, \xi_2, \cdots, \xi_L) &= y_{\text{measured}}^{(1)} \\ &\vdots \\ F_M(\xi_1, \xi_2, \cdots, \xi_L) &= y_{\text{measured}}^{(M)} \end{aligned} \right\} (L \gg M)$$

In the *Maximum Entropy (ME) Technique* we accept that particular solution of this system whose *Shannon entropy* is *maximal*.

Maximum Entropy (ME) Pore Shape Inversion

In an ME-based *rock-physical inversion* Doyen (1987) determined the *pore-shape distribution* in igneous rocks from hydraulic conductivity and/or dc electric conductivity measurements performed at a series of confining pressures. His rock model consisted of a random distribution of spherical pores connected with each other by tubes (“throats”). (This model is discussed in Korvin et al. 2014.)

At a given reference pressure $P = P_0$ the pores’ radii r are distributed according to the probability density function (*pdf*) $n(r)$; the throat-length l is distributed according to the *pdf* $n(l)$; the throats have elliptic cross-section with semi-axes b and c following the bivariate *pdf* $n(\alpha, c)$ where α is the aspect ratio $\alpha = b/c$. Using *Elasticity Theory* (Zimmerman 1991) to compute the deformation of voids under pressure P , the *pdf*’s for $r(P)$, $c(P)$, $l(P)$, $\alpha(P)$ can be determined for any pressure step from their values at reference pressure. The *pdf* $n(\alpha, c)$ satisfies the *self-consistency* equation (cf. Kirkpatrick 1973; see section “*Effective Medium (EM) Approximation*”) for all pressure steps P :

$$\sum_{\alpha, c} n(\alpha, c) \frac{g_e^*(P) - g_e(P, \alpha, c, l)}{g_e(P, \alpha, c, l) + \left(\frac{Z}{2} - 1\right) g_e^*(P)} \quad (18a)$$

$$\sum_{\alpha, c} n(\alpha, c) \frac{g_h^*(P) - g_h(P, \alpha, c, l)}{g_h(P, \alpha, c, l) + \left(\frac{Z}{2} - 1\right) g_h^*(P)} = 0 \quad (18b)$$

where Z is the average pore coordination number (number of throats incident to a pore); $g_e^*(P)$ and $g_h^*(P)$ are the measured bulk electric and hydraulic conductivities at pressure P , and $g_e = g_e(P, \alpha, c, l)$ and $g_h = g_h(P, \alpha, c, l)$ are the theoretical conductivities of an originally (α, c) -shaped throat of length l subjected to pressure P .

Elasticity Theory also provides a theoretical value $\beta_p(r, \Phi)$ for the compressibility of a rock at pressure P if it contains a volume-fraction Φ of r -sized pores; and an expression for its expected value $\beta_p(\{n(r)\}, \Phi)$. Similarly, for a volume fraction Φ of throats of length l and cross-sectional shape (α, c) both $\beta_p(\alpha, c, l, \Phi)$ and its expected value can be theoretically determined.

Subjecting the rock to pressure P , the pore parameters change as $r \rightarrow r(P)$, $l \rightarrow l(P)$, $c \rightarrow c(P)$, $\alpha \rightarrow \alpha(P)$, the measured bulk compressibility becomes $\beta(P)$. Assuming that the compressibilities due to pores and throats are independent and additive, one gets

$$\beta_p(\{n(r)\}, \Phi_p, P) + \beta_t(\{n(\alpha, c)\}, \Phi_t, l, P) = \beta(P) \quad (19)$$

Let $V(r)$ be the volume of a pore of radius r , and $V(\alpha, c, l)$ the volume of an (α, c) -shaped tube of length l . Then $n(r)$ and $n(\alpha, c)$ satisfy the additional constraints:

$$\begin{aligned}
\Sigma_r n(r) &= 1, \Sigma_{\alpha, c} n(\alpha, c) \\
&= 1, N_p \Sigma_r n(r) V(r) + N_t \Sigma_{\alpha, c} n(\alpha, c) V(\alpha, c, l) \\
&= \Phi_p + \Phi_t = \Phi
\end{aligned}$$

where N_p and $N_t = N_p \frac{Z}{2}$ are the numbers of pores, respectively throats, in a unit volume of rock. By simple geometry, $N_p, N_t, \Phi_p, \Phi_t, Z, l$ can be expressed in terms of $n(r), n(\alpha, c), \Phi$. As there are much more possible values of $n(r)$ and $n(\alpha, c)$ than pressure steps P , we assume that $n(r)$ and $n(\alpha, c)$ are *Maximum Entropy* solutions, that is,

$$\begin{aligned}
-\Sigma_r n(r) \log n(r) &= \max; \quad -\Sigma_r n(\alpha, c) \log n(\alpha, c) = \max \\
-\Sigma_r n(r) \log n(r) &= \max; \quad -\sum_{\alpha, c} n(\alpha, c) \log n(\alpha, c) = \max
\end{aligned} \quad (20)$$

subject to the conditions

$$\sum_{\alpha, c} n(\alpha, c) \frac{g_e^*(P) - g_e(P, \alpha, c, l)}{g_e(P, \alpha, c, l) + \left(\frac{Z}{2} - 1\right) g_e^*(P)} = 0 \quad (20a)$$

$$\sum_{\alpha, c} n(\alpha, c) \frac{g_h^*(P) - g_h(P, \alpha, c, l)}{g_h(P, \alpha, c, l) + \left(\frac{Z}{2} - 1\right) g_h^*(P)} \quad (20b)$$

$$\beta_p(\{n(r)\}, \Phi_p, P) + \beta_l(\{n(\alpha, c)\}, \Phi_l, l, P) = \beta(P) \quad (20c)$$

$$\sum_r n(r) = 1, \sum_{\alpha, c} n(\alpha, c) = 1,$$

$$N_p \sum_r n(r) V(r) + N_l \sum_{\alpha, c} n(\alpha, c) V(\alpha, c, l) = \Phi_p + \Phi_l = \Phi \quad (20d)$$

The constraining Equations (20a–d) have the common form

$$\sum_{r, \alpha, c} n(r, \alpha, c) A(r, \alpha, c, P) = 0 \text{ for } P = P_1, P_2, \dots, P_N \quad (21)$$

where (r, α, c) , satisfying $n(r, \alpha, c) \geq 0$; $\sum_{r, \alpha, c} n(r, \alpha, c) = 1$ is the joint *pdf* of $\{r, \alpha, c\}$ at the reference pressure. Its *Shannon entropy* is

$$S[n(r, \alpha, c)] = - \sum_{r, \alpha, c} n(r, \alpha, c) \log n(r, \alpha, c) \quad (22)$$

Using the Lagrange multiplier technique, we first find the vector which minimizes the function

$$f(\lambda_{P_1}, \dots, \lambda_{P_N}) = \log \left[\sum_{r, \alpha, c} \exp \left\{ - \sum_{k=1}^N \lambda_{P_k} A(r, \alpha, c, P_k) \right\} \right] \quad (23)$$

and then find the ME solution (called “pore spectra”) as

$$n(r, \alpha, c) = \frac{\exp \left\{ - \sum_{k=1}^N \lambda_{P_k} A(r, \alpha, c, P_k) \right\}}{\sum_{r, \alpha, c} \exp \left\{ - \sum_{k=1}^N \lambda_{P_k} A(r, \alpha, c, P_k) \right\}} \quad (24)$$

Shale Compaction and Maximal Entropy

By Athy’s law (Athy 1930) in thick pure shale porosity decreases with depth as

$$\Phi(z) = \Phi_0 \exp(-kz) \quad (25)$$

where $\Phi(z)$ is porosity at depth z , Φ_0 porosity at the surface, and k a constant. Assuming all pores have the same volume, the porosity of a rock is proportional to the number of pores in

a unit volume of the rock. Athy’s rule states in this case that the pores in compacted shales are distributed in such a manner that their number in a unit volume of rock exponentially decreases with depth. There are several analogies of this rule in Statistical Physics. The most familiar is the *barometric equation* of Boltzmann for the density $\rho(z)$ of the air at altitude z as $n\rho(z) = \rho(0) \cdot \exp\left[-\frac{mgz}{kT}\right]$, where m is the mass of a single gas molecule, g gravity acceleration, k Boltzmann’s constant, and T absolute temperature. In Statistical Physics (Landau and Lifshitz 1980: 106–114) Boltzmann’s equation is derived from the assumptions that the gas particles move independently of each other and the system tends toward its most probable (*maximum entropy*) state.

During compaction of shale, water is expelled and clay particles rearrange themselves toward a more dense system of packing. Korvin (1984) adopted Litwiniszyn’s model (1974) and considered shale compaction history as an *upward migration of pores*. Take a rectangular prism P of the present-day shale of unit cross-section reaching down to the basement at depth Z_0 , and suppose its *mean porosity* is Φ , that is, it contains a fractional volume ΦZ_0 of fluid and a volume $(1 - \Phi)Z_0$ of solid clay particles. Assume that the compaction process is *ergodic*, that is, it tends toward the *maximum-entropy final state*. Neglecting the actual depositional history we assume for time $t = 0$ an initial condition where a prism of water of unit cross-section, height ΦZ_0 and density ρ_1 had been overlain by solid clay of height $(1 - \Phi)Z_0$ and density ρ_2 , $\rho_1 < \rho_2$. Divide the prism of water into $\mathfrak{N}$ “particles” (water-filled pores), each of volume ΔV , which at $t = 0$ started to migrate upwards independently of each other, until the final (*maximum entropy*) state had been reached. The initial potential energy of the system had been

$$E = \mathfrak{N} \cdot V \cdot g(\rho_2 - \rho_1)(1 - \Phi)Z_0, \quad (26)$$

where initial porosity and particle number are connected by $\mathfrak{N} = \frac{Z_0 \Phi}{\Delta V}$. Divide the prism P into N equal slabs of thickness $\Delta z = Z_0/\Delta z$, denote the i th slab by γ_i ($i = 0, 1, \dots, N - 1$), and divide the prism P into $N^* = Z_0/\Delta V$ nonoverlapping small cubes. We have $\mathfrak{N} \ll N^*$ if Φ is sufficiently small. We rank the N^* possible positions, called “states,” of a pore into N groups: a pore is said to belong to the group γ_i if and only if its center (x, y, z) lies within the slab γ_i . This implies every group γ_i contains $G = \Delta z/\Delta V$ states. Suppose that N_i pores are found in state γ_i . The numbers N_i satisfy two constraints, the *conservation of pore-particle number*, and the *conservation of total potential energy*:

$$\sum_{i=0}^{N-1} N_i = \mathfrak{N} \quad (27a)$$

$$\sum_{i=0}^{N-1} \varepsilon_i N_i = E \quad (27b)$$

where E is the total energy (see Eq. 26); ε_i is the potential energy of a single pore particle in group γ_i , due to buoyancy:

$$\varepsilon_i = g(\rho_2 - \rho_1) \cdot \Delta V \cdot i \cdot \Delta z \quad (28)$$

The set of numbers $\{N_i\}$ determine the *macroscopic* distribution of pores inside the prism P . Apart from a constant factor, the entropy of the distribution is

$$S = \sum_{i=1}^{N-1} N_i \log \frac{eG}{N_i} \quad (29)$$

Denote the average number of pore particles in group γ_i by $\bar{n}_i$, then $\bar{n}_i = N_i/G$,

$$S = G \sum_{i=1}^{N-1} \bar{n}_i \log \frac{e}{\bar{n}_i} \quad (30)$$

and the constraints (27a, b) become

$$G \cdot \sum_{i=0}^{N-1} \bar{n}_i = \mathfrak{N}, \quad G \cdot \sum_{i=0}^{N-1} \bar{n}_i \varepsilon_i = E \quad (31)$$

The pore particles will migrate to a position such that the entropy (Eq. 29) be maximal. To maximize the entropy subject to the constraints (31), we introduce Lagrange multipliers α, β assume that

$$\frac{\partial}{\partial \bar{n}_i} (S + \alpha \frac{\mathfrak{N}}{G} + \beta \frac{E}{G}) = 0 \quad (i = 0, 1, \dots, N-1), \text{ that is}$$

$$\bar{n}_i = \exp(\alpha + \beta \varepsilon_i), \text{ wherefrom } \exp \alpha = \frac{\Phi}{1-\phi}, \beta = -\frac{1}{E} \text{ and}$$

$$\Phi(z) = \frac{\Phi}{1-\phi} \cdot \exp \left[-\frac{z}{(1-\phi)Z_0} \right] \quad (32)$$

Identifying the first factor in Eq. (32) with surface porosity Φ_0 , the equation becomes

$$\Phi(z) = \Phi_0 \cdot \exp \left[-\frac{(1+\Phi_0)z}{Z_0} \right] \quad (33)$$

an equation that reproduces Athy's compaction law.

Contact Force Distribution in Granular Media

K. Bagi (2003) determined the probability density functions of contact force components in the principal directions of average stress in random granular assemblies by maximizing the statistical entropy, and proved that the distribution of force magnitudes is exponential. She considered a body consisting of randomly packed, cohesionless grains. If external stresses

are acting on this body's boundary, contact forces arise between neighboring grains and the average stress of the assembly, $\bar{\sigma}_{ij}$, depends on the contact forces f_i^c between neighbor grains and on the vectors l_i^c pointing from the center of the first grain to the center of the second,

$$\bar{\sigma}_{ij} = \frac{1}{V} \sum_{c=1}^M l_i^c f_j^c \quad (34)$$

(M is the total number of contacts; the index “ c ” runs over all contacts; $i, j = 1, 2, 3$ indicate vector components; V is the total volume). If the coordinate axes (x_1, x_2, x_3) are parallel to the principal directions of $\bar{\sigma}_{ij}$, then $\bar{\sigma}_{ij} = 0$ if $i \neq j$. Consider Eq. (34) first for $i = j = 1$:

$$\bar{\sigma}_{11} = \frac{1}{V} \sum_{c=1}^M l_1^c f_1^c. \quad (35)$$

Since when two grains are in contact, any of them can be “first” or “second,” we can assume that $0 \leq f_1^c \leq f_{\max}$. We discretize this domain into equal (small) intervals: a particular f_1^c can fall into the k th interval where the medium value is F_{1k} with probability $p_k^{(1)}$, $k = 1, 2, \dots$, or into the second interval where the medium value is F_{12} , etc. The distribution of the f_1 -component is described by $p_1^{(1)}, p_2^{(1)}, \dots, p_k^{(1)}, \dots$; representing the probabilities that f_1^c falls into the 1st, 2nd, ... k th, ... interval. Next consider all those grain contacts where the f_1^c contact force component falls into a chosen k th interval (that is $f_1^c \approx F_{1k}$). Take the average of the l_1^c grain-center-to-grain-center vector components for these contacts, and denote it as $\bar{l}_{1k}^{(1)}$ where the superscript “(1)” expresses that the grouping was done according to the f_1 -components.

With these notations, Eq. (35) is averaged to

$$\bar{\sigma}_{11} = \frac{1}{V} \sum_{k=1}^{\infty} (M \cdot p_k^{(1)}) \bar{l}_{1k}^{(1)} F_{1k}. \quad (36)$$

and we similarly get all nine equations

$$\bar{\sigma}_{ij} = \frac{1}{V} \sum_{k=1}^{\infty} (M \cdot p_k^{(j)}) \bar{l}_{ik}^{(j)} F_{jk} \quad (37)$$

Suppose the geometrical data and average stress are given and we are searching for the *distribution* of the i th contact force component $\{f_i\}$. The unknown $\{p_1, p_2, \dots, p_k, \dots\}$ values satisfy

$$\sum_{k=1}^{\infty} p_k = 1 \quad (38)$$

and (for a fixed $i = j$)

$$\bar{\sigma}_{ij} = \frac{M}{V} \sum_{k=1}^{\infty} p_k \bar{l}_{ik}^{(j)} F_{jk}. \quad (39)$$

We need a geometric assumption on the $\bar{l}_{ik}^{(j)}$ values. Suppose that they are constant:

$$\bar{l}_{i1}^{(j)} = \bar{l}_{i2}^{(j)} = \bar{l}_{i3}^{(j)} = \dots = \bar{l}_i^{(j)} \quad (40)$$

The average stress then can be expressed (for fixed $i = j$) as

$$\bar{\sigma}_{ij} = \frac{M}{V} \bar{l}_i \sum_{k=1}^{\infty} p_k F_{jk}. \quad (41)$$

We are looking for that distribution which has the highest entropy:

$$-\sum_{k=1}^{\infty} p_k \log p_k = \max, \quad (42)$$

subject to the conditions

$$\sum_{k=1}^{\infty} p_k = 1 \quad (43)$$

$$\bar{\sigma}_{ij} = \frac{M}{V} \sum_{k=1}^{\infty} p_k \bar{l}_{ik}^{(j)} F_{jk} \text{ (for a fixed } i = j). \quad (44)$$

Introduce the Lagrange-multipliers α and β associated with the constraints (Eqs. 43 and 44). Easy algebra yields:

$$p_k = \frac{1}{\exp(1 + \alpha)} \cdot \exp(-\beta \bar{l}_i^{(j)} F_{jk}) \quad (45)$$

where α and β can be (numerically) determined from the system of nonlinear equations (using the given data $\bar{\sigma}_{ii}$, $\bar{l}_i$, V and M):

$$\sum_{k=1}^{\infty} F_{jk} \exp(-1 - \beta \bar{l}_i^{(j)} F_{jk} - \alpha) - \frac{V}{M \bar{l}_i^{(j)}} \bar{\sigma}_{ij} = 0 \text{ (} i = j) \quad (46)$$

$$\sum_{k=1}^{\infty} \exp(-1 - \beta \bar{l}_i^{(j)} F_{jk} - \alpha) - 1 = 0 \quad (47)$$

Equation (45) means that the *distribution of contact force components is exponential*, provided that Eq. (40) is an acceptable simplification.

Self-Consistency and Effective Media

Effective Medium (EM) Approximation

A good introduction (Kirkpatrick 1973) to classical *Effective Medium Theory* is the study of random resistors on the bounds

of a lattice. Apply an external field along one axis so that the voltages increase by a constant amount per row of nodes. The average effect of the random resistors can be described by an *effective medium*, that is, a hypothetical homogeneous medium in which the total field inside is equal to the external field. Consider an effective field which is made up of equal conductances, g_m , connecting nearest neighbors on a cubic mesh. To find g_m , we require that the local perturbations in this medium due to the voltages, induced when individual conductances g_{ij} replace g_m , should average to zero. Consider a conductance g_{AB} between randomly picked neighboring nodes A , B in the direction of the external field. Using Kirchhoff's laws it can be shown that if we change g_m to g_{AB} , the arising voltage, V_o , induced between A and B will be

$V_o = V_m(g_m - g_{AB}) / [g_{AB} + (\frac{Z}{2} - 1)g_m]$, where V_m is the constant increase in voltages per row, and Z is the coordination number.

If the conductivities g_{ij} are distributed according to the *pdf* $f(g)$, the requirement that the average of V_o should vanish yields an equation for g_m :

$\int dg f(g)(g_m - g) / [g + (\frac{Z}{2} - 1)g_m]$. (A special case of this equation was applied in section “[Maximum Entropy \(ME\) Pore Shape Inversion](#)”, as Eq. 18a, 18b.)

Another simple case of the *EM* approximation (Stroud 1998) concerns the random mixture of two types of grains, denoted by 1 and 2, which are present in relative volume fractions p and $1 - p$, and characterized by conductivities σ_1 and σ_2 . To calculate the effective conductivity σ_e of this composite, imagine that each grain, instead of being embedded in its actual random environment, lies in an homogeneous *effective medium*, whose conductivity σ_e will be determined *self-consistently*. If the grains are spherical, the electric field inside them is given in Electrostatics as $E_i = E_0 \frac{3\sigma_e}{\sigma_i + 2\sigma_e}$ ($i = 1$ or $i = 2$), where E_0 is the electric field far from the grain. *Self-consistency* requires that the *average electric field* within the grain should equal E_0 , that is

$p E_0 \frac{3\sigma_e}{\sigma_1 + 2\sigma_e} + (1 - p) E_0 \frac{3\sigma_e}{\sigma_2 + 2\sigma_e} = E_0$ leading to the quadratic equation

$p \frac{\sigma_1 - \sigma_e}{\sigma_1 + 2\sigma_e} + (1 - p) \frac{\sigma_2 - \sigma_e}{\sigma_2 + 2\sigma_e} = 0$. The equation has two solutions for σ_e , one physically sensible, the other unphysical.

The *EM* approximation is also used in Percolation Theory (see section “[Percolation Theory](#)”) where transport properties (such as conductivity or diffusivity) grow as a power-function above, but near to, the percolation threshold, p_c . For $p > p_c$ the electrical conductivity σ of a random network grows as $\sigma \propto (p - p_c)^t$, $p > p_c$ where p is the occupation probability for a site or bond, and the scaling exponent t is equal to 2 in 3D. In the *Effective Medium (EM)* approach (Ghanbarian et al. 2014) a randomly selected part of the heterogeneous medium is replaced by the homogeneous material. Replacing the actual transport coefficient (e.g., conductivity) by an *effective* one perturbs the local current. The effective conductivity is determined by requiring that the average of these

perturbations be zero. For the lattice percolation model of electric conductivity in rocks this gives (for two components, see Kirkpatrick 1973, Stroud 1998) $\sigma(p)/\sigma_b = 1 - \frac{1-p}{1-Z}$ where σ_b is the electrical conductivity of the brine, grains are nonconducting, and Z is the coordination number. In n -dimensional lattices the coordination number Z and p_c are approximately related as $Zp_c = d/(d-1)$ so in 2D one has $2/Z \approx p_c$, and $\frac{\sigma(p)}{\sigma_b} \approx \frac{p-p_c}{1-p_c}$.

Differential Effective Medium (DEM) Approximation

The elastic properties of porous or cracked rocks are commonly modeled by the *DEM* (Differential Effective Medium) approach (Mavko et al. 1998; Almquist et al. 2011; Haidar et al. 2018). In *DEM* a composite material is constructed step-wise by making infinitesimal changes to an already existing composite. For example, a two-phase porous or cracked rock can be built up by incrementally adding small fractions of pores or cracks to the rock matrix. As we insert new inclusions into the background model, it will continuously change as:

$$\begin{aligned} (1 - \Phi)[K^*(\Phi)] &= (K_2 - K^*)P^{(*)2}(\Phi) \\ (1 - \Phi)[\mu^*(\Phi)] &= (\mu_2 - \mu^*)Q^{(*)2}(\Phi) \end{aligned}$$

(see Mavko et al. 1998) where Φ is porosity; $K^*(\Phi)$ is the effective bulk moduli of DEM; K^* is bulk modulus of the matrix (phase 1); K_2 bulk modulus of the inclusion (phase 2); $P^{(*)2}$ is geometry factor for an inclusion of material 2 in a background medium with effective moduli K^* and μ^* ; $\mu^*(\Phi)$ is the effective shear moduli of DEM; μ^* shear modulus of the matrix (phase 1); μ_2 shear modulus of the inclusion (phase 2); $Q^{(*)2}$ geometry factor for an inclusion of material 2 in a background medium with effective moduli K^* and μ^* . The DEM equations (for K and μ) are coupled, as both depend on both the bulk and shear moduli of the composite through the geometry factors. The equations are numerically integrated starting from porosity $\Phi = 0$ with initial values $K^*(0) = K_m$ and $\mu^*(0) = \mu_m$ which are the mineral values for the single homogeneous solid constituent. Integration then proceeds from $y = 0$ to the desired highest value $y = \Phi_{\max}$ as the inclusions are slowly introduced into the solid. The method needs a background rock matrix, the geometry factors, the elastic moduli of the inclusions, and the fraction of inclusions as input. The value of the aspect ratios are also needed in the geometry factors. There are three ranges of aspect ratio values, for the interparticle pores, the stiff pores, and the cracks. There is a control step where one calculates the reference velocity V_p^{ref} using the DEM equations. (It depends on the aspect ratio of the interparticle pores, the fraction of inclusions, the elastic moduli of the rock matrix and of the inclusions.) The reference value V_p^{ref} serves to decide whether stiff pores or cracks are to be added to the rock matrix. If

$V_p^{\text{ref}} < V_p^{\text{meas}}$, then stiff pores are added, and if $V_p^{\text{ref}} > V_p^{\text{meas}}$, then cracks are added to the matrix. When the process is complete, the effective elastic moduli can be calculated.

In an interesting application (Almquist et al. 2011) of *DEM* to compute the elastic constants for calcite and muscovite mixtures the background medium was *anisotropic*. For anisotropic elastic media *DEM* is computed as

$$\frac{dC^{\text{DEM}}}{dV} = \frac{1}{1-V} (C_i - C^{\text{DEM}}) A_i$$

where C^{DEM} is the elastic tensor of the effective media, V is the volume fraction of the inclusions, C_i is the elastic tensor of the inclusion, and A_i , which relates the strain inside the inclusion with that of the background matrix, is defined as

$$A_i = [I + G(C_i - C^{\text{DEM}})]^{-1}$$

Here I is the fourth-rank unit tensor, G is a symmetric tensor Green's function, C_i is the elastic modulus of the inclusion, the tensor C^{DEM} is updated in each inclusion step. Inclusions are ellipsoids with given aspect ratio ($a \geq b \geq c$).

Thermodynamic Algorithms

Random Walk

Gaussian random walk is a mathematical model of the *Brownian motion* of small particles in liquids. For this motion the change of the position $X(t)$ of the randomly moving particle $\Delta X(t) = X(t + \Delta t) - X(t)$ obeys a Gaussian distribution with mean $\langle \Delta X(t) \rangle = 0$ and variance $\langle [\Delta X(t)]^2 \rangle^{1/2} \propto \Delta t$. Two characteristic rock-physical applications of random walk modeling are a study of nuclear magnetic relaxation (NMR) in brine-filled porous rocks (Jin et al. 2009), and random-walk-based determination of the electric conductivity in rocks (Ioannidis et al. 1997).

The *random-walk method* of simulating the *Nuclear Magnetic Resonance* (NMR) response of fluids in porous rocks can be applied on *grain-based* and *voxel-based representations* of the rock. In the grain-based approach a spherical grain pack is the input, where the solid surface is analytically defined; in the voxel-based approach the input is a computer-generated 3D image of the reconstructed porous media. The implementation of random walk is the same in both cases, but it has been found that spin magnetization decays much faster in the digitized models than in case of analytically given surfaces because of overestimating the irregular pore surface areas.

For fluids in porous media, total magnetization originates from diffusion and relaxation processes. Three processes are involved in the relaxation of spin magnetization (only

transverse relaxation is considered): (1) bulk fluid relaxation, with characteristic time T_{2B} due to dipole–dipole interactions between spins within the fluid; (2) surface relaxation, characterized by effective relaxation time T_{2S} , due to the additional interaction of protons at the pore-grain interface with paramagnetic impurities in the grains, and the motion of water molecules in a layer near the pore-grain interface; and (3) relaxation due to background magnetic field heterogeneities, T_{2D} . The effect of the pore surface on the relaxation time T_{2S} is proportional to specific surface area: $\frac{1}{T_{2S}} = \rho \frac{S}{V}$ where ρ is surface relaxivity, D_B diffusion coefficient of the bulk fluid, S is pore surface area, and V pore volume. The decay rate T_2 of the magnetization is given by $\frac{1}{T_2} = \frac{1}{T_{2B}} + \frac{1}{T_{2S}}$. Far away from a boundary, T_2 is equal to the bulk fluid transversal relaxation time, T_{2B} . For spins close to a pore boundary, the magnetization decay is locally enhanced by surface relaxation. Most random-walk simulations of NMR assume that the spin is “killed” with a probability p when it reaches a boundary, and relate this probability to surface relaxivity ρ as $p = 0.96A \frac{\rho \Delta r}{D_B}$ where Δr is the displacement the spin achieves during the interval $[t, t + \Delta t]$ when it reaches the boundary, and A is an empirical factor of order 1 (usually set to 2/3). In the discussed more realistic model (Jin et al. 2009) of local surface relaxation the spin is not “killed,” but its magnetization decreases by a factor $(1 - p) \approx \exp[-\Delta t(p/\Delta t)]$. Therefore, the total relaxation rate becomes $1/T_2 = 1/T_{2B} + p/\Delta t$. Equivalently, the relaxation time used in each step of the random walk can be written

$$\frac{1}{T_2} = \frac{1}{T_{2B}} + \chi \frac{3.84\rho}{\Delta r} \quad (48)$$

where in the grain-based model, $\chi = 1$ when the walker is located within a step of the relaxing boundary, and 0 otherwise.

The starting points of walkers are randomly picked in the pore-space occupied by fluids (water and/or oil). Each walker is allowed to walk for a sufficiently long time, whenever it attempts to step out of the pore space into the matrix, its magnetization decays with a rate determined by Eq. (48). At each time-step Δt , a random vector $\vec{n} = (n_x, n_y, n_z)$ is generated, whose independent components are normally distributed with unit variance and zero mean. The walker is then spatially displaced by $\Delta \vec{r} = \sqrt{(6D_B \Delta t)} \frac{\vec{n}}{\|\vec{n}\|}$. The time step Δt is dynamically changed along the walk so that it is smaller than any magnetic-field pulse duration, and the amplitude of the total displacement, $\Delta \vec{r}$, is much shorter than the relevant geometric length scales (e.g., pore throats, wetting film thickness, etc.). It was assumed that the pores are fully brine-saturated, the pulse sequence was of the Carr–Purcell–Meiboom–Gill (CPMG) type with an inter-echo spacing of 1 ms, the parameters were $T_{2B} = 2800\text{ms}$, $D_B = 2\mu\text{m}^2/\text{ms}$,

$\rho = 0.005$. One analytically treatable case, a cubic packing of spheres, a random packing of monodisperse and polydisperse spheres, and a digitized micro X-ray-CT image were used as test cases. The theoretical case was a single spherical pore of radius R where the magnetization decay in the fast diffusion limit ($\rho R/D_B < 1$) is given by $\frac{M(t)}{M_0} = \exp\left[-t\left(\frac{1}{T_{2B}} + \rho \frac{S}{V}\right)\right]$, M_0 being the initial transverse magnetization (equal to 1 in this study). Simulations were performed for the case $\frac{\rho R}{D_B} = 0.02$. Grain-pack based simulations were made for the cubic packing of spheres, and for random packing of monodisperse and polydisperse spheres. The voxel-based approach was run on the micro X-ray-CT image of a Fontainebleau sandstone sample with 21.5% porosity, and the image consisted of $512 \times 512 \times 512$ voxels, with $5.68 \mu\text{m}$ voxel size.

In another random-walk model (Ioannidis et al. 1997), in an attempt to find the formation factor F in Archie’s equation, instead of solving Laplace’s equation these authors exploited the formal analogy between random walks and the solutions of the Laplace equation (Kim and Torquato 1990). They calculated the mean square displacement a function of time for a large number of random walkers. A single walker starts at a random void voxel (i_0, j_0, k_0) . It can take random steps to any neighboring void, without penetrating solid grains. The time required for each step is, by Einstein’s rule $t_{v-v} = \frac{(\partial D)^2}{\sigma_0}$, where σ_0 is conductivity of the pore fluid (set to 1), ∂D is distance traveled in one step in a 3D lattice of spacing δ , so that it can be δ , $\delta\sqrt{2}$ or $\delta\sqrt{3}$. To avoid premature disappearance of walkers, periodic boundary conditions are assumed. The random walker travels a large number of steps ($\sim 10^6$) and there are many ($\sim 10^3$) independent walkers. The average displacement after time t will be $\langle R^2 \rangle = \left(\frac{\sigma}{\Phi_a}\right)t$ where $\sigma (= \sigma_0/F)$ is the effective electrical conductivity, and Φ_a is the fraction of voxels accessible for the walker. For such rocks (carbonates, shales) where solid-void, void-solid, and solid-solid transitions are also allowed, Einstein’s formula changes to $t_{s-s} = \frac{(\partial D)^2}{\sigma_s} = \frac{(\partial D)^2}{\sigma_0} F_s$ (F_s is the formation factor of the matrix). Steps between solid and void are only allowed with a given probability. In the model they used $p_{v-s} = \frac{1}{1+F_s}$ and $p_{s-v} = \frac{F_s}{1+F_s}$. The time to cross an interface is $\tau = \left(\frac{2}{1+F_s}\right) \cdot \frac{(\partial D)^2}{4\sigma_0} F_s$, and $t_{s-v} = \frac{t_{s-s}}{4} + \tau$, $t_{v-s} = \frac{t_{v-v}}{4} + \tau$.

Lattice Boltzmann Methods for Flow in Porous Media

The *Lattice Boltzmann Method (LBM)* is a numerical tool for simulating complex fluid flows or other transport processes (Ghassemi and Pak 2010), a *finite-difference method* to solve the *Boltzmann transport equation* on a lattice. It is a simplified *molecular dynamics model* where space, time, and particle-velocities assume discrete values, each lattice node is connected to its neighbors, there can be either 0 or 1 particles at a lattice node, particles move with a given velocity. In

discrete time instants, each particle moves to a neighboring node, and this process is called *streaming*. When more than one particle arrive at the same node from different directions, they *collide* and change their velocities according to *collision rules* conserving particle number (mass), momentum, and energy before and after the collision.

The LBM originated from Boltzmann's kinetic theory of gases, where the gas consists of a large number of particles in random motion. Transport in this model is described by the equation $\frac{\partial f}{\partial t} + \vec{u} \cdot \nabla f = \Omega$ where $f(\vec{x}, t)$ is the particle distribution function, $\vec{u}$ is particle velocity, Ω the collision operator. As compared to Boltzmann's gas dynamics the number of particles is much reduced, and they are confined to the nodes of a lattice. The different numerical realizations of LBM are classified by the $D_n Q_m$ scheme, where “ D_n ” stands for “ n dimensions,” “ Q_m ” stands for “ m speeds.” Here we only discuss $D_2 Q_9$, the 2D model where a particle moves (“streams”) in nine possible directions, including the possibility to stay at rest. These velocities are referred to as *microscopic velocities* and are denoted by e_i , $i = 0, \dots, 8$. In the $D_2 Q_9$ model the nine velocity vectors are:

i	0	1	2	3	4	5	6	7	8
$\vec{e}_i$	(0,0)	(1,0)	(0,1)	(-1,0)	(0,-1)	(1,1)	(-1,1)	(-1,-1)	(1,-1)

Each particle follows a discrete distribution of probabilities $f_i(\vec{x}, t)$, $i = 0 \dots 8$, of streaming in any one particular direction. The *macroscopic fluid density* is the sum of microscopic particle distribution functions, $\rho(\vec{x}, t) = \sum_{i=0}^8 f_i(\vec{x}, t)$, the *macroscopic velocity* $\vec{u}(\vec{x}, t)$ is the average of microscopic velocities $\vec{e}_i$ weighted with the distribution functions f_i , $\vec{u}(\vec{x}, t) = \frac{1}{\rho} \sum_{i=0}^8 c f_i \vec{e}_i$ where $c = \Delta x / \Delta t$ is the lattice speed.

The key steps in LBM are *streaming* and *collision*, which are given by

$$\begin{cases} f_i(\vec{x} + c \vec{e}_i \Delta t, t + \Delta t) - f_i(\vec{x}, t) & \text{(streaming)} \\ -\frac{[f_i(\vec{x}, t) - f_i^{eq}(\vec{x}, t)]}{\tau} & \text{(collision)} \end{cases}$$

where $f_i^{eq}(\vec{x}, t)$ is the *local equilibrium distribution*, and τ the relaxation time toward local equilibrium. For single-phase flows, it suffices to use the simplified Bhatnagar-Gross-Krook collision rules (Bhatnagar et al. 1954) where equilibrium distribution is defined as $f_i^{eq}(\vec{x}, t) = w_i \rho + \rho s_i(\vec{u}(\vec{x}, t))$

with $s_i(\vec{u}) = w_i \left[3 \frac{\vec{e}_i \cdot \vec{u}}{c} + \frac{9}{2} \frac{(\vec{e}_i \cdot \vec{u})^2}{c^2} - \frac{3}{2} \frac{\vec{u} \cdot \vec{u}}{c^2} \right]$ and the weights w_i are 4/9 for $i = 0$; 1/9 for $i = 1, 2, 3, 4$; and 1/36 for $i = 5, 6, 7, 8$. The fluid's kinematic viscosity in the $D_2 Q_9$

model is related to relaxation time by $\nu = \frac{2\tau-1}{6} \cdot \frac{(c\Delta x)^2}{\Delta t}$. Parameters are selected such that the lattice speed c be much larger than the maximum fluid velocity v_{max} expected to arise in the simulation. This is expressed by the “computational” *Mach number*, $Ma = v_{max}/c$. Theoretically, it is required that $Ma \ll 1$, in practice, Ma should be, at least, ≤ 0.1 .

The algorithm runs as follows: (1) Initialize $\rho, \vec{u}, f_i$ and f_i^{eq} ; (2) Streaming step: $f_i \rightarrow f_i^*$ in the direction of $\vec{e}_i$; (3) Compute the macroscopic ρ and $\vec{u}$ from f_i^* ; (4) Compute f_i^{eq} ; (5) Collision step: calculate the updated distribution $f_i = f_i^* - \frac{1}{\tau}(f_i^* - f_i^{eq})$; (6) Repeat steps 2–5.

During streaming and collision, the boundary nodes need special treatment. In classical fluid dynamics, the interface between solids and the fluid is assumed to be *non-slip*, but this would slow down computations. In most studies “*bounce-back*” conditions are used, so that any flux of fluid particles that hits a boundary simply reverses its velocity.

Simulated Annealing

Simulated Annealing (SA) is an optimization algorithm devised by Kirkpatrick et al. (1983), who realized the analogy between the thermodynamics of the *annealing* of metals (i.e., the metallurgic technology of alternating heating and controlled cooling of a metal to reduce its defects), and the optimization problem of finding the *global minimum* of a multiparameter objective function. In SA a random perturbation is applied to a system's *current configuration*, so that a *trial configuration* is obtained. Let E_c and E_t be the energy levels of the current and trial configurations. If $E_c \geq E_t$, then the trial configuration is accepted and it becomes the current configuration. On the other hand, if $E_c < E_t$, then the trial configuration is accepted with a probability given by $P(\Delta E) = \exp[-\Delta E/k_B T]$ where $\Delta E = E_t - E_c$, k_B is the Boltzmann constant, and T is temperature. This step helps the system to jump out from local minima. After a sufficient number of iterations at a fixed temperature, the system approaches equilibrium, where the free energy reaches its minimum value. By gradually decreasing T and repeating the simulation process (starting out every time from the equilibrium state found for the previous T value), new lower energy configurations are achieved. Two characteristic rock-physical applications (apart from *pore-space reconstruction*, see section “[Reconstructing 3D Pore Space from 2D Sections](#)”) are *determination of fluid distribution in a porous rock* (Politis et al. 2008) and finding the *soil-components' specific thermal conductivity* (Stefaniuk et al. 2016). In the first, pore space is represented by a set of cubic voxels of side l . Each voxel is labeled by an integer indicating its phase. The solid phase is labeled as 0, the fluid phases as 1, 2, 3, $\dots$, n if there are n different pore-fillers. The saturation S_i of phase i is the volume fraction of the total pore space occupied by phase i . The distribution of the n fluid phases in the pore space is

determined assuming that it minimizes the total interfacial free energy

$$G_s = \sum_{i=0}^{n-1} \sum_{j>i}^n A_{ij} \sigma_{ij}, \quad (49)$$

where A_{ij} is the area of the interface between phases i and j , σ_{ij} the interfacial free energy per unit area between phases i , j . The interfacial free energies obey Young's equation so that the following $n(n-1)/2$ equations are satisfied:

$$\sigma_{0i} = \sigma_{ij} \cos \theta_{ij} \quad (50)$$

where θ_{ij} is the contact angle of the ij -interface with the solid surface, $i \neq 0$, $j \neq 0$, $i < j$. Suppose n fluid phases are distributed in the pore space. The number of voxels belonging to each phase depends on the saturation S_i . At each minimization step n voxels, each from a different fluid phase, are randomly selected, and $n(n-1)/2$ random swaps are performed. Each trial swap causes a change of G_s by $\Delta G_{s,ij}$ where $i < j$. If at least one $\Delta G_{s,ij} < 0$, the trial swap with the minimum $\Delta G_{s,ij}$ is accepted. If every $\Delta G_{s,ij} > 0$, the swap $i \leftrightarrow j$ is accepted with probability

$$P_{ij} = \frac{\exp[-\Delta G_{s,ij}/G_{\text{ref}}]}{n(n-1)/2} \quad (51)$$

where G_{ref} stands instead of the $k_B T$ parameter.

After a sufficient number of iterations, the system approaches equilibrium for a specific G_{ref} value. By gradually decreasing G_{ref} , and repeating the process, lower energy levels of G_s will be approached. The "cooling schedule" in this example was $G_{\text{ref}} = \lambda^N G_{\text{ref},N}$ where N is iteration number, λ ($0 < \lambda < 1$) a tunable parameter.

In the *soil-physical* application the soil consists of organic matter, water in its pores, and some soil materials say clay, silt, and sand. Suppose the thermal conductivity of water σ_w and of the organic matter σ_{om} are known, and we want to determine the probability density functions $f_{Cl}(\sigma)$, $f_{Si}(\sigma)$, $f_{Sa}(\sigma)$ for the specific thermal conductivities of the clay, silt, and sand constituents, from the knowledge of the volumetric fractions Φ_{Cl} , Φ_{Si} , Φ_{Sa} , Φ_w , Φ_{om} and from the measured overall thermal conductivity σ^{meas} of the composite soil. The *pdf* of the thermal conductivity of the solid phase is given by the mixture rule (Korvin 1982)

$$f(\sigma) = [\Phi_{Cl} f_{Cl}(\sigma) + \Phi_{Si} f_{Si}(\sigma) + \Phi_{Sa} f_{Sa}(\sigma)] / (\Phi_{Cl} + \Phi_{Si} + \Phi_{Sa}). \quad (52)$$

We create a large cubic sample consisting of $\mathfrak{N} \gg 1$ voxels. Out of them, $\mathfrak{N} \cdot (1 - \Phi_{Cl} - \Phi_{Si} - \Phi_{Sa})$ randomly selected

voxels are kept for water and organic matter, we assign to them the known thermal conductivities $\sigma_w = 0.6 \text{ Wm}^{-1} \text{ K}^{-1}$ and $\sigma_{om} = 0.25 \text{ Wm}^{-1} \text{ K}^{-1}$. The remaining $\mathfrak{N} \cdot (1 - \Phi_w - \Phi_{om})$ voxels are randomly associated with clay, silt, or sand, for example, clay is taken with the probability $\Phi_{Cl} / (\Phi_{Cl} + \Phi_{Si} + \Phi_{Sa})$. The thermal conductivities assigned to these voxels are random values with the pdf (52).

To find the *effective thermal conductivity* σ for a realization ω we numerically solve the heat flow boundary value problem. The solution yields the effective thermal conductivity $\langle \sigma(\omega) \rangle$ and averaging over a sufficient number N of realizations we get the cube's homogenized thermal conductivity as

$$\sigma^{\text{hom}} = \frac{1}{N} \sum_{j=1}^N \langle \sigma(\omega_j) \rangle \quad (53)$$

The *SA* step is used to find such *pdfs* $f_{Cl}(\lambda)$, $f_{Si}(\lambda)$, $f_{Sa}(\lambda)$ for which σ^{hom} of Eq. (53) gives the closest match to the measured data. Starting from some initial guess of the *pdfs* we make them evolve toward optimal solutions that minimize the "energy" E which, at any step, is defined as $E = (\sigma^{\text{hom}} - \sigma^{\text{meas}})^2$. As the *pdf* changes through the steps, it modifies the energy function such that $E \rightarrow E^*$, $\Delta E = E^* - E$. The change of the *pdf* is accepted with the probability $P(\Delta E) = \begin{cases} 1 & \Delta E \leq 0 \\ \exp[-\Delta E/T] & \Delta E > 0 \end{cases}$ where T is a fictitious temperature, being decreased as $T^i = \alpha T^{i-1}$, where $0 < \alpha < 1$ is a control parameter. (In Stefaniuk et al. 2016, they had $\alpha = 0.9$.)

Computational Phase Transitions

Percolation Theory

We explain the method through a hydraulic permeability study (Korvin 1992a; for a percolation treatment of Archie's law of electric conductivity, see, e.g., Hunt 2004). *Percolation Theory* was invented by S. R. Broadbent in the 1950s who worked on the design of gas masks for use in coal mines. The masks contained porous carbon granules into which the gas could penetrate. Broadbent found that if the pores were large enough and sufficiently well connected, the gas could permeate the interior of the granules; but if the pores were too small or inadequately connected, the gas would not get beyond the granules' surface. There was a *critical* porosity and pore interconnectedness, above which the mask worked well and below which it was ineffective. Thresholds of this sort are typical of *percolation processes*. In the bond-percolation problem we assume that a fraction $1-p$ ($0 < p < 1$) of the bonds of a regular grid are randomly cut and a fraction p are left uncut. Then there exists a critical fraction p_c (called *percolation threshold*) such that there is no continuous connection along the bonds of the network between the opposite

Statistical Rock Physics, Table 1 Lattices with their percolation probabilities (From Korvin 1992a, b: 22)

Lattice	Dimension	Coordination n
---------	-----------	----------------

Statistical Rock Physics, Table 2 Tortuosity exponents (after Korvin 1992a, b: 29)

Model of the percolation path	γ	Note
Straight line through the correlation length ξ	0	3D percolation
Minimum path	0.25	3D percolation
Conductive path	0.29	3D percolation-conduction
Self-avoiding random walk on uncut bonds	0.58	3D percolation
Brownian motion in 3D	0.83	
Brownian motion on a d_f -dimensional fractal	$0.83(1.5d_f - 1)$	Using the Alexander-Orbach conjecture

faces for $p < p_c$, and there exists a connection with probability 1 for $p > p_c$. For the two-dimensional square lattice the percolation threshold is 0.5. In the more general case the percolation threshold depends on the dimensionality of the network, d , and on its coordination number Z (the average number of bonds connected to any node of the network), but it is independent of the detailed structure of the network. Table 1 lists coordination numbers and percolation thresholds for some common networks. In n -dimensions, the percolation thresholds and coordination numbers conform closely to the empirical rule: $Zp_c = d/(d - 1)$.

Close to the percolation threshold ($p > p_c$) the nodes that are connected with each other by continuous paths form large clusters of average size ξ , called the *correlation length*. The correlation distance diverges for $p \rightarrow p_c$, $p > p_c$ as $\xi \propto (p - p_c)^{-\nu}$, for three-dimensional networks $\nu = 0.83$, independently of the coordination number. Percolation between two opposite nodes of a cluster, a distance ξ apart, takes place along tortuous zig-zag paths. Near the percolation threshold the length $L(\xi)$ of a typical flow path will grow as a power of ξ : $L(\xi) \propto (p - p_c)^\alpha$ for $p \rightarrow p_c$, $p > p_c$. As the correlation length ξ is the natural length scale in percolation problems, we define the tortuosity of the percolation path as: $\tau = L(\xi)/\xi \propto \xi^{\alpha-1} := (p - p_c)^{-\gamma}$ where, for different models of the percolation path the tortuosity exponents γ are compiled in Table 2.

Percolation theory was applied to compute the permeability of kaolinite-bearing sandstone samples (see Korvin 1992a, b: 28–33. Kaolinite is a “discrete-particle” clay, and it is preferentially deposited in the throats of the sandstone’s pores, completely blocking them.) If the pore structure of a sedimentary rock is converted to a discrete lattice by letting pores correspond to nodes, and throats to bonds, then the continuous Darcy flow becomes a lattice percolation. For kaolinite-bearing sandstones, if a given throat is completely blocked by kaolinite the corresponding bond will be considered as “cut.” If any throat is open with probability p and blocked by kaolinite particles with probability $q = 1 - p$, then in the equivalent lattice

percolation problem a fraction q of the bonds is randomly cut. There exists a *percolation threshold* such that the fluid cannot flow through the sample for $p < p_c$ and percolation starts for $p > p_c$. At the onset of percolation, the fluid particles follow zig-zag paths; the closer p is to p_c , the greater will be the length $L(x)$ of a typical path between two nodes, which are geometrically a distance x apart.

Expressing the *Kozeny-Carman (KC)* equation in terms of the hydraulic radius as

$k = \frac{R_{HYD}^2}{b} \cdot \Phi \cdot \frac{1}{\tau^2}$, (more precisely, $k[md] = \frac{(R_{HYD}[mm])^2}{b} \cdot \Phi \cdot \frac{1}{\tau^2} \cdot 10^9$), let λ denote the volume fraction of kaolinite, Φ porosity, then the ratio of open pore space to the total space filled by pores or clays is $p = \frac{\Phi}{\Phi + (1 - \Phi)\lambda}$. The tortuosity tends to infinity with $p \rightarrow p_c$, $p > p_c$ as $\tau \propto (p - p_c)^{-\nu}$, that is $\frac{1}{\tau^2} \propto (p - p_c)^{\nu^2}$. Define a percolation function PERC as

$$PERC = \begin{cases} 0 & \text{if } p \leq p_c \\ C_0(p - p_c)^{\nu^2} = C_0(p - p_c)^{PERC} & \text{if } p > p_c \end{cases}$$

where $PERC = \nu^2$, the normalizing constant C_0 is chosen such as to make $PERC(1) = 1$, that is $C_0 = 1/(1 - p_c)^{PERC}$. To find the prefactor in the asymptotic law $\frac{1}{\tau^2} \propto (p - p_c)^{\nu^2}$, we consider clean sand with $\lambda = 0$ kaolinite content, in which case $p = 1$ and $PERC(1) = 1$, that is for τ_0 we can choose a reasonable average tortuosity for clean sands, say $\tau_0 = 4$. Geometrical considerations give $R_{HYD} = \frac{1}{3} \cdot \frac{\Phi + (1 - \Phi)\lambda}{(1 - \Phi)(1 - \lambda)} \cdot r \sqrt{p}$ (where Φ , $\lambda \neq 1$, r is mean grain radius). The final expression for k becomes, as function of Φ , r , Z , λ (porosity, grain radius, coordination number, and kaolinite volume content):

$$k = \begin{cases} \frac{R_{HYD}^2}{b\tau_0^2} \cdot \Phi \cdot 10^9 \cdot \left(\frac{p - p_c}{1 - p_c}\right)^{PERC} & p \geq p_c \\ 0 & p < p_c \end{cases} \quad (54)$$

with $b = 2$, $\tau_0 = 4$, $p_c = 1.5/Z$; $p = \frac{\Phi}{\Phi + (1 - \Phi)\lambda}$; $R_{HYD} = \frac{1}{3} \cdot \frac{\Phi + (1 - \Phi)\lambda}{(1 - \Phi)(1 - \lambda)} \cdot r \sqrt{p}$.

Renormalization Group (RNG) Models of Rock Failure

How can we find the *critical* percolation probability p_c , or the *critical number* of cracks per unit volume that make an originally solid rock fall to pieces? Such questions of rock physics can be solved by the *Renormalizations Group* method of Statistical Physics (where it had been adopted from *Quantum Field Theory*). In the *RNG* approach, to study the critical behavior of a large system, we first consider its smaller subsystems, find their behavior, and then we put them together and express in terms of them the behavior of a larger-scale subsystem. Suppose that each of the n th order subsystems can be in one of the states $S_1, S_2, \dots, S_K$ with probabilities $p_1^{(n)}, p_2^{(n)}, \dots, p_K^{(n)}$ ($\sum_{i=1}^K p_i^{(n)} = 1$). Similarly, any $(n+1)$ st-order subsystem belongs to state S_i with probability $p_i^{(n+1)}$, ($\sum_{i=1}^K p_i^{(n+1)} = 1$). The distributions $\{p_i^{(n+1)}\}$ and $\{p_i^{(n)}\}$ are related by a system of K nonlinear equations

$$\begin{array}{rcl} p_1^{(n+1)} & = & F_1 \left[p_1^{(n)}, p_2^{(n)}, \dots, p_K^{(n)} \right] \\ \dots & \dots & \dots \\ p_K^{(n+1)} & = & F_K \left[p_1^{(n)}, p_2^{(n)}, \dots, p_K^{(n)} \right] \end{array} \quad (55)$$

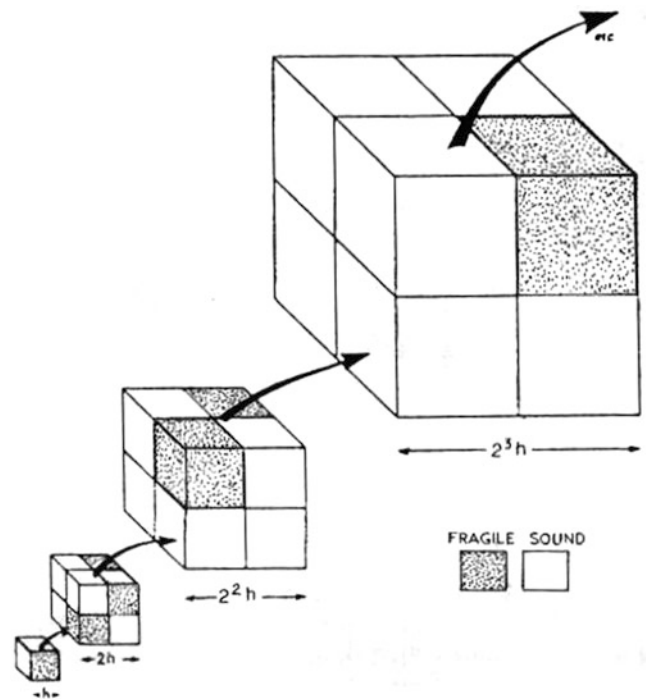
where the functions F_i depend on the physical properties of the states, and on the geometry as the $(n + 1)$ st-order subsystems are put together from the n th order ones (“*coarse-graining*”). The *critical probabilities* are solutions of Eq. (55) in the limit $n \rightarrow \infty$:

$$\begin{aligned} p_1^{(\text{crit})} &= F_1 \left[p_1^{(\text{crit})}, p_2^{(\text{crit})}, \dots, p_K^{(\text{crit})} \right] \\ &\vdots \quad \quad \quad \vdots \\ p_K^{(\text{crit})} &= F_K \left[p_1^{(\text{crit})}, p_2^{(\text{crit})}, \dots, p_K^{(\text{crit})} \right] \end{aligned} \quad (56)$$

In the case of *rock damage* (Allègre et al. 1982) there are only two states (fragile/sound) with probabilities $p^{(n)}, q^{(n)}$; $p^{(n)} + q^{(n)} = 1$ in the n th step of coarse-graining, that is Eqs. (55 and 56) become $p^{(n+1)} = F[p^{(n)}]$ and, for $n \rightarrow \infty$, $p_c = F(p_c)$.

Coarse-graining is based on the observation, that rock failure is consequence of fractures, a fracture is consequence of microfissures, and a microfissure is consequence of microcracks. This is shown in Fig. 1 where each block in the n th step can be *fragile* or *sound* with probabilities $p^{(n)}$ and $1 - p^{(n)}$.

Define a cubic cell “fragile” if it contains no continuous pillar connecting any two opposite faces. Otherwise, the cell is “sound.” Summing up the probabilities of all possible fragile configurations (with multiplicities because of topological equivalence) and letting $n \rightarrow \infty$ the critical probability satisfies the algebraic equation $p = p^4(3p^4 - 8p^3 + 4p^2 + 2)$, whose nontrivial solution is $p_c = 0.896$. We note that a



Statistical Rock Physics, Fig. 1 Coarse graining. (From Allègre et al. 1982)

different definition (Turcotte 1986) of “fragility” would lead to another equation $p = 3p^8 - 32p^7 + 88p^6 - 96p^5 + 86p^4$ whose acceptable solution yields quite a different (and more realistically!) $p_c = 0.49$.

Discrete Scale Invariance

In May, 2006, a mud volcano erupted in Java, triggered by the drilling of an oil exploration well. The pressure created by the drilling was sufficient to break an about 1800 m long connected path for the deep mud to propagate up to the surface. Can we understand, simulate, and predict such overall failures of solid rock massifs using *Renormalization Group (RNG)* and *Phase Transition Theory*?

Suppose there exists a function $F(t)$ of time (or of temperature, pressure, porosity, etc.) that can be continuously measured and which behaves as an *order parameter* in the sense of *Landau's theory of second order phase transitions*, that is, it is identically zero on one side of the phase transition, while nonzero and increasing as a power function on the other side, changing its behavior at a critical time instant t_c . How can one find such an $F(t)$ and the critical time-instant t_c ?

A way to solve this is to assume *Discrete Scale Invariance* (Saleur et al. 1996; Sornette 1998) at the critical time t_c . Suppose a system undergoes phase transition (nonconductive $\rightarrow$ conductive, predictable $\rightarrow$ chaotic, solid $\rightarrow$ damaged, etc.). In

Statistical Rock Physics, Table 3 Lopatin's TT index and corresponding stages of CH generation. (From Waples 1980)

TTI	Stage
15	Onset of oil generation
75	Peak oil generation
160	End of oil generation
~500	40° oil preservation deadline
~1000	50° oil preservation deadline
~1500	Wet gas preservation deadline
>65,000	Dry gas preservation deadline

case of *scale invariance*, the order parameter $F(t)$ satisfies $F(0) = 0$ and $F(\lambda|t - t_c|) = \mu(\lambda) F(|t - t_c|)$ for all $\lambda > 0$. This is Pexider's functional equation (Korvin 1992b: 75–76), whose solution is $\mu(\lambda) = \lambda^\alpha$; $F(|t - t_c|) = |t - t_c|^\alpha$, that is F behaves as a *power function*. In case of *discrete scale invariance* $F(\lambda|t - t_c|) = \mu(\lambda) F(|t - t_c|)$ only holds for a *finite set* of the λ values, $\lambda = \lambda_1, \lambda_2, \dots, \lambda_n$. For the case $\lambda_k = \lambda^k$ and letting $\tau = |t - t_c|$ we have

$$F(\tau) = \tau^\alpha \cdot \Theta\left(\frac{\log \tau}{\log \lambda}\right)$$

where Θ is a periodic function. Fourier expanding to first order:

$F(\tau) = A + B\tau^\alpha + C\tau^\alpha \cos(\omega \cdot \log \tau - \varphi)$ with $\omega = 2\pi/\log \lambda$, that is at phase transition the power-function-like increase is “decorated” with *logperiodic corrections*. The unknown parameters of the Fourier expansion, $A, B, C, t_c, \alpha, \lambda, \varphi$ can be found with nonlinear fitting. (An approximate value of t_c might be found from *RNG*.)

Arrhenius Law and Source-Rock Maturity

In Physical Chemistry, according to the *Arrhenius equation*, the rate r of chemical processes exponentially grows with temperature: $r = A \cdot \exp[-E_a/RT]$ where A is a constant, E_a activation energy, R the universal gas constant, and T absolute temperature. In *Basin Analysis* (Waples 1980; Lerche 1990), Arrhenius equation is our main tool to estimate the *maturation of hydrocarbon* in source rocks, in the knowledge of the basin's burial-temperature-time history. The summary effect of increasing temperature $T(t)$ during a geological time window $[t_1, t_2]$ is given by the *maturation integral* $A \cdot \int_{t_1}^{t_2} \exp[-E_a/RT(t)] dt + C_0$. For hydrocarbon maturation, the constant E_a/R is selected such as to satisfy the empirical finding that in the 50°C – 250°C temperature range, and for typical reservoir pressures, the rate of decomposition of organic materials doubles at every 10°C temperature rise. If one selects $A = 1$, $C_0 = 0$, and the integral is substituted by a discrete sum, it will go over to Lopatin's *Time Temperature Index (TTI)* = $\sum_k \Delta t_k \cdot 2^k$ where Δt_k is the time (in million years)

spent in the heat-window $100^\circ\text{C} + 10 \cdot k^\circ\text{C} \leq T_0 \leq 100^\circ\text{C} + 10 \cdot (k+1)^\circ\text{C}$; $k = 0, \pm 1, \pm 2, \dots$. N.V. Lopatin used TTI in the 1970s to predict the thermal maturation of coal. Both for coal and HC good correlation has been established between TTI and vitrinite reflection, which is a microscopically determinable measure of maturity. The threshold values given in Table 3 connect Lopatin's TT index to the distinct stages of CH generation.

Conclusions

Apart from living organisms, rocks are the most complex objects in the world. Prediction of their *effective* physical properties (such as elastic moduli, hydraulic or electric conductivity) based on their mineralogical composition and internal geometry as function of geologic age, depth, pressure, and temperature is the basic task of Geophysics, Geodynamics, and Geomechanics. Because of the inherent randomness of rocks, this task requires the powerful tools developed for treating random systems of an abundant number of degrees of freedom. This chapter described some recent Rock Physical applications of *Stochastic Geometry* and *Statistical Physics*. To help readers to follow recent developments, and to apply these techniques in their own research, each technique was described in a self-contained manner, and illustrated by actual examples. We foresee that similar to the Edwards-style *Statistical Physics of Granular Materials* (Makse et al. 2004) *Statistical Rock Physics* would become an important, independent part of *Condensed Matter Physics*.

Cross-References

- Autocorrelation
- Computational Geoscience
- Entropy
- Flow in Porous Media
- Geocomputing
- Geologic Time Scale Estimation
- Geomechanics
- Grain Size Analysis
- Inversion Theory in Geoscience
- Maximum Entropy Method
- Optimization in Geosciences
- Partial Differential Equations
- Porosity
- Pore Structure
- Porous Medium
- Probability Density Function
- Rock Fracture Pattern and Modeling

- [Scaling and Scale Invariance](#)
- [Simulated Annealing](#)
- [Simulation](#)
- [Spatial Autocorrelation](#)
- [Spatial Statistics](#)
- [Statistical Computing](#)
- [Stochastic Geometry in the Geosciences](#)

Acknowledgments This report could not have been prepared without the excellent facilities provided by my “second home,” the *Oriental Collection of the Hungarian Academy of Sciences’ Library*. When working on the first revision, I learned the sad news about the death of Dr. Károly Posgay, geophysicist, on December 25, 2019, at the age of 95. Twenty years under his guidance, 1966–1986, changed me, a budding applied mathematician, to an Earth Scientist. I dedicate this contribution to his memory.

Bibliography

- Allègre CJ, Le Mouél JL, Provost A (1982) Scaling rules in rock fracture and possible implications for earthquake prediction. *Nature* 297:47–49
- Almqvist BSG, Mainprice D, Madonna C, Burlini L, Hirt AM (2011) Application of differential effective medium, magnetic pore fabric analysis, and X-ray microtomography to calculate elastic properties of porous and anisotropic rock aggregates. *J Geophys Res* 116:B01204. <https://doi.org/10.1029/2010JB007750>
- Athy LI (1930) Compaction and oil migration. *Bull Am Ass Petrol Geol* 14:25–35
- Bagi K (2003) Statistical analysis of contact force components in random granular assemblies. *Granul Matter* 5:45–54
- Bhatnagar P, Gross E, Krook M (1954) A model for collisional processes in gases I: small amplitude processes in charged and neutral one-component system. *Phys Rev A* 94:511–524
- Doyen PE (1987) Crack geometry of igneous rocks: a maximum entropy inversion of elastic and transport properties. *J Geophys Res* 92(B8):8169–8181
- Ghanbarian B, Hunt AG, Ewing RP, Skinner TE (2014) Universal scaling of the formation factor in porous media derived by combining percolation and effective medium theories. *Geophys Res Lett* 41(11):3884–3890
- Ghassemi A, Pak A (2010) Pore scale study of permeability and tortuosity for flow through particulate media using lattice Boltzmann method. *Int J Numer Anal Meth Geomech*. <https://doi.org/10.1002/nag.932>
- Haidar MW, Reza W, Iksan M, Rosid MS (2018) Optimization of rock physics models by combining the differential effective medium (DEM) and adaptive Batzle-Wang methods in “R” field, East Java. *Sci Bruneiana* 17(2):23–33
- Hunt AG (2004) Continuum percolation theory and Archie’s law. *Geophys Res Lett* 31:L19503
- Ioannidis MA, Kwiecen MJ, Chatzis I (1996) Statistical analysis of porous microstructure as a method for estimating reservoir permeability. *J Pet Sci Eng* 16:251–261
- Ioannidis MA, Kwiecen MJ, Chatzis I (1997) Electrical conductivity and percolation aspects of statistically homogeneous porous media. *Transp Porous Media* 29(1):61–83
- Jin G, Torres-Verdin C, Toumelin E (2009) Comparison of NMR simulations of porous media derived from analytical and voxelized representations. *J Magn Reson* 200:313–320
- Kim IC, Torquato S (1990) Determination of the effective conductivity of heterogeneous media by Brownian motion simulation. *J Appl Phys* 68:3892–3903
- Kirkpatrick S (1973) Percolation and conduction. *Rev Mod Phys* 45(4):574–588
- Kirkpatrick S, Gelatt CD Jr, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220(4598):671–680
- Korvin G (1982) Axiomatic characterization of the general mixture rule. *Geos exploration* 19(4):267–276
- Korvin G (1984) Shale compaction and statistical physics. *Geophys J R Astron Soc* 78(1):35–50
- Korvin G (1992a) A percolation model for the permeability of kaolinite-bearing sandstone. *Geophys Trans* 37(2–3):177–209
- Korvin G (1992b) *Fractal models in the earth sciences*. Elsevier, Amsterdam
- Korvin G (2016) Permeability from microscopy: review of a dream. *Arab J Sci Eng* 41(6):2045–2065
- Korvin G, Oleschko K, Abdurhaheem A (2014) A simple geometric model of sedimentary rock to connect transfer and acoustic properties. *Arab J Geosci* 7(3):1127–1138
- Landau LD, Lifshitz EM (1980) *Statistical physics. Part1. (Vol. 5 of Course of theoretical physics)*. Pergamon Press, Oxford, pp 106–114
- Lerche I (1990) *Basin analysis. Quantitative methods. Part I*. Academic, San Diego
- Litwiniszyn J (1974) *Stochastic methods in the mechanics of granular bodies*. International centre for mechanical sciences. Courses and lecture notes no 93. Springer, Wien
- Makse HA, Brujic J, Edwards SF (2004) *Statistical mechanics of jammed matter*. Wiley – VCH Verlag GmbH, Berlin
- Mandelbrot B (1982) *The fractal geometry of nature*. W.H. Freeman, New York
- Matyka M, Khalili A, Koza Z (2008) Tortuosity-porosity relation in porous media flow. *Phys Rev E* 78:026306
- Mavko G, Mukerji T, Dvorkin J (1998) *The rock physics handbook: tools for seismic analysis in porous media*. Cambridge University Press, Cambridge
- Politis MG, Kainourgiakis ME, Kikkinides ES, Stubos AK (2008) Application of simulated annealing on the study of multiphase systems. In: Tan CM (ed) *Simulated annealing*. I-Tech Education and Publishing, Vienna, pp 207–226
- Saleur H, Sammis CG, Sornette D (1996) Discrete scale invariance, complex fractal dimensions, and log-periodic fluctuations in seismicity. *J Geophys Res* 101(88):17661–17677
- Sornette D (1998) Discrete scale invariance and complex dimensions. *Phys Rep* 297:239–270
- Stefaniuk D, Różański A, Łydźba D (2016) Recovery of microstructure properties: random variability of soil solid thermal conductivity. *Stud Geotech Mech* 38(1):99–107
- Stroud D (1998) The effective medium approximations: some recent developments. *Superlattice Microst* 23(3/4):567–573
- Turcotte DL (1986) Fractals and fragmentation. *J Geophys Res* B91:1921–1926
- Waples DW (1980) Time and temperature in petroleum formation: application of Lopatin’s method to petroleum exploration. *AAPG Bull* 64:916–926
- Yeong CLY, Torquato S (1998) Reconstructing random media. *Phys Rev E Stat Nonlinear Soft Matter Phys* 57:495–506
- Zimmerman RW (1991) *Compressibility of sandstones*. Elsevier, Amsterdam

Statistical Seismology

Jiancang Zhuang

The Institute of Statistical Mathematics, Research
Organization of Information and Systems, Tokyo, Japan

Synonyms

Earthquake statistics; Seismometrics; Seismostatistics

Definition

The terminology statistical seismology can be considered from two different perspectives or scopes: special and general. The special scope is similar to that used in statistical physics. Statistical seismology aims to develop stochastic models that describe the patterns of seismicity, along with related statistical inferences, to understand the physical mechanisms of earthquake occurrence and produce probability forecasts of earthquakes. The general scope of statistical seismology includes the use of statistical methods related to traditional seismology, such as Bayesian methods in geophysical inversion. Synonyms of statistical seismology include “seismometrics” and “seismostatistics.” Generally, statistical seismology falls under the special scope. Knowledge of this interdisciplinary subject has been widely applied in earthquake physics, earthquake forecasting, and earthquake engineering. This chapter only considers topics under the special scope.

Historical Overview

The term “statistical seismology” first appeared in a paper, titled “An Application of the Theory of Fluctuation to Problems in Statistical Seismology,” by Kishinouye and Kawasumi (1928) in the *Bulletin of the Earthquake Research Institute, Tokyo Imperial University*, 4, 75–83 (Fig. 1). This term was then used by Keiti Aki in a paper, titled “Some problems in statistical seismology” in Japanese on *Zisin* in 1956. In 1995, David Vere-Jones gave lectures on the statistical analysis of seismicity at the Graduate School of Chinese Academy of Science (now the University of Chinese Academy of Science) in Beijing. As suggested by Yaolin Shi, he named the lecture “statistical seismology” (Vere-Jones 2001). At present, this subject has become an important branch of seismology, serving as a bridge between seismic-waveform dominant traditional seismology and tectonics/geodynamics by providing theoretical and technical tools for analyzing seismicity.

The history of statistical seismology can be roughly divided into three episodes:

- Episode I: Exploration and accumulation of simple individual empirical laws
- Episode II: Construction of a theoretical framework and development of time-dependent forecasting models
- Episode III: Model development and forecasting and testing attempts

This chapter presents an episode-wise discussion of the concepts, theories, and methods in statistical seismology according to the above timeline.

Episode I: Exploration and Accumulation of Simple Individual Empirical Laws

This period started at the beginning of the twentieth century. In 1982, John Milne, James Ewing, and Thomas Cray installed the first model seismometer in Japan, marking the start of modern *seismology*. Seismometers enable the detection of the occurrence of global earthquakes, such that we can calculate their occurrence time and hypocenter locations and compile relatively complete earthquake catalogs. The primary applications of statistics in earthquake studies during this period were simple statistical techniques, such as linear regression and point estimates, among others, scattered among individual studies on different topics. The following subsections summarize the main findings during this stage.

The Gutenberg-Richter Law for the Magnitude-Frequency Relationship

In 1944, Gutenberg and Richter published a formula describing the relationship between any magnitude, e.g., m , and the number of earthquakes in any given region and period (Gutenberg and Richter 1944):

$$\log_{10} N(\geq m) = a - bm, \quad \text{or,} \quad N = 10^{a-bm}, \quad (1)$$

where $N(\geq m)$ is the number of earthquakes with a magnitude no less than m in the given region and period and b is the so-called Gutenberg-Richter b -value. In probability terms:

$$\begin{aligned} \Pr\{\text{magnitude} \geq m \mid \text{magnitude} \geq m_c\} &\approx \frac{N(\geq m)}{N(\geq m_c)} \\ &= \frac{10^{a-bm}}{10^{a-bm_c}} = 10^{-b(m-m_c)}. \end{aligned} \quad (2)$$

Therefore, the cumulative distribution function (c.d. f. or cdf) for the magnitude distribution is as follows:

Statistical Seismology,
Fig. 1 Head portion of the first
 page of Kishinouye and
 Kawasumi's 1928 paper

統計地震學に於ける Schwankung の 理論の應用

所員 岸 上 冬 彦
河 角 廣

An Application of the Theory of Fluctuation to Problems in Statistical Seismology.

by

Fuyuhiko KISHINOUE and Hiroshi KAWASUMI.

The object of this investigation is to introduce the application of the theory of fluctuation to some problems of statistical seismology. We have investigated the fluctuation of the number of earthquakes instrumentally registered in some units of time, taking the statistical after-effect into consideration.

The theory here applied is that given in R. Fürth: Schwankungserscheinungen in der Physik (1920, Sammlung Vieweg, Heft 48), especially in the chapter dealing with "die Schwankungen mit Wahrscheinlichkeitsnachwirkung bei intermittierende Beobachtung." A brief sketch of the theory concerned in our investigation is as follows.

Suppose that a trial consists of a great number, N , of events, and the probability of a favourable event is a very small number p , while the number of favourable events in a trial, $n = Np$, is finite. Then the probability of obtaining n favourable events in a trial, $W(n)$, is given by Poisson's formula,

$$W(n) = \frac{e^{-\nu} \nu^n}{n!}, \text{ where } \nu \text{ is the mean value of } n.$$

If we take a great number of trials in succession and the interval of each trial be constant (τ), the velocity of fluctuation $\mathcal{A} = n_1 - n_2$, the mean duration T , and the time of recurrence Θ of n favourable events, can be calculated theoretically, if we know the law which governs the appearance of n_2 favourable events after n_1 favourable events. This after-effect of n_1 on n_2 is here assumed to be effective on the next trial only, and the probability that a favourable event in a trial turns out unfavourable in the next is also assumed to be a constant P .

Then the probability that $n \pm k$ favourable events occur after n favourable events, $W(n, n \pm k)$, as well as $\bar{\mathcal{A}}(n)$, $\bar{\mathcal{A}}^2$, $T(n)$, $\Theta(n)$, and the probability of lasting

$$\begin{aligned}
F(m) &= \Pr\{\text{magnitude} \leq m\} \\
&= 1 - \Pr\{\text{magnitude} \geq m\} \\
&= \begin{cases} 1 - 10^{-b(m-m_c)}, & \text{if } m \geq m_c, \\ 0, & \text{otherwise,} \end{cases} \quad (3)
\end{aligned}$$

and the corresponding probability density function (p.d.f. or pdf) is as follows:

$$\begin{aligned}
f(m) &= \begin{cases} b 10^{-b(m-m_c)} \ln 10, & \text{if } m \geq m_c; \\ 0, & \text{otherwise,} \end{cases} \\
&= \begin{cases} \beta e^{-\beta(m-m_c)} & \text{if } m \geq m_c; \\ 0, & \text{otherwise,} \end{cases} \quad (4)
\end{aligned}$$

where β is linked with b by $\beta = b \ln 10 \approx 2.3026b$.

The Omori-Utsu Formula for the Aftershock Frequency

The Omori-Utsu formula describes the decay of the aftershock frequency with time after the mainshock, as an inverse power law. Omori (1894) examined the aftershocks of the 1891 $M_s 8.0$ Nobi earthquake, first attempting to use an exponential decay function to fit the data, but unsatisfactory results were obtained. Afterward, it was found that the number of aftershocks occurring each day can be described by the equation:

$$n(t) = K(t + c)^{-1}, \quad (5)$$

where t is the time from the occurrence of the mainshock, and K and c are constants.

Utsu (1957) postulated that the decay of the aftershock numbers can vary, proposing the following equation:

$$n(t) = K(t + c)^{-p}, \quad (6)$$

which yields better fitting results. Equation (6) is referred to as the modified Omori formula or the Omori-Utsu formula.

The Omori-Utsu formula has been used extensively to analyze, model, and forecast aftershock activity. Utsu et al. (1995) reviewed the values of p for more than 200 aftershock sequences, finding that this parameter ranges from 0.6 to 2.5, with a median of 1.1. No clear relationship between the estimates of the p -values, and the mainshock magnitudes was found.

Not only do mainshocks trigger aftershocks, but large aftershocks can also trigger their own aftershocks. To model such phenomena, Utsu (1970) used the following *multiple Omori-Utsu formula*:

$$\lambda(t) = K/(t - t_0 + c)^{-p} + \sum_{i=1}^{N_T} \frac{K_i H(t - t_i)}{(t - t_i + c_i)^{-p_i}}, \quad (7)$$

where t_0 is the occurrence time of the mainshock, t_i , $i = 1, \dots, N_T$ define the occurrence times of the triggering aftershocks, and H is the Heaviside function. The likelihood for the multiple Omori-Utsu formula is slightly more complicated than that for the simple Omori-Utsu formula but can be written in a similar manner.

One difficulty in applying the multiple Omori-Utsu formula is to determine which earthquakes have triggered an event. The largest aftershocks often (but not always) have secondary aftershocks. We can use the techniques of residual analysis described in the section as a diagnostic tool, to understand which events have secondary aftershocks.

The Båth Law for the Maximum Magnitude of Aftershocks and Other Scaling Laws

Another well-known empirical law in earthquake clusters is the Båth law, which asserts that the magnitude difference between the mainshock and the largest aftershock has a median of 1.2 (Båth 1965). A more general form proposed by Utsu (1961) has the following form:

$$M_0 - M_1 = c_1 M_0 + c_2, \quad (8)$$

where c_1 and c_2 are constants and M_0 and M_1 are the magnitudes of the mainshock and its largest aftershock, respectively. This law is useful for evaluating the possible loss caused by aftershocks.

Many empirical scaling laws related to earthquake magnitude were established during this period, including the following:

1. Several empirical laws for the relationship between the Richter magnitude, M_L Richter 1935, and the amplitude, A , of waves recorded at seismographs. For example, the Lilie empirical formula is as follows:

$$M_L = \log_{10} A - c_1 + c_2 \log_{10} \Delta, \quad (9)$$

where Δ is the epicenter distance and c_1 and c_2 are constants.

2. Relationship among the shaking intensity, I_0 , at the epicenter, magnitude, M , and focal depth, h (e.g., Gutenberg and Richter 1942):

$$M = a_0 I_0 + a_1 \log h - a_2, \quad (10)$$

where a_0 , a_1 , and a_2 are constants.

3. Scaling relationship between the number, N , of aftershocks and the size, M , of the mainshock (Yamanaka and Shimazaki 1990):

$$\log_{10} N = c \log_{10} M - d, \quad (11)$$

where c and d are constants.

4. Seismic moment, M_0 , scaling with the fault area, S (Kanamori and Anderson 1975):

$$\log_{10} S = a_0 \log_{10} M_0 - a_1, \quad (12)$$

or fault length, L (Shimazaki 2013):

$$\log_{10} L = b_0 \log_{10} M_0 - b_1, \quad (13)$$

where a_0 , a_1 , b_0 , and b_1 are constants.

Most empirical studies have been based on simple linear regressions.

Episode II: Construction of a Theoretical Framework and Development of Time-Dependent Forecasting Models

This stage began in the 1970s with two milestones: the introduction of the point process model and the development of the theory of conditional intensity in the point process. First, the development of stochastic models for earthquake risks was a requirement for earthquake engineering. Building codes, with respect to the design of a building structure, required the consideration of the probability of the occurrence of the largest ground-shaking event for a specific period into the future. In earthquake engineering, the stationary Poisson model (also referred to as the time-independent model) was often used to estimate the future earthquake hazard.

Considered as a classic reference, Vere-Jones (1970) proposed the point process to describe the process of earthquake occurrence times, focusing on the tools required to generate a functional and spectrum analysis. Vere-Jones (1973) introduced the use of the conditional intensity in statistical seismology, which is defined as the expectation of earthquake occurrence under the condition of previous knowledge on the earthquake process and/or external observations.

The core idea of model development is bringing into existing stochastic models with more nonrandomness based on physical theory and observations. During this episode, for long-term earthquake hazard evaluations, renewal models were developed by modifying the inter-event time in the Poisson model to more general random distributions. The stress release model was developed by adding Reid's elastic rebound theory to the rate function of the Poisson model. For short-term earthquake forecasting, the Reasenberg and Jones

model and the ETAS model were developed based on the Omori-Utsu formula.

Basic Formulation

Point process modeling is the key tool in statistical seismology for describing the occurrence of earthquakes, where some probabilistic rules are specified for the occurrence of earthquakes in time and/or space. When some other quantities, such as intensities or magnitudes, are attached to each event, the point process is then referred to as a marked point process.

Introduction to Conditional Intensity

Denote a point process, in time by N , and a certain temporal location, by t . We assume in the following discussion that we have known observations up to time t . The most important characteristic is the waiting time u to the next event from time t . We can therefore consider the following cumulative probability distribution function of the waiting time:

$$F_t(u) = \Pr\{\text{from } t \text{ onward, waiting time to next event} \leq u \mid \text{observations before } t\}, \quad (14)$$

with the corresponding probability density function:

$$f_t(u)du = \Pr\{\text{from } t \text{ onward, waiting time is between } u \text{ and } u + du \mid \text{observations before } t\}, \quad (15)$$

The survival function can be defined as follows:

$$S_t(u) = \Pr\{\text{from } t \text{ onward, waiting time} > u \mid \text{observations before } t\} \\ = \Pr\{\text{No event occurs between } t \text{ and } t + u \mid \text{observations before } t\}. \quad (16)$$

The hazard function can be defined as follows:

$$h_t(u)du = \Pr\left\{\begin{array}{l} \text{next event occurs between} \\ t+u \text{ and } t+u+du \end{array} \middle| \begin{array}{l} \text{Observations before } t \text{ and that no} \\ \text{event occurs between } t \text{ and } t+u \end{array}\right\} \quad (17)$$

Finally, the cumulative hazard function can be defined as

$$\Lambda_t(u) = \int_0^u h_t(s) ds. \quad (18)$$

These functions are related in the following manner:

$$S_t(u) = 1 - F_t(u) = \int_u^\infty f_t(s) ds = \exp\left[-\int_0^u h_t(s) ds\right] \\ = \exp[-\Lambda_t(u)], \quad (19)$$

$$\begin{aligned} f_t(u) &= \frac{dF_t}{du} = -\frac{dS_t}{du} = h_t(u) \exp \left[-\int_0^u h_t(s) ds \right] \\ &= \exp[-\Lambda_t(u)] \frac{d\Lambda_t(u)}{du}, \end{aligned} \quad (20)$$

$$\begin{aligned} h_t(u) &= \frac{f_t(u)}{S_t(u)} = -\frac{d}{du} [\log S_t(u)] \\ &= -\frac{d}{du} [\log(1 - F_t(u))] = \frac{d\Lambda_t(u)}{du}, \end{aligned} \quad (21)$$

$$\begin{aligned} \Lambda_t(u) &= -\log S_t(u) = -\log[1 - F_t(u)] \\ &= -\log \left[1 - \int_0^u f_t(s) ds \right]. \end{aligned} \quad (22)$$

Using the above formulas, if a random variable, W , represents the waiting time, which has a cumulative distribution function of $F_t(\cdot)$, then $F_t(W)$ belongs to a uniform distribution on $[0, 1]$; consequently, $\Lambda_t(W)$ follows an exponential distribution of the unit rate.

In the above concepts, the hazard function h_t has the property of additivity; the hazard function of the superposition of two subprocesses is the sum of the two hazard functions of these subprocesses. Additionally, if there is no event occurring between t and $t + u$, and $v \geq u \geq 0$, then $h_t(v) = h_{t+u}(v - u)$. Thus, $h_t(0)$ is identical to $h_{t'}(t - t')$ for any time t' between t and the occurrence time of the previous event. Assuming left-continuity with t , the function $h_t(0)$ is referred to specifically as the conditional intensity, denoted by $\lambda(t)$ or $\lambda(t | \mathcal{H}_t)$, where $\mathcal{H}_t$ represents the observation history up to time t , but does not include t . Based on the definition of the hazard function, we can obtain the following:

$$\lambda(t)dt = \Pr\{\text{one or more events occur in } [t, t + dt) | \mathcal{H}_t\}. \quad (23)$$

This is typically used as the definition of *conditional intensity*. When the point process, N , is simple, i.e., there is at most one event occurring at the same time and location, then $\Pr\{N[t, t + dt) > 1\} = o(dt)$. Using this relationship, $\lambda(t)$ can also be defined as follows:

$$\mathbf{E}[N[t, t + dt) | \mathcal{H}_t] = \lambda(t)dt + o(dt). \quad (24)$$

This process becomes a stationary Poisson process when $h_t(u)$ is a constant.

Likelihood, Maximum Likelihood Estimate, and AIC

Given an observed dataset of a point process, N , say $\{t_1, t_2, \dots, t_n\}$, in a given time interval $[S, T]$, the likelihood function, L , is the joint probability density of the waiting times for each of these events, which can be written as

$$L(N; S, T) = \left[\prod_{i=1}^n \lambda(t_i) \right] \exp \left[-\int_S^T \lambda(u) du \right], \quad (25)$$

or in logarithm form:

$$\log L(N; S, T) = \sum_{i=1}^n \log \lambda(t_i) - \int_S^T \lambda(u) du. \quad (26)$$

A direct use of the likelihood function is parameter estimation: When the model involves some unknown regular parameters, e.g., θ , we can estimate θ by maximizing the likelihood, i.e., the MLE (maximum likelihood estimate) is

$$\hat{\theta} = \arg_{\theta} \max L(N; S, T; \theta). \quad (27)$$

If several models fit to the same dataset, the optimal model can be selected using the Akaike information criterion (AIC, see Akaike 1974). The statistic

$$AIC = -2 \max_{\theta} \log L(\theta) + 2k_p \quad (28)$$

is computed for each model fit to the data, where k_p is the total number of estimated parameters for a given model. Based on a comparison of models with different numbers of parameters, the addition of $2k_p$ roughly compensates for the additional flexibility provided by the extra parameters. The model with the lowest AIC value is the optimal choice for forward prediction purposes.

Transformed Time Sequence and Residual Analysis

Residual analysis is a powerful and widely used tool for assessing the fit of a particular model to a set of occurrence times (e.g., Ogata 1988). Assume that we have a realization of a point process with event times denoted by $t_1, t_2, \dots, t_n$. We can then calculate the transformed event times, denoted by $\tau_1, \tau_2, \dots, \tau_n$, in such a manner that they have the same distributional properties as the homogeneous Poisson process with a unit rate parameter. If we suppose that the conditional intensity is $\lambda(t)$, the transformed time sequence for $i = 1, 2, \dots, n$, is calculated with

$$\tau_i = \int_0^{t_i} \lambda(u) du. \quad (29)$$

The sequence $\{\tau_i : i = 1, 2, \dots, n\}$ then forms a Poisson process with a unit rate as the above transformation modifies the waiting time into an exponential random variable with a unit rate.

This method can be used to test the goodness-of-fit of the model. If the fitted model, $\hat{\lambda}(t)$, is close enough to the true model, then $\{\hat{\tau}_i = \int_0^{t_i} \hat{\lambda}(u) du : i = 1, 2, \dots, n\}$ is similar to the

standard Poisson model. The deviation of the rate in the transformed time sequence from the unit rate indicates either increased activity or quiescence relative to the seismic rate based on the original model.

Models Developed in Episode II

Renewal/Recurrence Models

A renewal process is a generalization of the Poisson process. The Poisson process has independent identically distributed waiting times that are exponentially distributed before the occurrence of the next event; however, a renewal process has a more general distribution for the waiting times. This indicates that the time interval between any two adjacent events is independent from the other intervals; the occurrence of the next event depends only on the time of the last event, but not on the full history.

Renewal models yield the characteristic recurrence of earthquakes along a fault or in a region. This class of models is widely used in seismicity and seismic hazard analysis. For example, Field (2007) summarized how the Working Group on California Earthquake Probabilities (WGCEP) estimates the recurrence probabilities of large earthquakes on major fault segments using various recurrence models to produce the official California seismic hazard map. These models sometimes use the elastic rebound theory proposed by Reid (1910). According to this theory, large earthquakes release elastic strain that has accumulated since the last large earthquake. Some seismologists have deduced that longer quiescence periods indicate a higher probability of an imminent event (e.g., Nishenko and Buland 1987), whereas others contend that the data contradict this view (e.g., Kagan and Jackson 1995). Renewal models are often used to quantitatively demonstrate whether earthquakes occur temporally in clusters or quasiperiodically.

If we denote the density function of the waiting times by $f(\cdot)$, which is also usually referred to as the renewal density, the conditional intensity of the renewal process is the same as the hazard function:

$$\lambda(t) = \frac{f(t - t_{N(t_-)})}{1 - F(t - t_{N(t_-)})}, \quad (30)$$

where $t_{N(t_-)}$ is the occurrence time of the last event before t and F is the cumulative probability function of f .

The following probability functions are often selected as the renewal densities:

- The **gamma renewal** model has a density defined as follows:

$$f(u; k, \theta) = u^{k-1} \frac{e^{-u/\theta}}{\theta^k \Gamma(k)}, \quad \text{for } u \geq 0 \text{ and } k, \theta > 0, \quad (31)$$

where θ is the scale parameter, k is the shape parameter, and Γ and Γ_α are the gamma and incomplete gamma functions, respectively, defined by $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ and $\Gamma_\alpha(x) = \int_x^\infty t^{x-1} e^{-t} dt$.

- The **log-normal renewal** model has a density function defined as follows:

$$f(u; \mu, \sigma) = \frac{1}{u\sigma\sqrt{2\pi}} e^{-\frac{(\ln(u)-\mu)^2}{2\sigma^2}}, \quad \text{for } u \geq 0, \quad (32)$$

where μ and σ are the mean and standard deviation of the variable's natural logarithm, respectively, and erf is the error function.

- The Weibull distribution, used in the **Weibull renewal** model, is one of the most widely used lifetime distributions in reliability engineering. The probability density function of a Weibull random variable, X , is defined as follows:

$$f(u; \mu, k) = \begin{cases} \frac{k}{\mu} \left(\frac{u}{\mu}\right)^{k-1} e^{-(u/\mu)^k} & u \geq 0, \\ 0 & u < 0, \end{cases} \quad (33)$$

where $k > 0$ is the shape parameter and $\mu > 0$ is the scale parameter of the distribution.

- Brownian passage time model Kagan and Knopoff (1987) used an inverse Gaussian distribution to model the evolution of stress as a random walk with tectonic drift. Matthews et al. (2002) then proposed the *Brownian passage-time model* based on the properties of the Brownian relaxation oscillator (BRO). In this conceptual model, the loading of the system has two components: (1) a constant-rate loading component, λt , and (2) a random component, $\varepsilon(t) = \sigma W(t)$, defined as a Brownian motion (where W is a standard Brownian motion and σ is a nonnegative scale parameter. The Brownian perturbation process for the state variable, $X(t)$, (see Matthews et al. 2002) is defined as:

$$X(t) = \lambda t + \sigma W(t).$$

An event will occur when $X(t) \geq X_c$. These event times can be observed as the “first Passage” or “hitting” times of Brownian motion with drift. The BRO is a family of stochastic renewal processes defined by four parameters: the drift or

mean loading (λ), perturbation rate (σ^2), ground state (X_0), and failure state (X_f). In fact, the recurrence properties of the BRO (response times) can be described by a Brownian passage-time distribution, which is characterized by two parameters: (1) the mean time or period between events, (μ), and (2) the aperiodicity of the mean-time, α , which is equivalent to the coefficient of variation (defined as the ratio of the variance to the mean occurrence time). The probability density for the Brownian passage-time (BPT) model is given by

$$f(t; \mu, \alpha) = \left(\frac{\mu}{2\pi\alpha^2 t^3} \right)^{\frac{1}{2}} \exp \left\{ -\frac{(t - \mu)^2}{2\alpha^2 \mu t} \right\}, \quad t \geq 0 \quad (34)$$

with a cumulative probability function

$$F(t; \mu, \alpha) = \Phi \left(\frac{t - \mu}{\alpha \sqrt{\mu x}} \right) + \exp \left(\frac{2}{\alpha^2} \right) \Phi \left(-\frac{t + \mu}{\alpha \sqrt{\mu x}} \right), \quad (35)$$

where Φ is the cumulative probability distribution function of a standard normal random variable, i.e., F , which is the inverse Gaussian or Wald distribution.

Among the models described above, the BPT model has become the most popular, which has been used to obtain interesting results. However, there is no clear evidence showing that one model is superior to the others. This may be caused by the relatively large standard errors for the parameter estimates owing to scarce data points in the historical catalogs. Another is that this may also be due to the nature of renewal processes (e.g., Daley and Vere-Jones 2003, pages 77–79): The superposition of independent nontrivial renewal processes does not result in a renewal process unless all of the superposed processes are simply Poisson processes. This property requires that the renewal process be applied to a fault system that is relatively disjoint in behavior from the surrounding environment or is loaded by a constant tectonic stress rate. However, the identification of such fault systems appears to be technically difficult in practice.

Stress Release Model

The stress release model was proposed in a series of studies by Vere-Jones and others (see, for example, Zheng and Vere-Jones 1991) by introducing randomness into the elastic rebound theory (Reid 1910). This model assumes that the stress level in a specific region gradually builds in a linear manner with time as a result of tectonic movements, dropping suddenly with the occurrence of an earthquake, i.e., the stress level at t , $X(t)$, can be written as follows:

$$X(t) = X(0) + \rho t - S(t), \quad (36)$$

where $X(0)$ is the initial stress level, ρ is the constant loading rate from the tectonic movement, and $S(t)$ is the accumulated

stress release from earthquakes during period $[0, t)$, such that we have the following:

$$S(t) = \sum_{i: t_i < t} s_i. \quad (37)$$

Stress release related to individual earthquakes is assumed to be independent of the stress level and to scale with the magnitude m_i :

$$s_i = 10^{0.75 m_i} \quad (38)$$

as indicated by the Benioff strain. The conditional intensity is a monotonically increasing function of the stress level, usually selected as:

$$\begin{aligned} \lambda(t) &= \Psi(X) = e^{\gamma X} = \exp[\gamma(X(0) + \rho t - S(t))] \\ &= \exp[a + bt - cS(t)], \end{aligned} \quad (39)$$

where γ is a single parameter representing the strength and heterogeneity of the regional crust and $a = \gamma X(0)$, $b = \rho\gamma$, and $c = \gamma$ are the final model parameters. The assumption of an exponential distribution agrees with laboratory experiments on stress corrosion, where the mean waiting time until fracture can be approximated by an exponential function of the negative applied stress.

Reasenbergs-Jones Model

During the 1980s and 1990s, Reasenbergs and Jones (1989) combined the Gentengburg-Richter magnitude-frequency relationship and Omori-Utsu formula to forecast aftershock activities.

The conditional intensity for the full model can be expressed as:

$$\lambda(t, m) = \frac{Ks(m)}{(t + c)^p}, \quad (40)$$

where $s(m)$ is the magnitude probability density function, usually taking the form of the Gutenberg-Richter magnitude-frequency relation.

Self- and Mutually Exciting Models

Hawkes' mutual excitation model, also referred to as the self-exciting and mutually exciting earthquake models, was developed by Utsu and Ogata (see Utsu and Ogata 1997) based on the Hawkes process (Hawkes (1971)). The conditional intensity function can be written as follows:

$$\lambda(t) = \mu(t) + \sum_{i: t_i < t} g(t - t_i) + \sum_{j: s_j < t} g(t - s_j) \quad (41)$$

where $\mu(t)$ represents the background rate, which may be a function of time, but it is deterministic, i.e., it does not depend on the observation history, and $g(t)$ and $h(t)$ represent the self-exciting and external excitation terms of earthquakes inside and outside of the region, respectively. The excitation terms, $g(t)$ and $h(t)$, are in the form of a sum of the Laguerre polynomials:

$$g(t) = e^{-\alpha t} \sum_{k=0}^{K_1} p_k t^k, \quad h(t) = e^{-\beta t} \sum_{k=0}^{K_2} q_k t^k, \quad (42)$$

where K_1 and K_2 are nonnegative integers and $p_1, p_2, \dots, p_{K_1}$, $\alpha, q_1, q_2, \dots, q_{K_2}$, β are model parameters for estimation. The determination of the number of parameters, K_1 and K_2 , in the Laguerre expansions can be performed using the AIC. This model was used in earlier studies primarily to investigate possible triggering effects between different types of seismicity or the existence of precursory information in observations from non-seismic geophysical variables.

ETAS Model

In most cases, the aftershock activity consists of a secondary aftershock clustering, as described by the multiple Omori-Utsu formula (7). However, separating triggering earthquakes from others is difficult. Ogata (1988) generalized this model by removing the distinguishment between triggering events and the other events. Each event, irrespective of whether it is a small or large event, can, in principle, trigger its own aftershocks. This model incorporates the Omori-Utsu formula into the Hawkes self-exciting process.

More precisely, the temporal conditional intensity of this model is:

$$\lambda(t) = \mu + \sum_{i: t_i < t} \kappa(m_i) g(t - t_i), \quad (43)$$

where $s(m) = \beta e^{-\beta(m-m_0)}$, $m \geq m_0$ is the PDF form of the Gutenberg-Richter relation, in which m_0 is the magnitude threshold, $\kappa(m) = A \exp[\alpha(m - m_0)]$ is the mean number of events directly triggered by an event of magnitude m , and $g(u) = (p - 1)(1 + t/c)^{-p}/c$ is the probability density function (p. d. f.) of the time difference between the parent event and its child, i.e., the p.d.f. form of the Omori-Utsu formula. Ogata (1988) termed (43) the ETAS model based on an analogy with the spread of epidemics. The magnitude of an aftershock does not have to be smaller than its parent in the ETAS model. The ETAS model also belongs to a more general class of self-exciting Hawkes processes (1971).

A condition for the stability of the ETAS model is that the process must be subcritical. In data analysis, this important condition justifies whether the estimated model parameters are reasonable or irrational. For the ETAS model, the critical parameter is as follows:

$$\varrho = \int_{m_0}^{\infty} s(m) \kappa(m) dm = \frac{A\beta}{\beta - \alpha}.$$

When $\varrho < 1$, the ETAS model is stable and stationary, which requires $\beta \geq \alpha$ and $A \leq 1 - \alpha/\beta$. When $\varrho \geq 1$, there is a finite probability that the number of events in a unit time interval becomes infinite as t increases to infinity. Zhuang and Ogata (2006) discuss more details on the behavior of the ETAS model.

Simulation, Forecasting, and Performance Evaluation

Generic Simulation Method for Temporal Point Processes

Simulations allow for the visualization of a statistical model. In comparison, one can obtain ideas of how to improve the model based on the difference between the simulated catalog and real seismicity. Seismicity models specified by the conditional intensity can be simulated by at least two methods: one is the inversion method and the other is the thinning method.

The inversion method uses the relationship between the conditional intensity function and hazard function, i.e., $\lambda(t + \tau) = h_t(\tau)$, $\tau \geq 0$, and the relation between the hazard function and probability distribution function:

$$h_t(\tau) = -\frac{d}{d\tau} [\log(1 - F_t(\tau))]. \quad (44)$$

The waiting time, W , to the next event from the current time, t , can be obtained by solving for W in:

$$U = 1 - F_t(W) = \exp \left[- \int_t^{t+W} \lambda(v) dv \right],$$

where U is a uniform random variable on $[0, 1]$.

Ogata (1981) introduced the thinning method to simulate a temporal point process, N , described by a bounded conditional intensity, $\lambda(t)$. Daley and Vere-Jones (2003) modified this method for the case where $\lambda(t)$ increases boundlessly between events (such as the stress release model). A concise proof of the thinning method is provided in Zhuang and Touati (2015). This algorithm is outlined below:

If we simulate a temporal point process, N , equipped with a conditional intensity, $\lambda(t)$, in the time interval $[t_{\text{start}}, t_{\text{end}}]$, then we have the following:

- Step 1. Set $t = t_{\text{start}}$ and $i = 0$.
- Step 2. Select a positive number, $\Delta \leq t_{\text{end}} - t$, and find a positive M such that $\lambda(t) < M$ on $[t, t + \Delta]$ under the condition that no event occurs on $[t, t + \Delta]$.
- Step 3. Simulate the waiting time, τ , according to a Poisson process of rate M , i.e., $\tau \sim \text{Exp}(M)$.
- Step 4. Generate a random variate, $U \sim \text{uniform}(0, 1)$.

- Step 5. If $U \leq \lambda(t + \tau)/M$ and $\tau < \Delta$, set $t_{i+1} = t + \tau$, $t = t + \tau$, and $i = i + 1$. Otherwise, set $t = t + \min(\tau, \Delta)$, i.e., rejecting the event and progressing with time.
- Step 6. Return to step 2 if t does not exceed t_{end} .
- Step 7. Return $\{t_1, t_2, \dots, t_n\}$, where n is a sequential number from the last event.

Probability Forecasts and Performance Evaluation

Another direct use of simulation is forecasting. Vere-Jones (1998) provides a standard scheme for earthquake probability forecasting based on random simulations, relating the scoring procedure for evaluating the forecasting performance. The conditional intensity is natural for forecasting. Given observations in the past and future, the model parameters in the conditional intensity function can be obtained by fitting the model to the past observations. Once the model parameters are obtained, the occurrence process of the events in the given future interval can be simulated according to the conditional intensity and past observations. The forecasting probability can then be estimated based on

$$\Pr\{\text{phenomenon } A \text{ occurs}\} \approx \frac{\text{number of simulations in which } A \text{ occurs}}{\text{total number of simulations}}. \quad (45)$$

The evaluation of the performance of such forecasts can be performed in a measured and statistical manner using the framework of probability gain or information gain. First, a reference model is selected, usually the Poisson model. Suppose that for a future space-time-magnitude cell, the probability that an earthquake with a target magnitude occurs within it is $p^{(0)}$ according to the reference model, where p is given by the forecasting model. The probability of this forecast is:

$$p_g = X \frac{p}{p^{(0)}} + (1 - X) \frac{1 - p}{1 - p^{(0)}}, \quad (46)$$

and the information gain is as follows:

$$I_g = \log p_g, \quad (47)$$

where $X = 1$ if the target event occurs; otherwise, $X = 0$. Such an information gain can be summed over many forecast intervals:

$$I_g = \sum_{i=1}^K \left[X_i \log \frac{p_i}{p_i^{(0)}} + (1 - X_i) \log \frac{1 - p_i}{1 - p_i^{(0)}} \right], \quad (48)$$

where $p_i, p_i^{(0)}$ and $i = 1, 2, \dots, K$, are the probabilities of event occurrence based on the forecasting and reference models, respectively. The average information gain per cell and the

average information per event, defined by I_g/K and I_g/N , where K is the total number of cells over which to forecast and N is the total actual number of events occurring in the forecasting space-time-magnitude range, are often used as measures of the forecasting performance.

Episode III: Advanced Models, Modelling Techniques and Applications in Forecasting and Testing

This stage began in the late 1990s. First, models were extended for more complicated and high-dimensional cases. The stress-released model was extended to the linked (coupled) stress release model (e.g., Bebbington and Harte 2003), which is used for long-term seismic hazard estimation. The ETAS model has also been extensively examined and widely used for analyzing seismicity data. Ogata (1998) extended this model to a space-time form while Zhuang et al. (2002) proposed the ETAS model-based method for declustering an earthquake catalog, referred to as the stochastic declustering method. During this period, the importance of statistical seismology was also recognized by seismologists. An increasing number of seismologists accepted that the ETAS model should be the null hypothesis for testing hypotheses related to seismicity patterns, replacing the Poisson model (complete randomness). Another feature of the research conducted during this period was that the rate- and state-dependent friction laws were incorporated with the ETAS model to explain different phenomena in microseismicity. Additionally, rigorous statistical testing methods for earthquake forecasting were also developed. The international project ‘‘Collaboratory for Studies of Earthquake Predictability (CSEP)’’ was established by the Southern California Earthquake Center, performing rigorous statistical testing for submitted earthquake forecasting methods. Japan, Europe, New Zealand, and China also installed their own CSEP projects.

Developments Related to Stress Release Models

The linked stress release model is an extension of the univariate stress release model (e.g., Bebbington and Harte (2003)). This model considers the stress transfer and interaction between several regions. Assume that we have k tectonic subregions. The evolution of stress can be written as:

$$X_i(t) = X_i(0) + \rho_i t - \sum_{j=1}^k c_{ij} S^{(j)}(t). \quad (49)$$

For each subregion i , $S^{(j)}(t)$ is the accumulated stress released in region j over the period $(0, t)$, and the coefficient c_{ij} measures the fixed proportion of the stress drop that occurs

in region j that is transferred to region i . If $c_{ij} = 0$ for the case where $i \neq j$, this model can be reduced to the simple stress release model in each subregion. Identical to the simple stress release model, the risk in each subregion is assumed to depend exponentially on the stress, i.e., the conditional intensity takes the following form:

$$\begin{aligned}\lambda_i(t) &= \exp[\mu_i + v_i X_i(t)] \\ &= \exp\left\{a_i + b_i t - \sum_{j=1}^k C_{ij} S^{(j)}(t)\right\},\end{aligned}\quad (50)$$

where $a_i (= \mu_i + v_i X_i(0))$, $b_i = v_i \rho_i$, and $C_{ij} = v_i c_{ij}$ are the model parameters to be estimated.

ETAS Model: From Temporal to Spatiotemporal

Formulation

Not many spatiotemporal models describe seismicity. One reason is that complicated numerical computations are always involved in the implementation. Among the few spatiotemporal models that have been proposed, only the spatiotemporal ETAS model has been extensively used and widely accepted as the standard model in the context of short-term earthquake clustering. Mathematically, this model is in the class of marked branching processes with immigration. This subsection focuses on issues related to spatiotemporal ETAS models, such as the estimation of the background rate, stochastic declustering, and stochastic reconstruction, among others.

Before Ogata generalized the temporal ETAS model to a space-time version, Musmeci and Vere-Jones (1992) used space-time diffusion clustering models to analyze the seismicity in Italy. Conditional intensity functions for these models are the:

$$\lambda(t, x, y, M) = \mu(x, y) + A \sum_{t_i < t} \frac{e^{\alpha M_i} e^{-c(t-t_i)}}{2\pi\sigma_x\sigma_y t} \exp\left\{-\frac{1}{2t} \left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)\right\}, \quad (51)$$

and

$$\begin{aligned}\lambda(t, x, y, M) &= \mu(x, y) \\ &+ \frac{A e^{\alpha M} e^{-ct} t^2 C_x C_y}{\pi^2 (x^2 + t^2 C_x^2) (y^2 + t^2 C_y^2)},\end{aligned}\quad (52)$$

where A , α , σ_x , σ_y , C_x , and C_y are constants. For a fixed point (x, y) , when $t \rightarrow \infty$, the aftershocks in (51) and (52) decay with time according to $t^{-1}e^{-ct}$ and $t^{-2}e^{-ct}$, respectively.

Ogata (1998) used the following conditional intensity, which is generally used at present for the space-time ETAS model:

$$\begin{aligned}\lambda(t, x, y) &= \mu(x, y) \\ &+ \sum_{i: t_i < t} \kappa(m_i) g(t - t_i) f(x - x_i, y - y_i; m_i),\end{aligned}\quad (53)$$

where:

$$\kappa(m) = A e^{\alpha(m-m_0)} \quad (54)$$

is the expected number of aftershocks generated from a mainshock of magnitude M ,

$$g(t) = \begin{cases} (p-1)c^{p-1}(t+c)^{-p}, & \text{for } t > 0, \\ 0, & \text{otherwise,} \end{cases} \quad (55)$$

is a probability density function of the lagged time distribution of aftershocks, derived from the Omori-Utsu formula:

$$f(x, y; m) = \frac{1}{\pi\sigma(m)} f_0\left(\frac{x^2 + y^2}{\sigma(m)}\right) \quad (56)$$

is the density function of the aftershock locations from a mainshock at the origin with magnitude m . Different functions have been used for $f_0(x, y)$ and $\sigma(m)$. Usually, f_0 has a 2D gaussian density, i.e., $f_0(\omega) = \frac{1}{2D^2} e^{-\frac{\omega^2}{2D^2}}$, or inverse power, i.e., $f_0(\omega) = \frac{q-1}{D^2} \left(1 + \frac{\omega^2}{D^2}\right)^{-q}$, and $\sigma(m) \propto \kappa(m)$ or $\sigma(m) \propto [\kappa(m)]^{\frac{1}{2}}$, where D , q , and γ are extra model parameters. As shown in Zhuang et al. (2004), the inverse power form of f_0 and $\sigma(m) \propto [\kappa(m)]^{\frac{1}{2}}$ usually fit earthquake data better than other forms. Anisotropic response functions, i.e., replacing $\frac{1}{1-\rho^2} (vx^2 + 2\rho xy + y^2/v)$ with $x^2 + y^2$ in (56), have also been used to model the anisotropic shape of aftershock zones (e.g., Ogata and Zhuang 2006).

Stochastic Declustering

The proportion of the contribution from event i at the occurrence of (t_j, x_j, y_j) is as follows:

$$\rho_{ij} = \begin{cases} \frac{\xi(t_j, x_j, y_j; t_i, x_i, y_i)}{\lambda(t_j, x_j, y_j)}, & \text{when } j > i, \\ 0, & \text{otherwise.} \end{cases} \quad (57)$$

This quantity can be explained as the probability that event j is triggered by the i th event. Moreover, the probability that the j th event belongs to the background is:

$$\varphi_j = \frac{\mu(x_j, y_j)}{\lambda(t_j, x_j, y_j)}. \quad (58)$$

If we select each event, j , with probabilities ρ_{ij} or φ_j , we can form a new process consisting of the events triggered by the i th event or a new process consisting of the background events, respectively.

The above algorithm resolves the difficulties in testing hypotheses associated with earthquake clustering features, which are caused by the complicated mixture of the background seismicity and different earthquake clusters in both space and time. Different stochastic versions separate the earthquake clusters by repeating the stochastic declustering procedure for an extended period. The nonuniqueness of such realizations is an advantage of stochastic declustering as it illustrates the uncertainty in determining earthquake clusters; thus, repetition can aid in the evaluation of the significance of some properties of seismic clustering patterns. However, we can also implement these tests by directly using the probabilities, i.e., φ_j and ρ_{ij} . For example, Zhuang et al. (2005) defined the cumulative background seismicity in a certain region

$$S(t) = \sum_{i:t_i < t} \varphi_i. \quad (59)$$

This function can be used to verify whether the background seismicity in a region is stationary or not.

Estimation of Space-Time ETAS Model

Given an earthquake catalog, Zhuang et al. (2002) proposed an algorithm to estimate the model parameters and nonparametric background rate in the space-time ETAS model by iterating the following steps.

- Given some background rate, μ , we can maximize the likelihood function:

$$\begin{aligned} \log L(\theta) = & \sum_{i:t_i \in [0, T], (x_i, y_i) \in S} \log \lambda_{\theta}(t_{ki}, x_{ki}, y_{ik}) \\ & - \int_0^T \int_S \lambda_{\theta}(t, x, y) dx dy dt, \end{aligned} \quad (60)$$

where:

$$\begin{aligned} \lambda(t, x, y) = & \nu \mu(x, y) \\ & + \sum_{i:t_i < t} \kappa(M_i) g(t - t_i) f(x - x_i, y - y_i; M_i). \end{aligned} \quad (61)$$

- Calculate $\varphi_i = \frac{\mu(x_i, y_i)}{\lambda(t_i, x_i, y_i)}$ for i across all of the events.

- Obtain an improved estimation of $\mu(x, y)$. For example, Zhuang et al. (2002) used the weighted variable kernel estimate:

$$\hat{\mu}(x, y) = \frac{1}{T} \sum_i \varphi_i K(x - x_i, y - y_i, h_i), \quad (62)$$

where $K(x, y, h)$ is a 2D Gaussian kernel function with bandwidth h and the variable bandwidth, h_i , is the distance from the i th event to its n_p closest neighboring event or a fixed value if this distance is smaller than it; n_p usually ranging from 3 to 6.

This algorithm has been further developed as nonparametric estimates of the earthquake clustering model or the more complicated Hawkes models.

Simulation of the ETAS Model

Instead of using the generic inversion or thinning method, the simulation of the space-time ETAS model can be implemented based on its natural branching structure (Zhuang et al. 2004; Zhuang and Touati 2015). Suppose that we have observations, $\{(t_i, m_i) : i = 1, 2, \dots, K\}$, up to time t_{start} with the objective of forward simulating the process along time interval $[t_{\text{start}}, t_{\text{end}}]$.

Step 1. Generate a background catalog as a stationary Poisson process with the estimated background intensity, μ , on $[t_{\text{start}}, t_{\text{end}}]$, together with the original observations, $\{(t_i, m_i) : i = 1, 2, \dots, K\}$, before t_{start} , recorded as Generation 0, i.e., $G^{(0)}$.

Step 2. Set $P = 0$.

Step 3. For each event i , i.e., (t_i, m_i) , in the catalog, $G^{(\ell)}$, simulate its $N_i^{(\ell)}$ offspring: $O_i^{(\ell)} = \{(t_k^{(i)}, m_k^{(i)}) : i = 1, 2, \dots, N_i^{(\ell)}\}$, where $N_i^{(\ell)}$ is a Poisson random variable with a mean of $\kappa(m_i) t_k^{(i)}$ is generated from the probability density, $g(t - t_i)$, and $m_k^{(i)}$ is generated according to the methods described in Sect. 4.

Step 4. Set $G^{(\ell+1)} = \cup_{i \in G^{(\ell)}} O_i^{(\ell)}$.

Step 5. Remove all events whose occurrence times are after t_{end} from $G^{(\ell+1)}$.

Step 6. If $G^{(\ell+1)}$ is not empty, let $\ell = \ell + 1$ and return to Step 3; otherwise, return $\cup_{j=0}^{\ell} G^{(j)}$.

The above algorithm is computationally more efficient (and thus quicker), as the sum over past events is only performed within a single aftershock cascade for each simulated event, not over the entire history, as in the first method. This also allows events to be labelled so as to indicate their position in the branching process, such that the “family tree”

information of a background event and multiple generations of aftershocks are preserved and available for analysis.

Applications of the ETAS Model

Many interesting topics have attracted the attention of statistical seismologists.

The Båth Law

The first topic is the relation between the ETAS model and *Båth's law*. Båth's law (1965) suggests that the median magnitude difference between the a mainshock and its largest aftershock is approximately 1.2. Utsu (1957) expressed the median of this difference as a linear function of the mainshock magnitude. Many studies have shown that Båth's law can be explained as a double exponential distribution generated by the ETAS model (e.g., Luo and Zhuang (2016)).

Testing the Foreshock Hypothesis

Seismologists have extensively examined whether foreshocks can be explained by a clustering model for aftershocks, i.e., whether foreshocks are mainshocks whose aftershocks happen to be bigger. Several researchers have suggested that the probability of foreshock phenomena does not exceed the probability of a foreshock expected by an ETAS-like generic clustering model (e.g., Saichev and Sornette (2005) and Zhuang and Ogata (2006)). Many simulation studies have been conducted by different researchers; however, unlike the conclusions for Båth's law, the conclusions diverge.

Induced Seismicity

Induced seismicity caused by human activities, such as waste fluid disposal and reservoir impoundment, has attracted public attention and scientific interest. The ETAS model has been applied to this area to quantify the difference between induced seismicity and natural earthquakes, as well as to provide an assessment of related seismic hazards (Lei et al. 2008; Llenos and Michael 2013).

ETAS Model and the Rate-and-State Friction Law

The ETAS model can be used to detect the change point in the total seismicity rate through a transformed time sequence-based residual analysis (Ogata 1988, or in The background seismicity rate (Zhuang et al. 2005). The rate-and-state dependent fraction law has been used to explain such changes (e.g., Jia et al. 2014).

Further Extension and Recent Developments of the Space-Time ETAS Model

Impact of Anisotropy of Earthquake Clusters and Hypocenter Depth

Aftershock clusters of larger mainshocks are usually not isotropic around the epicenter of the mainshock but are rather

aligned along the mainshock rupture. Hainzl et al. (2008) demonstrated via ETAS simulations that ignoring this fact can lead to a biased parameter estimation, particularly to an underestimation of the (ν parameter). To account for anisotropy, (Ogata 1998) used an elliptical aftershock distribution. Recently, Guo et al. (2015b) developed a finite-source ETAS model that considers the influence that the rupture geometry of large earthquakes has on the aftershock locations. In the model formulation, for an earthquake that has a finite source with a projection area, S , on Earth's surface, the spatial response is modified as follows:

$$f(x, y; S) = \frac{\iint_S f(x-u, y-v) \tau(u, v) du dv}{\iint_S \tau(u, v) du dv} \quad (63)$$

where (u, v) is a location on S and $\tau(u, v)$ is the inhomogeneous productivity density of the mainshock along the rupture area, where the integral in the denominator is for normalization. Guo et al. (2015a) also developed a 3D ETAS model to incorporate the hypocenter depth in the model formulation, where the spatial response is as follows:

$$f(x, y, z; m_i, z_i) = f(x, y; m_i) h(z, z_i) \quad (64)$$

where $f(x, y; m_i) = \frac{q-1}{\pi D'^2} \left(1 + \frac{x^2+y^2}{D'^2}\right)$, where D' is a newly introduced parameter defined as follows:

$$h(z, z_i) = \frac{\left(\frac{z}{H}\right)^{\eta_H} \left(1 - \frac{z}{H}\right)^{\eta(1-\frac{z}{H})}}{HB \left(\eta \frac{z}{H} + 1, \eta \left(1 - \frac{z}{H}\right) + 1\right)} \quad (65)$$

where B is the beta function, $B(p, q) = \int_0^1 t^{p-1} (1-t)^{q-1} dt$, t , and H is the maximum depth. Zhuang et al. (2019) showed that the 3D and finite-source ETAS models performed better than the 2D point source ETAS model when fitting them to the Italian catalog.

Location-Dependent ETAS Parameters

Ogata (2004) developed a hierarchical space-time ETAS model to describe the difference in the seismicity clustering structure at different locations. Ogata (2004) assumed that each parameter in the model is a function of location, which has a smoothness prior. Ogata (2004) developed powerful Bayesian tools with penalized likelihoods to estimate the changes in the clustering characteristics in the form of spatial variation in the model parameters. Zhuang (2015) developed the weighted likelihood method to estimate the spatial variations in the space-time ETAS model, applying it to the seismicity of Japan.

Self-Similar ETAS Models

Vere-Jones (2005) developed a self-similar ETAS model to avoid problems caused by the cutoff magnitude in the ETAS model and enable fully self-similar features. Owing to the missing data problem for immediate after-shocks and small events; however, this model remains theoretically underdeveloped and has not yet been applied to actual seismicity data. Another development in self-similarity has been the introduction of the branching after-shock sequence (BASS) model (e.g., Turcotte et al. 2007). However, such an extension cannot be successful as the Gutenberg-Richter magnitude frequency relationship owing to the destruction of all of the events in the process (see Zhuang et al. 2013).

Other Space-Time Models: The EEPAS and Double Branching Models

The EEPAS (Every Earthquake is a Precursor According to Scale) model was developed by Evison and Rhoades (2004) and Rhoades and Evison (2004) in their studies on foreshock swarms. Its conditional intensity is as follows:

$$\lambda(t, x, y, m) = \mu\lambda_0(t, x, y, m) + \sum_{i:t_i < t} w_i \eta(m_i) r(m|m_i) f(t-t_i) g(x-x_i, y-y_i|m_i), \quad (66)$$

where f , g , and r are normal or lognormal densities, i.e.:

$$f(u|m_i) = \frac{1}{u\sigma_T\sqrt{2\pi}} \exp\left[-\frac{(\log u - a_T - b_T m_i)^2}{2\sigma_T^2}\right], \quad (67)$$

$$g(x, y|m_i) = \frac{1}{2\pi\sigma_A^2 10^{m_i b_A}} \exp\left[-\frac{x^2 + y^2}{2\sigma_A^2 10^{m_i b_A}}\right], \quad (68)$$

$$r(m|m_i) = \frac{1}{\sigma_M\sqrt{2\pi}} \exp\left[-\frac{(m - a_T - b_T m_i)^2}{2\sigma_M^2}\right] C \quad (69)$$

The weights, w_i , are usually set to one, but they can be also set according to the probability of being a background event, which is obtained from an initial stochastic declustering *REJECT* using the ETAS model. This model has a similar form as the ETAS model but is different in detail. The conditional intensity function is not applied to the entire catalog, but only for events above a second higher threshold while the summation is performed for all of the events above a primary lower threshold. This model appears to produce high average probability gains or entropy scores for moderate-term forecasts of medium to large events. Here, we refer to Rhoades and Evison (2004) for detailed applications.

Marzocchi and Lombardi (2008) proposed the double-branching model, where the triggering between earthquakes can be divided into two steps: one is short-term clustering owing to the elastic response in the upper layers of Earth, which can be well represented by the ETAS model, and the other is the interaction among events over longer time-space scales. The conditional intensity of this model has the following form:

$$\lambda(t, x, y) = \mu_2(x, y) + \sum_{i:t_i < t} \left[\frac{K_1 e^{\alpha_1(M_i - M_0)}}{(t - t_i + c)^p} \frac{C_1}{r_i^2 + d_1^2} \right] + \sum_{i:t_i < t} \left[K_2 e^{\alpha_2(M_i - M_0)} e^{-(t-t_i)/\tau} \frac{C_2}{r_i^2 + d_2^2} \right]$$

where K_1 , K_2 , α_1 , α_2 , c , p , τ , d_1 , d_2 , q_1 , and q_2 are model parameters, and C_1 and C_2 are normalization constants. Marzocchi and Lombardi (2008) used a two-step approach to estimate this model: They first used Algorithm A to obtain the background probabilities based on the stochastic declustering method, followed by fitting a self-exciting model with the second summation term in Eq. (70) to the catalog of background events. They found that the background seismicity, obtained by removing short-term clustering, is characterized by a second-step long-term clustering with a characteristic time compatible with post-seismic relaxation.

Conclusion and Future Prospects

As a subject that aims to bridge the gap between physical and statistical models, statistical seismology has developed rapidly during the last several decades. A significant achievement is the formulation of conditional intensity models for quantifying time-varying seismicity rates. The conditional intensity is natural for forecasting and the evaluation of the forecasting performance can be completed in a measured and statistical manner using the probability gain framework. The ETAS model has especially become a de facto standard model, or null hypotheses, for comparison with other models and ideas. This suggests that improvements to our understanding of earthquake clustering can be quantified by developing new ideas into models and then comparing them to ETAS, or other models, using statistical hypothesis testing. Ultimately, the rigorous testing of forecast models is necessary to improve our ability to forecast seismic hazards.

Currently, owing to our inability to observe many of the fundamental processes of the system, as well as its inherent randomness, it is difficult to deterministically predict individual earthquakes. Therefore, statistical seismology, which places a larger focus on probabilistic forecasting, represents the best quantification method with respect to our state of

knowledge. Furthermore, to provide more reliable earthquake forecast models, the challenge becomes the construction of models that can yield increased information gain with respect to a reference model, such as ETAS or complete randomness.

With the rapid development of observation technologies, an increasing amount of observational data has been obtained. These new observations provide new theories and approaches to help us understand seismicity. Based on these new observations, seismologists can develop new methods to more efficiently analyze these data and new models to connect them to the earthquake process and tectonic environments, thus providing more knowledge on the earthquake occurrence process and the ability to obtain reliable forecasts.

Acknowledgement I thank my coauthors for their contributions in writing CORSSA articles (<http://www.corssa.org/>). A large portion of these chapters were adapted from these articles. I also apologize that many important references are not cited due to limitations on the number of references.

Bibliography

- Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Autom Control* 19(6):716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- Báth M (1965) Lateral inhomogeneities in the upper mantle. *Tectonophysics* 2:483–514
- Bebbington M, Harte D (2003) The linked stress release model for spatio-temporal seismicity: formulations, procedures and applications. *Geophys J Int* 154:925–946
- Daley DD, Vere-Jones D (2003) An introduction to theory of point processes – volume 1: elementary theory and methods, 2nd edn. Springer, New York
- Evison FF, Rhoades DA (2004) Demarcation and scaling of long-term seismogenesis. *Pure Appl Geophys* 161:21–45. <https://doi.org/10.1007/s00024-003-2435-8>
- Field EH (2007) A summary of previous working groups on California earthquake probabilities. *Bull Seismol Soc America* 97(4):1033–1053. <https://doi.org/10.1785/0120060048>
- Guo Y, Zhuang J, Zhou S (2015a) A hypocentral version of the space-time ETAS model. *Geophys J Int* 203(1):366. <https://doi.org/10.1093/gji/ggv319>
- Guo Y, Zhuang J, Zhou S (2015b) An improved space-time ETAS model for inverting the rupture geometry from seismicity triggering. *J Geophys Res Solid Earth* 120(5):3309–3323. <https://doi.org/10.1002/2015JB011979>
- Gutenberg B, Richter CF (1942) Earthquake magnitude, intensity, energy, and acceleration. *Bull Seismol Soc Am* 32(3):163–191
- Gutenberg B, Richter CF (1944) Frequency of earthquakes in California. *Bull Seismol Soc Am* 34:184–188
- Hainzl S, Christophersen A, Enescu B (2008) Impact of earthquake rupture extensions on parameter estimations of point-process models. *Bull Seismol Soc Am* 98(4):2066–2072. <https://doi.org/10.1785/0120070256>
- Hawkes AG (1971) Point spectra of some mutually exciting point processes. *J R Statist Soc B (Statistical Methodology)* 33(3):438–443
- Jia K, Zhou S, Zhuang J, Jiang C (2014) Possibility of the independence between the 2013 Lushan earthquake and the 2008 Wenchuan earthquake on Longmen Shan Fault, Sichuan, China. *Seismol Res Lett* 85(1):60–67. <https://doi.org/10.1785/0220130115>
- Kagan YY, Jackson DD (1995) New seismic gap hypothesis: five years after. *J Geophys Res* 100(B3):3943–3959
- Kagan YY, Knopoff L (1987) Random stress and earthquake statistics - time-dependence. *Geophys J R Astron Soc* 88(3):723–731
- Kanamori H, Anderson DL (1975) Theoretical basis of some empirical relations in seismology. *Bull Seismol Soc Am* 65(5):1073–1095
- Kishinouye F, Kawasumi H (1928) An applicatoin of the theory of fluctuation to problems in statistical seismology, *The Bulletin of the Earthquake Research Institute, Tokyo Imperial University* 4:75–83
- Lei X, Yu G, Ma S, Wen X, Wang Q (2008) Earthquakes induced by water injection at 3 km depth within the rongchang gas field, Chongqing, China. *J Geophys Res Solid Earth* 113(B10). <https://doi.org/10.1029/2008JB005604>
- Llenos AL, Michael AJ (2013) Modeling earthquake rate changes in Oklahoma and Arkansas: possible signatures of induced seismicity. *Bull Seismol Soc Am* 103(5):2850–2861. <https://doi.org/10.1785/0120130017>
- Luo J, Zhuang J (2016) Three regimes of the distribution of the largest event in the critical etas model. *Bull Seismol Soc Am* 106:1364–1369. <https://doi.org/10.1785/0120150324>
- Marzocchi W, Lombardi A (2008) A double branching model for earthquake occurrence. *J Geophys Res* 113(B08317). <https://doi.org/10.1029/2007JB005472>
- Matthews MV, Ellsworth WL, Reasenber PA (2002) A Brownian model for recurrent earth-quakes. *Bull Seismol Soc Am* 92(6):2233–2250. <https://doi.org/10.1785/0120010267>
- Musmeci F, Vere-Jones D (1992) A space-time clustering model for historical earthquakes. *Ann Inst Stat Math* 44:1–11. <https://doi.org/10.1007/BF00048666>
- Nishenko SP, Buland R (1987) A generic recurrence interval distribution for earthquake forecasting. *Bull Seismol Soc Am* 77(4):1382–1399
- Ogata Y (1981) On Lewis' simulation method for point processes. *IEEE Trans Inf Theory* 27(1):23–31
- Ogata Y (1988) Statistical models for earthquake occurrences and residual analysis for point processes. *J Am Stat Assoc* 83(401):9–27. <https://doi.org/10.1080/01621459.1988.10478560>
- Ogata Y (1998) Space-time point-process models for earthquake occurrences. *Ann Inst Stat Math* 50(2):379–402. <https://doi.org/10.1023/A:1003403601725>
- Ogata Y (2004) Space-time model for regional seismicity and detection of crustal stress changes. *J Geophys Res* 109(B3):308. <https://doi.org/10.1029/2003JB002621>
- Ogata Y, Zhuang J (2006) Space-time ETAS models and an improved extension. *Tectonophysics* 413(1–2):13–23
- Omori F (1894) On the aftershocks of earthquakes. *J Coll Sci Imp Univ Tokyo* 7:111–200
- Reasenber PA, Jones LM (1989) Earthquake hazard after a mainshock in California. *Science* 243:1173–1176
- Reid H (1910) The mechanics of the earthquake, the California earthquake of April 18, 1906, report of the state investigation commission, vol 2. Carnegie Institution of Washington, Washington, DC, pp 16–28
- Rhoades DA, Evison FF (2004) Long-range earthquake forecasting with every earthquake a precursor according to scale. *Pure Appl Geophys* 161:47–72. <https://doi.org/10.1007/s00024-003-2434-9>
- Richter CF (1935) An instrumental earthquake magnitude scale. *Bull Seismol Soc Am* 25:1–32
- Saichev A, Sornette D (2005) Distribution of the largest aftershocks in branching models of triggered seismicity: theory of the universal bâth law. *Phys Rev E* 71(5):056,127. <https://doi.org/10.1103/PhysRevE.71.056127>

- Shimazaki K (2013) Small and large earthquakes: the effects of the thickness of Seismogenic layer and the free surface. *American Geophysical Union (AGU)*, pp 209–216. <https://doi.org/10.1029/GM037p0209>
- Turcotte D, Holliday J, Rundle J (2007) BASS, an alternative to ETAS. *Geophys Res Lett* 34(12)
- Utsu T (1957) Magnitude of earthquakes and occurrence of their aftershocks. *Zisin (J Seismol Soc Jap)* 10:35–45. (in Japanese)
- Utsu T (1961) A statistical study on the occurrence of aftershocks. *Geophys Magazine* 30:521–605
- Utsu T (1970) Aftershock and earthquake statistics (ii) – further investigation of aftershocks and other earthquake sequences based on a new classification of earthquake sequences. *J Faculty Sci, Hokkaido University, Ser VII (Geophysics)* 3:197–266
- Utsu T, Ogata Y (1997) Statistical analysis of seismicity. In: Healy J, Keilis-Borok V, Lee W (eds) *Algorithms for earthquake statistics and prediction*. International Association of Seismology and Physics of the Earth's Interior (IASPEI) library, vol 6. IASPEI, Menlo Park, pp 13–94
- Utsu T, Ogata Y, Matsu'ura RS (1995) The centenary of the Omori formula for a decay law of aftershock activity. *J Phys Earth* 43(1): 1–33. <https://doi.org/10.4294/jpe1952.43.1>
- Vere-Jones D (1970) Stochastic models for earthquake occurrence. *J Roy Stat Soc B (Methodological)* 32(1):1–62. (with discussion)
- Vere-Jones D (1973) The statistical estimation of earthquake risk. *N Z Statistician* 8:7–16
- Vere-Jones D (1998) Probability and information gain for earthquake forecasting. *Comput Seismol* 30:248–263
- Vere-Jones D (2001) The marriage of statistics and seismology. *J Appl Probab* 38A:9–13
- Vere-Jones D (2005) A class of self-similar random measure. *Adv Appl Probab* 37(4):908–914
- Yamanaka Y, Shimazaki K (1990) Scaling relationship between the number of aftershocks and the size of the main shock. *J Phys Earth* 38(4):305–324. <https://doi.org/10.4294/jpe1952.38.305>
- Zheng X, Vere-Jones D (1991) Application of stress release models to historical earthquakes from North China. *Pure Appl Geophys* 135(4): 559–576. <https://doi.org/10.1007/BF01772406>
- Zhuang J (2015) Weighted likelihood estimators for point processes. *Spatial Statist* 14(B):166–178. <https://doi.org/10.1016/j.spasta.2015.07.009>
- Zhuang J, Ogata Y (2006) Properties of the probability distribution associated with the largest event in an earthquake cluster and their implications to foreshocks. *Phys Rev E* 73(046):134. <https://doi.org/10.1103/PhysRevE.73.046134>
- Zhuang J, Touati S (2015) Stochastic simulation of earthquake catalogs. Community Online Resource for Statistical Seismicity Analysis. Available at <http://www.corssa.org>. <https://doi.org/10.5078/corssa-43806322>
- Zhuang J, Ogata Y, Vere-Jones D (2002) Stochastic declustering of space-time earthquake occurrences. *J Am Stat Assoc* 97(3):369–380
- Zhuang J, Ogata Y, Vere-Jones D (2004) Analyzing earthquake clustering features by using stochastic reconstruction. *J Geophys Res* 109(B05):301. <https://doi.org/10.1029/2003JB002879>
- Zhuang J, Chang C-P, Ogata Y, Chen Y-I (2005) A study on the background and clustering seismicity in the Taiwan region by using a point process model. *J Geophys Res* 110:B05S13. <https://doi.org/10.1029/2004JB003157>
- Zhuang J, Werner MJ, Harte DS (2013) Stability of earthquake clustering models: criticality and branching ratios. *Phys Rev E* 88(062):109. <https://doi.org/10.1103/PhysRevE.88.062109>
- Zhuang J, Murru M, Falcone G, Guo Y (2019) An extensive study of clustering features of seismicity in Italy from 2005 to 2016. *Geophys J Int* 216(1):302–318. <https://doi.org/10.1093/gji/ggy428>

Stereographic Projections

Frits Agterberg

Central Canada Division, Geomathematics Section,
Geological Survey of Canada, Ottawa, ON, Canada

Definition

Stereographic projections are geometrical mapping functions to project a sphere that contains directional features of interest onto a plane. This type of mapping is either conformal in that it preserves the angles at which different curves meet or it is area-preserving in that two sectors with the same area on the sphere have the same areas after two-dimensional projection. Both types of projections are often used in geosciences. In this chapter the emphasis will be on directed or undirected linear features that are subject to significant uncertainty. Such features can provide important information, especially in paleomagnetism and in structural geology. The differences between conformal and area-preserving projections are important in applications of remote sensing and geomatics (topographic map-making) but less important in geosciences for mapping features with significant directional uncertainties.

Introduction

Two types of graph paper associated with informal and area-preserving mapping, respectively, are known as Wulff net and Schmidt net. George Wulff (1902) was a Russian mineralogist and Adolf Schmidt was a German geophysicist (cf. Bartels 1947). Comprehensive reviews of hemisphere partition into cells with projections onto the plane can be found in Beckers and Beckers (2012) and Snyder (1993).

In a 3D space, the unit sphere is a set of points (x, y, z) with the property $x^2 + y^2 + z^2 = 1$. Using Cartesian coordinates (x, y, z) for points on the sphere and (X, Y) for the corresponding points on the plane with $z = 0$, the projection and its inverse are given by the equations:

$$(X, Y) = \left(\frac{x}{1-z}, \frac{y}{1-z} \right); \quad (x, y, z) \\ = \left(\frac{2X}{1+X^2+Y^2}, \frac{2Y}{1+X^2+Y^2}, \frac{X^2+Y^2-1}{1+X^2+Y^2} \right).$$

An alternative formulation of this result is obtained by using spherical coordinates (ϕ, θ) on the sphere where the angle ϕ (with $0 \leq \phi \leq \pi$) has azimuth θ ($0 \leq \theta \leq 2\pi$). With polar coordinates (R, Θ) on the plane, the projection and its inverse become:

$$(R, \Theta) = \left(\frac{\sin \varphi}{1 - \cos \phi}, \theta \right) = (\cot \varphi/2, \theta); \quad (\phi, \theta) = s.$$

Because of closure of the system, $\phi = \pi$ when $R = 0$. Several other conventions for this type of projection have been developed (see e.g., Gelfand et al. 1963).

Wulff Net

The required calculations introduced in the previous section can be carried out by computer, but, in geoscientific applications, it has become a common practice to use the Wulff net that is the stereographic projection of the grid of parallels and meridians on the bottom half of the sphere centered at a point on the equator. The images of the parallels and meridians on the Wulff net intersect one another at right angles. Any plane in three-dimensional space is characterized by its pole that passes through the origin and is perpendicular to the plane, which becomes a single uniquely defined point on its stereographic projection.

Further discussion and examples of Wulff net applications are provided in Agterberg (1961), a book that can be freely downloaded from the Internet by typing in its title. In a section on unit vectors in space (Agterberg 1974, Fig. 106, adapted from Whitten 1966), the use of the Wulff net is illustrated by using minor fold axes at 35 different localities in Paleozoic quartz phyllites near Monguelfo, Northern Italy. The vector mean of these directional features was also calculated. It dips 54° to the east and is shown in two Wulff net diagrams (Agterberg 1961, Figs. 17 and 18). The first of these two plots shows five cones (with top angles of 6° , 12° , 18° , 24° , and 30°) with the mean unit vector as their common axis. These five cones plot as circles. Together, they depict a so-called Fisher distribution (cf. Fisher 1953) that for small values of the standard deviation σ of deviations from the mean cannot be distinguished from an isotropic two-dimensional normal distribution (cf. Agterberg 1974, Sect. 14.3). In structural geology (cf. Ramsay 1967), it is common practice to contour frequency distributions of this type on the Wulff net (and also on the Schmidt net discussed in next section) by moving a small unit circle (e.g., 1% of total hemisphere area) across the net and counting the number of points it contains. For the 35 minor folds of the Monguelfo example, the resulting contour pattern (Agterberg 1961, Fig. 18) is a conical plot.

Schmitt Net

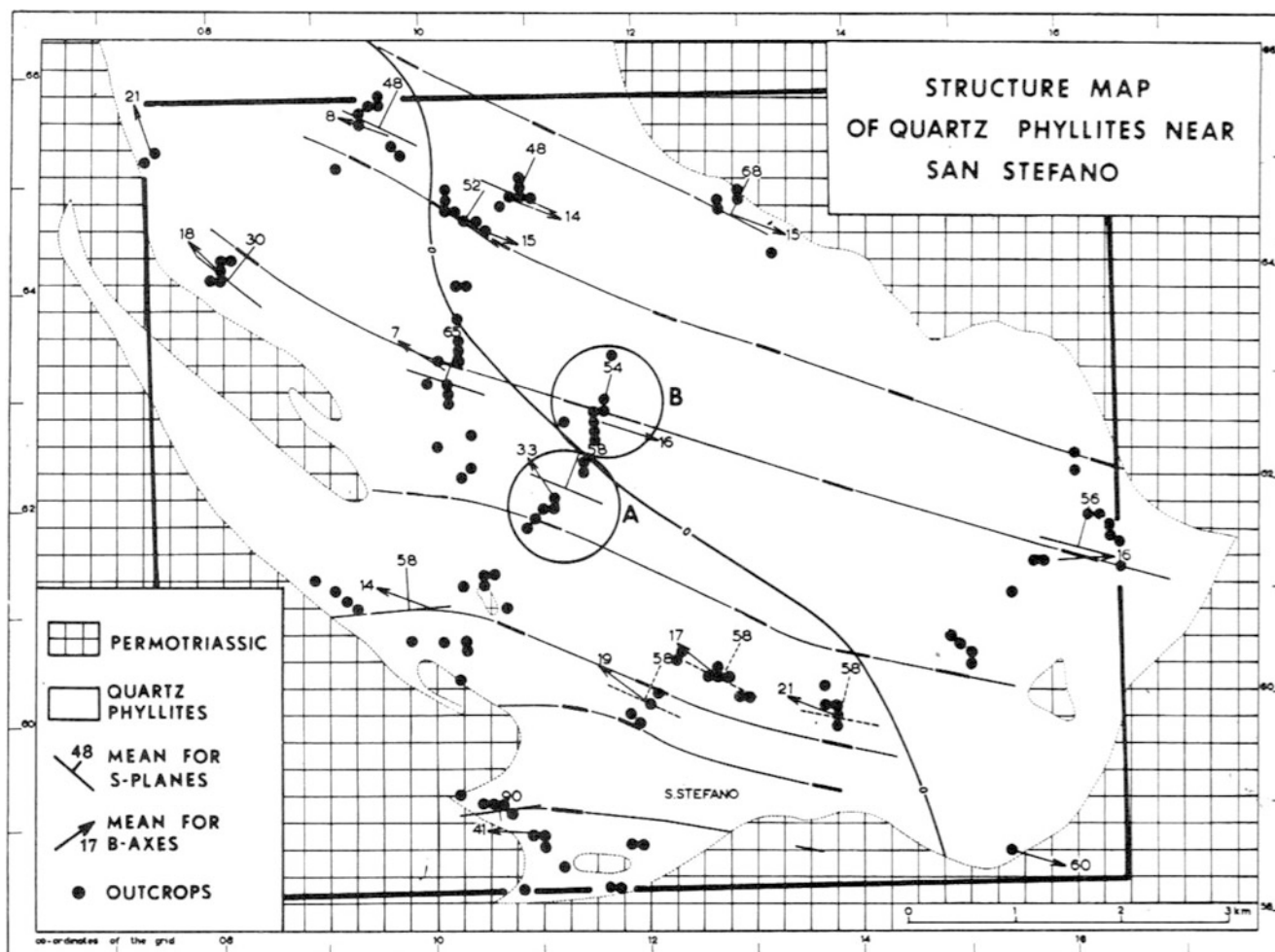
Formally, the Schmitt net is a manual drafting method for the Lambert azimuthal equal-area projection using graph paper.

Worldwide, it results in a lateral hemisphere of the Earth with its grid of parallels and meridians. The theory of hemispherical partition into equal area cells is covered in detail by Beckers and Beckers (2012). This type of plot is widely used in geophysics (especially paleomagnetism) and structural geology (Ramsay 1967; Borradaile 2003). An advantage of the Schmitt net with respect to the Wulff net is that it preserves areas. For example, two circular areas on the sphere project as two circles with the same area when the Schmitt net is used but their areas are slightly different on the Wulff net. In most geophysical and structural geologic applications, this difference in properties between the Wulff net and the Schmitt net is of no consequence. The following example after Agterberg (1985) illustrates the use of the Schmitt net in structural geology.

Figure 1 is a structure map of strongly deformed Paleozoic quartz phyllites in the San Stefano area to the east of the Dolomites in northern Italy. These quartz phyllites form the core of a later formed (Alpine) anticlinal structure. Sedimentary rocks to the northeast and southwest of the anticlinal core dip to the northeast and southwest, respectively. They are Permian in age commencing with the Permian Verrucano conglomerate that contains boulders of the Paleozoic quartz phyllites showing schistosity and minor folds similar to those found in the core of the Alpine San Stefano anticline proving that the schistosity and minor folds predated the Alpine orogeny.

Various methods of unit vector field analysis have been developed using the Alpine San Stefano anticline for testing purposes. These include two-dimensional polynomial trend-surface fitting applied separately to the three direction cosines of the measurements on minor fold axes in the outcrops shown as black dots in Fig. 1 followed by combining the estimated values to construct a unit vector field. Mathematical derivation of the underlying theory was presented in Agterberg (2012). Another approach to solve the same problem was based on fitting Fisher distributions on directional features in relatively large overlapping areas (Agterberg 1985). These later, more sophisticated methods essentially confirmed the results shown in Fig. 1 that were originally obtained using relatively simple statistical methods.

The San Stefano quartz phyllite is a soft rock that is only locally exposed, essentially along streams. Its outcrops are of two types: almost everywhere Hercynian minor folds are clearly exposed, and their azimuths and dips can be measured. However, only in relatively few outcrops can the strike and dip of the Hercynian schistosity for the entire outcrop be measured as well, because frequently it too strongly folded to determine its average attitude. Schmidt net plots for the two circular areas (A and B) in Fig. 1 are shown in Fig. 2. The mean schistosity attitudes for these two subareas (with strikes and dips shown in Fig. 1 and as $\square$ in Fig. 2) are similar, although they are based on relatively few individual



Stereographic Projections, Fig. 1 Preferred attitudes of schistosity planes and B-lineations with extrapolated regional pattern for crystalline basement of eastern Dolomites near San Stefano (after Agterberg 1961).

observations (shown as ○). The average minor fold axes are shown as arrows along with their angles of dip in Fig. 1. As shown in Fig. 2, they are more abundant than the schistosity plane measurements. Their 1% contours of total hemisphere area are shown on these two Schmitt net projections.

During Alpine orogeny, the San Stefano crystalline formed the core of an anticline in which movements along the Hercynian schistosity planes were reactivated. The trace of the resulting Alpine NW-SE trending axial plane intersects the approximately WNW-ESE striking Hercynian schistosity planes (Fig. 1). This style of folding is analogous to that described earlier by Ramsay (1967) for gneisses at Arnisdale, western Highlands of Scotland of which an example is shown in Fig. 3 where the pattern is for the azimuths of deformed B-lineations. Axial traces (intersections of axial planes with the topographic surface) for the later folds are indicated clearly in Fig. 3. The patterns of this type can become almost completely obscured when the surfaces, on which the

Cells of 100 m on a side, where one or more B-lineations were measured, are shown by dots. Domains for Fig. 2 are indicated by circles (A and B). (Source: Agterberg 1974, Fig. 110)

lineation was developed, were themselves variably oriented or when the amount of compressive strain accompanying the original folding was variable in space as in the San Stefano example of Figs. 1 and 2.

Summary and Conclusions

Two-dimensional stereographic projections of features on the sphere are either conformal preserving the angles at which different curves meet, or they are area-preserving in that two sectors with the same area on the sphere have the same areas after stereographic projection. The differences between conformal and area-preserving projections are important in applications of remote sensing and geomatics (topographic map-making) but less important in the geosciences for mapping features with significant directional uncertainties as are encountered in geophysics and structural geology. Properties

Stereographic Projections,

Fig. 2 Lineations (dots) and *s*-poles (circles) plotted on equal-area Schmidt nets, lower hemisphere, for domains A and B in central part of Fig. 1. For further explanations, see text (Source: Agterberg 1974, Fig. 111)

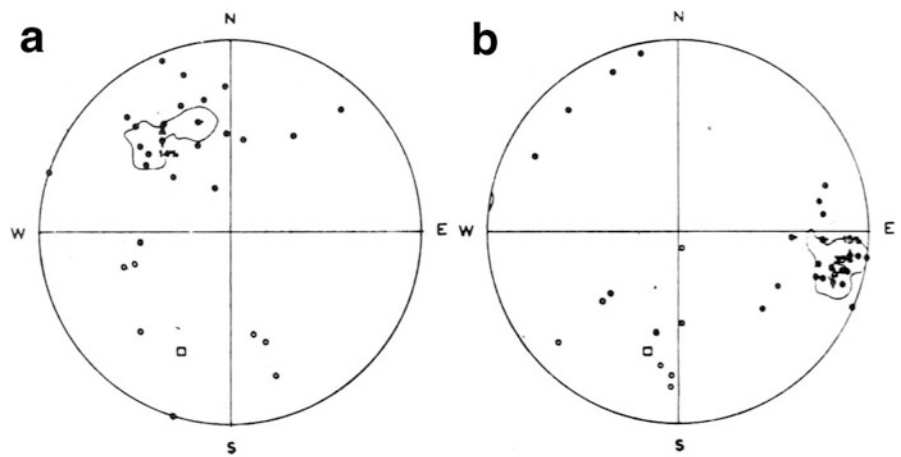
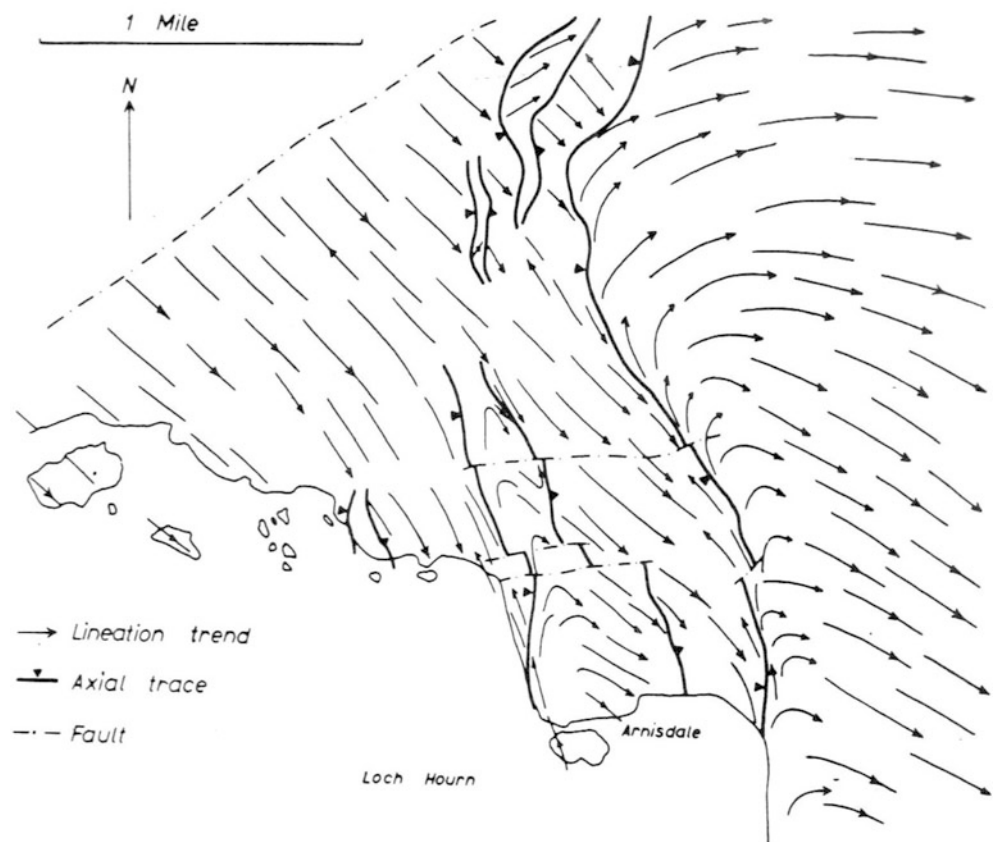
**Stereographic Projections,**

Fig. 3 Traces of deformed lineations in gneisses at Arnisdale, western Highlands of Scotland (after Ramsay 1967). (Source: Agterberg 1974, Fig. 108)



of the Wulff net and the Schmidt net were summarized in this chapter. The example discussed in detail was deciphering repeated movement of mass along schistosity planes during Hercynian and Alpine orogenies in the San Stefano quartz phyllite east of the Dolomites in North Italy.

Cross-References

- [Circular Error Probability](#)
- [Structuring Element](#)

Bibliography

- Agterberg FP (1961) Tectonics of the crystalline basement of the Dolomites in North Italy. Kemink, Utrecht
- Agterberg FP (1974) Geomathematics. Elsevier, Amsterdam
- Agterberg FP (1985) Spatial analysis in the earth sciences. In: Glaesner PS (ed) The role of data in scientific progress. Proceedings of the 9th international CODATA conference. Elsevier, Amsterdam, pp 9–18
- Agterberg FP (2012) Unit vector field fitting in structural geology. *Zeitschrift der Deutsche Geologischen Wissenschaft* 40(4/6):197–211
- Bartels J (1947) Adolf Schmidt, 1860–1944. Wiley Online Library
- Beckers B, Beckers P (2012) A general rule for disk and hemisphere partition into equal-area cells. *Comput Geom* 45(7):275–283
- Borradaile GF (2003) Statistics of Earth science data. Springer-Verlag, Berlin
- Fisher RA (1953) Dispersion on a sphere: Proceedings Royal Society of London. Series A 217:295–305
- Gelfand IM, Minlos RA, Shapiro ZY (1963) Representations of the rotation with Lorentz groups and their applications. Pergamon Press, New York
- Ramsay JG (1967) Folding and fracturing of rocks. McGraw-Hill, New York
- Snyder JP (1993) Flattening the Earth. University of Chicago Press
- Wulff HG (1902) Untersuchungen im Gebiete der optischen Eigenschaften isomorpher Kristalle. *Zeitschrift Kristallographie* 36:1–28

Stereology

Leszek Wojnar
Krakow University of Technology, Krakow, Poland

Roots and Definition

For the first time the word stereology was coined on May 11, 1961 during an informal meeting on Feldberg Mountain in the Black Forest, Germany. It was a result of discussion on invention of a single word describing three-dimensional (3D) interpretation of flat images, such as sections and projections (Underwood 1987). Soon, the International Society for Stereology, currently International Society for Stereology and Image Analysis, was founded.

One of the first definitions has expressed stereology as a body of procedures, mainly geometrico-statistical, which has the aim of obtaining information about three-dimensional structure from two-dimensional, flat images. It could also be defined as extrapolation from two- to three-dimensional space (Weibel 1987). Surprisingly, this old definition seems to be the most adequate to nowadays times.

In the monograph by Hilliard and Lawson (2003) one can find introductory remarks starting with a sentence which is very important for geosciences: Stereology began with the examination of rocks. The authors know the problem of precise definition of stereology and dynamic changes within

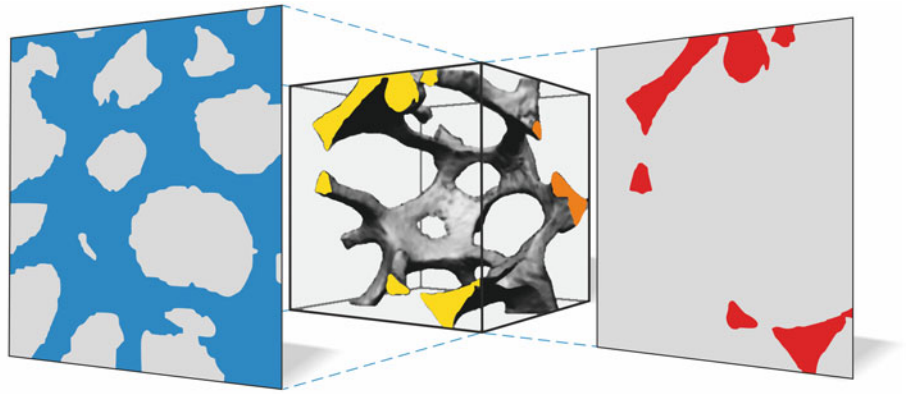
it: “The problem of defining stereology is existential: since existence precedes essence, we will not know exactly what stereology is until stereology is dead; then we will know what stereology was” (Hilliard and Lawson 2003).

The abovementioned monograph informs about 21 separate definitions for stereology. It is too much for analysis in a short chapter, so the review has to be limited. Currently at least four different definitions can be found at the top of the Internet browsers, the most popular and easily accessible source of information:

- Stereology is a branch of science concerned with inferring the three-dimensional properties of objects or matter ordinarily observed two-dimensionally (Stereology 2020b).
- Stereology is the three-dimensional interpretation of two-dimensional cross sections of materials or tissues. It provides practical techniques for extracting quantitative information about a three-dimensional material from measurements made on two-dimensional planar sections of the material (Stereology 2020a).
- Stereology is the process of sampling and counting material using specific protocols to obtain an estimate of a quantitative parameter, such as number, length, volume, etc. Obtaining accurate estimates is a critical component of producing a study outcome with statistical validity. By using appropriate sampling procedures and applying counting or measuring protocols (known as probes), stereology makes it possible to avoid potential artifacts and errors that could significantly skew results (Peterson 2020).
- Even the International Society for Stereology and Image Analysis gives a definition which is far from precision and completeness: It is an interdisciplinary field that is largely concerned with the three-dimensional interpretation of planar sections of materials or tissues. It provides practical techniques for extracting quantitative information about a three-dimensional material from measurements made on two-dimensional planar sections of the material (ISSIA 2020).

From the above-listed definitions it is difficult to find close relation to modern techniques based on computer-aided image analysis and/or tomography. Each of these definitions is generally correct but none of them seems to be complete. In spite of the passage of time clear explanation of the borders of stereology, which have significantly widened, is still problematic. Equipment for automatic image analysis was commercialized several years after introduction of stereology. Consequently, stereology has dealt initially only with manual methods which are very laborious and time consuming. As a result, many researchers and practitioners recognize stereology as an outdated tool, not applicable to modern, computerized, and often automated methods.

Stereology, Fig. 1 Three-dimensional structure with its projection (objects marked in blue) and exemplary section (objects marked in red)
(Source: Wojnar L (2020))



Contemporary stereology can be defined in the following words:

Stereology is an interdisciplinary branch of science comprising theoretical background, rules, and procedures for quantitative characterization of geometrical properties of real or virtual objects detectable in any kind of images. The most traditional task of stereological analysis is three-dimensional interpretation of flat images, being sections or projections of the examined structures (see Fig. 1).

Scope and Limitations

Classical methods of stereology are applicable to virtually any types of objects in 3D space, like particles of various minerals, fibers, edges, pores, fracture surfaces, etc. Nevertheless, in contrast to tomography, which can effectively reconstruct the complete three-dimensional geometry of objects analyzed, the goal of stereology is not exact reconstruction but obtainment of objective, quantitative, and possibly simple description of the structure tested, suitable for classification and comparison with other structures. It is noteworthy that in medicine tomography and stereology are often used together (Hilliard and Lawson 2003).

Inspection of any three-dimensional structure leads to different datasets which are limited to three categories, namely projections, sections, and 3D reconstructions (compare Fig. 1):

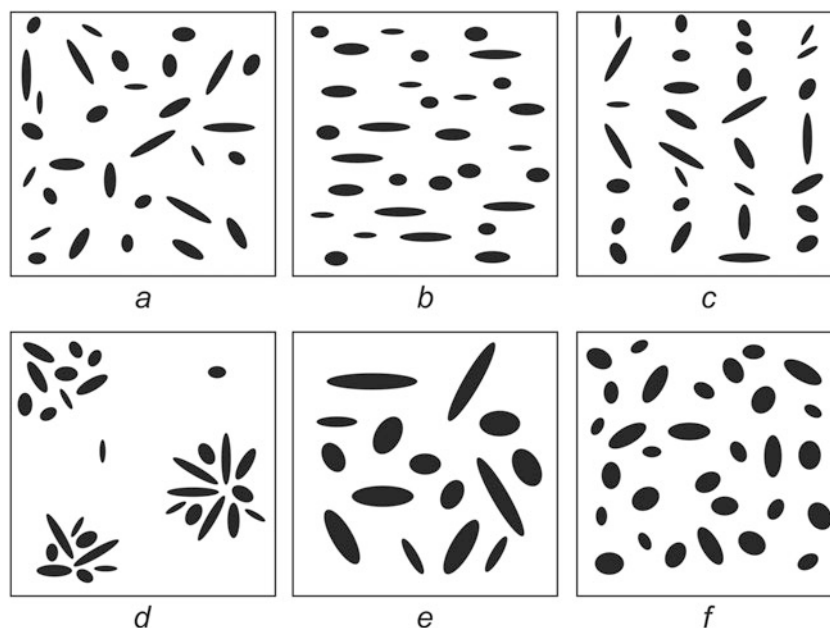
- Projection is the most commonly available kind of visual information. Photos taken by cameras or smartphones are in fact projections. Our sense of sight provides some 3D information after analysis of two slightly different projections. Transparency of objects or X-ray inspection allow for some insight into the internal arrangement of the investigated structures but it also provides projections.

- Section can be either physical or virtual as a subset of information from tomographic image. Single section usually does not deliver any information of 3D character.
- Three-dimensional (3D) reconstruction ensures the most complete information about the 3D structure but requires special apparatus and has numerous limitations connected with resolution and relatively high cost of analysis. Moreover, X-ray tomography cannot differentiate objects with similar atomic mass.

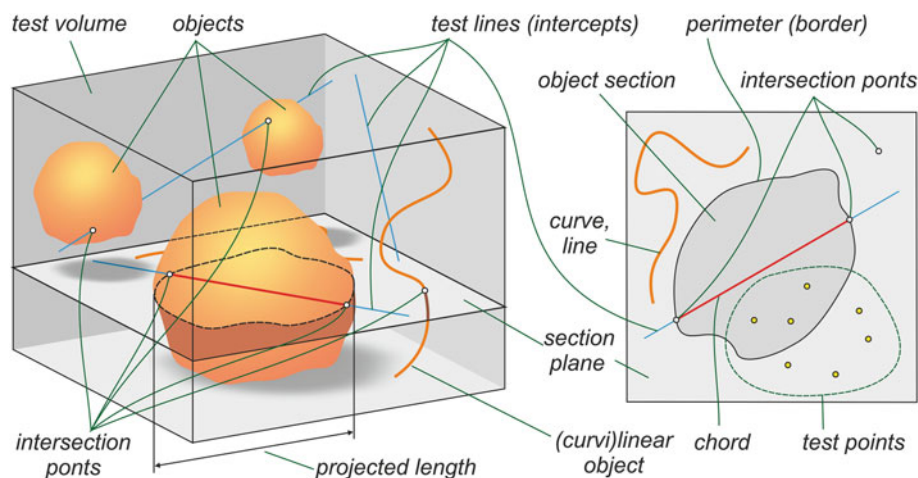
Main limitations of stereological methods are clearly visible in Fig. 1. Stereology does not allow to characterize spatial connectivity or three-dimensional shape and orientation of objects under consideration. This kind of information can be effectively obtained using tomographic inspection. Consequently, it makes an impression that stereology is no longer necessary. On the other hand, tomography is very effective in detection of pores but simultaneously useless in detection of the boundaries in polycrystalline, single-phase materials. Sections and projections can be obtained using numerous methods: photographic documentation, light microscopy, including polarization or DIC (differential interference contrast), transmission (TEM), and scanning (SEM) electron microscopy, EBSD (electron backscatter diffraction), FIB-SEM (focused ion beam), radiography, ultrasonography, AFM (atomic force microscopy), radar and sonar, etc. Multitude of these techniques assures the necessity of stereological analysis.

Stereology is based on statistical analysis of the spatial properties of objects and returns statistically exact estimators of various parameters. However, derivation of the basic relations requires some assumptions concerning properties of the examined structure. The most important statement tells that the arrangement of objects should be IUR, it means independent, uniform, and random. This assumption assures that all the sections are statistically identical and enables derivation of numerous equations. Unfortunately, in many cases the investigated structure is not IUR (compare Fig. 2b–d).

Stereology, Fig. 2 Models of different arrangements of objects: (a) IUR objects, (b) linearly oriented objects, (c) periodic arrangement of objects, (d) clustered objects, (e) different size, (f) different shape. Details in the text (Source: Wojnar L (2020))



Stereology, Fig. 3 Basic notions used in stereological analysis (Source: Wojnar L (2020))

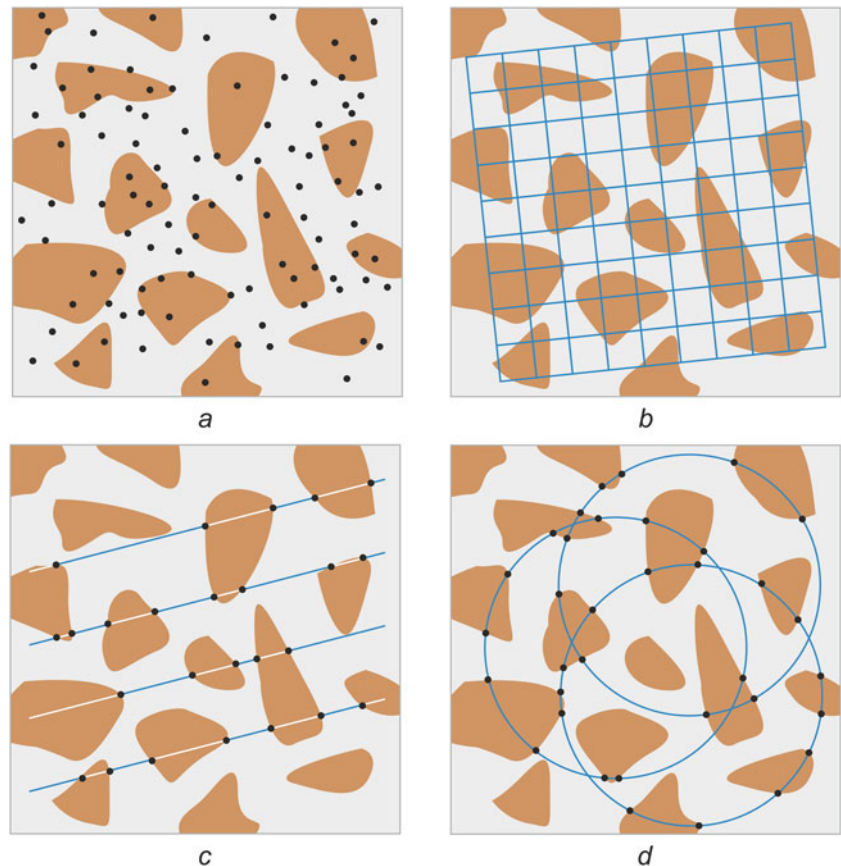


Possible solutions to such problems are presented later in this chapter.

This is postulated that for full stereological description of a group of objects it is necessary and sufficient to quantify four characteristics:

- **Amount:** This is the most popular and relatively simple for evaluation parameter. The amount of objects is quantified by their total volume in 3D space or by the total surface area of sections in 2D images. Number of objects is not a measure of amount. All the images in Fig. 2 have the same amount of black objects.
- **Size:** The most natural measure of an individual object size is its volume, but this measure is usually impossible for precise evaluation except of tomographic images. There are different, relatively easy for evaluation, measures of size. However, usually there is not exact relation between different size characteristics. Size of objects in Fig. 4 is the same in all the images. Greater objects can be found only in Fig. 4e.
- **Shape:** This parameter is very important but simultaneously very difficult for quantification due to the fact that even from a small piece of plasticine an infinite number of different shapes can be created. Consequently, any parameter unambiguously describing shape would have an

Stereology, Fig. 4 Manual analysis in stereology: (a) random test points, (b) networked test points as knots of the network, (c) linear method, (d) curvilinear test lines (Source: Wojnar L (2020))



unlimited value and this is obviously impossible. To partially solve this problem, the so-called shape factors are used. In Fig. 4a–e all the images contain objects of the same shape. Only objects in Fig. 4f have different shape.

- Arrangement (in the sense of location): This factor can have decisive impact on properties of various materials, like, for example, rocks or soils. Simultaneously, quantification of arrangement is the most difficult task. Object distribution is approximately uniform and random only in Figs. 4a, e, and f. In Figs. 4b, c, and d objects are oriented, periodically placed or clustered, respectively. This is also noteworthy that every structure, if only observed at sufficiently large magnification, becomes inhomogeneous. Therefore, quantification of arrangement has no universal measures.

Basic Principles and Methods

Basic notions used in the further part of this chapter are presented in Fig. 3. Stereological analysis is limited to a volume of space, called test volume. It contains objects, which can be three-dimensional, flat, or linear. They are examined using section planes, test lines (intercepts), and

test points. Test lines cut the surfaces of objects or edges of object sections at locations called intersection points. Line segments lying within an object, with the length limited by intersection points, are called chords.

Symbols used in stereological analysis are internationally recognized and used (Table 1). If the symbol is denoted by a capital letter, it refers to the whole test space or the whole set of tested objects. Lowercase letters denote either single object or mean values for a set of objects. In some cases, for example, in medicine, different notion for the same quantities can be found.

The most commonly used 12 stereological parameters are called global as they describe the whole sets of objects, not individual ones. These parameters have the symbol X_Y , where X denotes the measured quantity (for example, surface area) and the subscript index Y denotes reference space (for example, test volume). In turn mean values characterizing single objects are called local parameters (Table 2).

Local parameters, being simply mean values of various parameters, are easier for understanding and interpretation. Unfortunately, it can be completely misleading. For example, in the case of two, approximately equinumerous sets of large and small objects the mean value can describe a nonexistent object. It can lead to erroneous conclusions. In such a case a

Stereology, Table 1 Basic symbols and quantities used in stereology

Symbol	SI unit	Explanations and comments
A, a	m^2	Area; surface area of a flat surface (object, section, test area, etc.)
B, b	m	Border; length of an object outline, perimeter, or sum of such outlines
D, d	m	Diameter; diameter of a circle, sphere, object, etc. Sometimes the so-called equivalent diameter is traditionally used for characterization of irregular figures (usually sections). It is equal to the diameter of a circle with surface area equal to the surface area of the analyzed object
L, l	m	Length; length of test lines, chords, projections, etc. Also used for curved lines, like borders, perimeters, central lines of elongated objects, etc.
N, n	–	Number; number of objects, chords, etc.
P, p	–	Point; number of test or intersection points, knots of test grids, etc.
S, s	m^2	Surface; surface area of a curvilinear surface of an object or set of objects
V, v	m^3	Volume; volume of an object, set of objects, or test volume

Stereology, Table 2 Selected parameters used in stereology

Symbol	SI unit	Explanations and comments
Global parameters referred to a three-dimensional space		
V_V	–	Volume fraction; ratio of the total volume of tested objects to the total volume of tests space
S_V	m^{-1}	Specific surface area; ratio of the total surface area of tested objects to the total volume of test space
L_V	m^{-2}	Specific length; ratio of the total length of tested lineal objects to the total volume of test space
N_V	m^{-3}	Number of objects per unit test volume
Global parameters referred to a two-dimensional space		
A_A	–	Area fraction; ratio of the sum of section areas to the total area of the test plane
L_A	m^{-1}	Length per unit test area; ratio of the total length of examined linear objects to the area of test surface containing these objects
N_A	m^{-2}	Number of objects per unit test area; ratio of the total number of examined objects to the area of test surface containing these objects
P_A	m^{-2}	Number of points per unit test area; ratio of the total number of intersection or test points to the area of test surface containing these points
Global parameters referred to a one-dimensional space		
L_L	–	Linear fraction; ratio of the sum of chords of tested objects to the total length of test lines
N_L	m^{-1}	Number of objects per unit test line length
P_L	m^{-1}	Number of points per unit test line length
Global parameters referred to a zero-dimensional space		
P_P	–	Point fraction; ratio of the number of points hitting the tested objects and total number of test points
Local parameters		
$\bar{v}$	m^3	Mean object volume
$\bar{a}$	m^2	Mean surface area of a flat object or section
$\bar{s}$	m^2	Mean object surface area in 3D
$\bar{l}$	m	Mean chord or projection length
$\bar{d}$	m	Mean object diameter
$\bar{\lambda}$	m	Mean free path; mean distance between objects measured along a single direction
$\bar{h}$	m	Mean object height or spacing

safe solution is the use of statistical distributions of measured values.

Global and local parameters are closely related (see Eqs. (1)–(5), Table 3). Some of the relations presented in Table 3 are older than stereology but only development of this new branch of science allowed for their orderly organization, classification, and theoretical foundations. Surprisingly, in spite of the passage of time, the 55-years-old monograph by Underwood is still often cited as a source of

information on classical stereological methods (Underwood 1970).

The roots of stereology are firmly connected with the idea of three-dimensional interpretation of two-dimensional cross sections. Surprisingly, only four relations from Table 3 devoted to evaluation of V_V , S_V , L_V , and N_V give parameters which are strictly related to three-dimensional structures. It explains the need for upgrading the definition of stereology, presented in the introductory section.

Stereology, Table 3 Selected equations used in stereology

Equation number	Equation	Explanations and comments
(1)	$\bar{v} = \frac{V_V}{N_V}$	Exact relation of limited applicability due to problems with evaluation of N_V
(2)	$\bar{s} = \frac{S_V}{N_V}$	Easy tool for estimation of the mean surface area of an object of any shape. The use of Eq. (7) is necessary. Limited applicability due to problems with evaluation of N_V
(3)	$\bar{a} = \frac{A_A}{N_A}$	Simple relation, very convenient in the case of single-phase materials where $L_L = 1$
(4)	$\bar{l} = \frac{L_L}{N_L}$	Simple relation, very convenient in the case of single-phase materials where $L_L = 1$
(5)	$\lambda = \frac{1-L_L}{N_L}$	Characterizes object spacing in a single direction, parallel to the test line(s). Allows for measurements in different directions, and so is suitable for analysis of oriented structures
(6)	$V_V = A_A = L_L = P_P$	The simplest possible, very elegant, and very useful stereological equation. Probably the most popular stereological relation. It is restricted to IUR structures only
(7)	$S_V = 2P_L$	One of the most important relations in stereology. Unbiased estimation of the surface area of curvilinear objects in the space. Not applicable for oriented sets of surfaces
(8)	$L_V = 2P_A$	Unbiased estimation of the length of curvilinear objects in the space. Not applicable for oriented sets of lines
(9)	$N_V = \frac{N_A}{\bar{h}}$	Relation difficult for practical application due to problems with evaluation of the $\bar{h}$ value. $\bar{h}$ denotes mean height of an object, i.e., its projected length in direction perpendicular to the test plane
(10)	$L_A = \frac{\pi}{2}P_L$	Unbiased estimation of the length of curvilinear objects in the plane. Not applicable for oriented sets of lines
(11)	$N_L = \frac{P_L}{2}$	Practically obvious and straightforward relation which allows unbiased automatic evaluation of the number of objects cut by a section line
(12)	$N_A = \frac{1}{A}(n_i + 0.5n_e + 0.25n_c)$	Known as Jeffries' method for a rectangle. Used for unbiased estimation of the number of objects. Symbols denote: n_i – number of internal objects, n_e – number of objects cut by the image edge, n_c – number of objects at the corners of the image. Gives small, negligible errors in the case of objects placed outside corners but simultaneously cut by two edges of the image

Figure 4 illustrates classical, manual methods of analysis. One of the main achievements in stereology is the possibility to evaluate complex quantities, like volume, by simple measurements or even counts. An excellent example of the beauty and cleverness of stereological methods is Eq. (6) where volume fraction can be estimated on the basis of simple counts. Test points are placed randomly over the analyzed image (Fig. 4a) and volume fraction is estimated as a ratio of the number of points hitting the analyzed objects to the total number of points.

The use of randomly distributed points is correct, but in practice inconvenient. Much attention has to be paid to control which points have already been analyzed. Much faster and more convenient is application of a network of lines, where test points are simply nodes of the network (Fig. 4b). If the objects are IUR, application of the networked points preserves randomness of the whole analysis. During manual analysis some points seem to be placed just on the boundary between objects and the matrix. In such a doubtful case the point should be counted as $\frac{1}{2}$ instead of 1. This simple rule allows to avoid systematic errors.

In the case of repeated measurements performed on the same image the test grid can be applied at various locations and rotated by different angles. Consequently, some scatter of results will be observed. The relative error γ of V_V estimation using point method can be expressed as:

$$\gamma = u_\alpha \sqrt{\frac{1 - V_V}{P_T \cdot V_V}} \quad (13)$$

where:

α – significance level, usually 0.05,

U – statistics of the normal distribution; for $\alpha = 0.05$

$$U_\alpha = 1.96 \cong 2$$

P_T – number of test points

So, if $V_V = 0.1$, then 3600 points have to be tested in order to ensure 10% relative error. It should be stressed that Eq. (13) returns relative error of the method only, without any relation to improper sampling, nonhomogeneity of the material tested, etc. The number of points necessary to be tested is really very high. Therefore automated methods, usually using computerized image analysis, are used whenever possible. In principle the method is the same in manual and automated measurements. The main difference depends on the fact that automated methods use all the points (pixels) in the image and therefore relative error of the method reaches the value of zero. In spite of it errors in automatic analysis can be even larger than in the case of manual methods. The main source of problems is improper segmentation. Due to the digital character of images even a relatively small change in threshold values can introduce (in the case of small objects) more than 100% difference in the results.

Linear methods (compare Fig. 4c) are less popular in estimation of volume fraction. Counting is not sufficient in this case and one has to measure the lengths of chords, seen as white line segments in Fig. 4c. Such measurements can take more time and introduce more errors than simple point counting. This is a consequence of errors in length measurements. On the other hand, linear analysis is a perfect tool for estimation of the specific surface area (Eq. (7)). It is noteworthy that P_L estimation can be executed with the help of test lines of any shape, for example, circular (Fig. 4d). Irregular shapes of test lines are not used because it is difficult to fix their lengths. Relative error of the method of volume fraction evaluation using linear analysis can be expressed as:

$$\gamma = U_\alpha \sqrt{\frac{2(1 - V_V)^2}{N_T}} \quad (14)$$

where

N_T – number of chords taken for analysis

Although shape of objects building the internal structure of solid bodies is one of the most important factors affecting their properties, it is difficult to find a satisfactory definition. Possibly the best one can be found in Wikipedia Web page: the shape of a set of points is all the geometrical information that is invariant to translations, rotations, and size changes (Shape 2020).

Due to the reasons presented earlier, it is impossible to characterize any shape unambiguously by single numbers. The problem is partially successfully solved by introduction of the so-called shape coefficients. They are dimensionless numbers, thus invariant with size changes, describing how similar a given shape is to the ideal, model shape. Consequently, any shape factor has only a limited applicability, related to the reference object. In most cases circle is chosen as a reference shape for objects observed in 2D planes. There exists no universally accepted symbol for shape factor and here the letter q is chosen.

The simplest shape factor $q1$ is suitable for characterization of the elongation of ellipsoidal objects. It is a ratio of the

longest projection length a and projection length b measured in direction perpendicular to the longest projection (Fig. 5a):

$$q1 = \frac{a}{b} \quad (15)$$

This shape factor reaches the value $q1 = 1$ for a circle and corresponds well with rough judgement done without measurements by a human observer. However, it can reach very high values for significantly elongated objects. Therefore a modified elongation parameter $q2$ is used, for example, in image analysis software:

$$q2 = \frac{a - b}{a + b} \quad (16)$$

where $q2$ is equal to 0 for a perfect circle and goes asymptotically to 1 for very long and thin objects. Parameters $q1$ and $q2$ are not sensitive to differences in shapes without any elongation (Fig. 5b). For such shapes the next shape factor, called circularity, can be defined:

$$q3 = 4\pi \frac{A}{L^2} \quad (17)$$

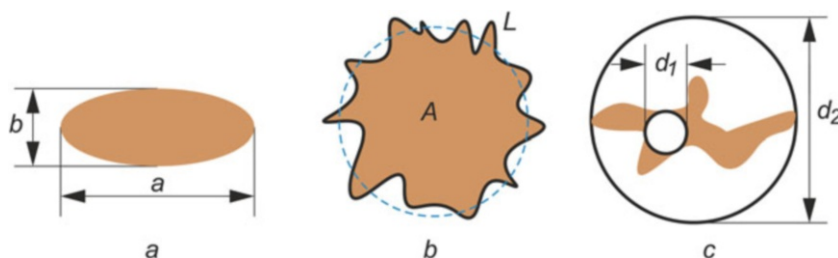
The use of the second power of the perimeter assures constant value for objects of the same shape but different sizes. Circularity reaches the value of 1 for a perfect circle and lower values for other shapes. Extremely irregular borders can lead to circularity values close to zero. This parameter is suitable for characterization of the shapes of islands or lakes.

One can meet also objects with shapes being a combination of elongation and irregularity of the border line. For characterization of such objects one can use the next shape factor, $q4$, which is a ratio of the circumscribed circle diameter d_2 to the inscribed circle diameter d_1 (Fig. 5c):

$$q4 = \frac{d_2}{d_1} \quad (18)$$

In practice a countless number of shape factors can be invented. However, the examples presented above are among the most popular.

Stereology, Fig. 5 Parameters for evaluation of shape factors (Source: Wojnar L (2020))



Sampling Strategy and Testing Objects of Any Shape or Arrangement

Stereology established sampling strategies which can be successfully applied in any research, including even the most modern tools connected with tomography and computerized image analysis. Sections (or images) for analysis can be taken in any way under assumption that the structure is IUR and the section covers enough large area. Definition how large this area should be is usually problematic. Therefore, sections taken either in a fully random way or systematic sampling design prior to experiments should be applied. Images taken by human observers without the above-mentioned restrictions often introduce significant bias to the analysis. This is a consequence of the fact that human observers intuitively choose places which look interesting, contain large objects, offers good quality, etc. Moreover, in linear analysis the test lines should be placed away in a distance which assures avoidance of multiple hitting the same object.

During image acquisition the fundamental problem is their resolution. It should be good enough for clear recognition of the smallest objects analyzed. These objects should contain several dozen pixels in order to be correctly classified. In the case of microscopic examination, the proper choice of magnification and, consequently, resolution is of the highest importance. Higher magnification allows for observation of smaller objects. Unfortunately, the images become less homogeneous and we lose information of large-scale arrangement of objects or even whole largest objects. Proper choice of microscopic magnification is always a compromise between details and global information.

One of the most common questions is: How many objects or images should be analyzed? There is no single rule for a proper choice of necessary number of observations. It is recommended to control the scatter of results. If the standard deviation of the results of measurements remains stable in spite of enlargement of the number of measurements, this number is sufficient.

Another very big problem is caused by structures which are not IUR. Independent, uniform, and random arrangement of objects can be found in ingenious rocks, like granite. In contrast, sedimentary rocks or various types of soil usually exhibit layered structure. Even more complicated structure of rocks, with curvilinear layers, is observed in the regions of massive rock formation movements. To make quantification even more complicated, tectonic faults can be taken under consideration.

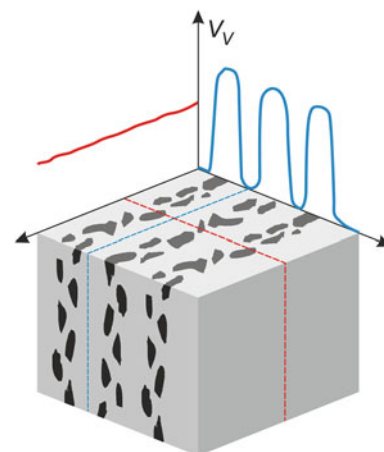
Sometimes finding appropriate solution is quite easy. In order to evaluate the total border length of elongated and oriented objects (like in Fig. 2b) it is sufficient to apply circular test lines (Fig. 4d) and Eq. (10) can be applied.

Another example of simple tricks which allow to avoid systematic errors is quantification of the volume fraction of

objects in layered structures (Fig. 6). If one chooses sections parallel to the layers (traces of cutting marked in blue), a saw-tooth wave will illustrate changes in estimated volume fraction as a function of section location (blue line in Fig. 6). It is sufficient to choose sections perpendicular to the layers (exemplary traces of such section are marked in red in Fig. 6) in order to get correct overall volume fraction of dark objects. Approximately, the same results will be obtained at any location of the section plane. Variation in the results will depend in this last case only on the section area and inhomogeneity of the structure examined. Sections perpendicular to the layers can be used also for estimation of the saw-tooth relation obtained from sections parallel to the layers. For estimation of this relation it is sufficient to evaluate volume fraction with the help of linear method with test lines parallel to the traces of layers. Another possible solution is to apply a set of equally distant points placed along a line parallel to the traces of layers and point counting method.

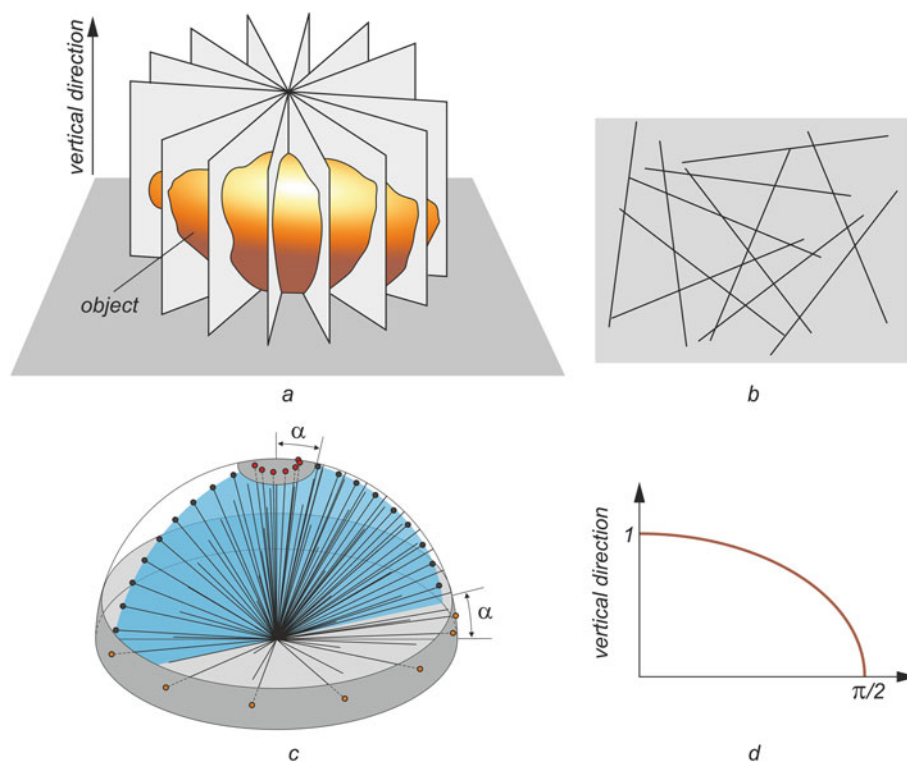
Specific surface area S_V is a very important stereological parameter. It can affect various properties and is one of the measures of size, because it increases when objects tend to divide into smaller segments. The method for estimating S_V on the basis of simple counts and linear method (Eq. (7)) was invented in the year 1945 by S. Saltykov. Unfortunately, this method works well only in the case of IUR objects. The scientific community has been waiting 42 years for generalization of this method, which allows for analysis of structures with objects of any shape and spatial arrangement (Baddeley et al. 1987).

Lest us analyze the method of vertical sections which allows for unbiased estimation of the surface area of objects of any shape. The result of analysis will be unbiased if the test lines (secants) are IUR, i.e., uniform, independent, and random. Intuitively it can be assured if we take a number of coaxial, uniformly distributed test planes (Fig. 7a). If we need



Stereology, Fig. 6 Evaluation of volume fraction in layered structures (Source: Wojnar L (2020))

Stereology, Fig. 7 Assessment of the surface area of an object of any shape: (a) object and vertical sections, (b) random test lines in a plane, (c) spatial distribution of test lines, and (d) exemplary cycloidal test line (Source: Wojnar L (2020))



to prepare physical sectioning, the object has to be put on a horizontal plane and the sections, as well as the reference axis, should be perpendicular to this plane. Consequently, they will be vertical. It explains the name: method of vertical sections, sometimes called a method of coaxial sections. The intuitive reasoning tells: if we have uniformly distributed vertical (coaxial) sections and the test lines in every section will be IUR (Fig. 7b), the sampling should be also IUR. Unfortunately, this not a case. If all the test lines fit a single point we can analyze their spatial distribution by analyzing distribution of the intersections with the surface of a hemisphere (Fig. 7c). Now, it becomes clear that the surface density of intersection points is much higher near the pole than close to the equator.

To solve this problem one should apply more secants parallel to the basic plane in comparison with the number of secants parallel to the vertical direction. And the number of secants with middle orientation should gradually change. Baddeley, Gundersen, and Cruz-Orive (1987) have invented a tricky and simple solution: a curvilinear test line which is a cycloid (Fig. 7d). During evaluation of the number of intersection PL with cycloids one should remember that the length of a cycloid of height 1 is equal to 2. Analysis of the vertical cross sections with the help of a set of cycloids allows for the direct use of Eq. (7) and assures unbiased estimation of the specific surface area.

Other Stereological Methods

The methods presented above belongs to a group of basic, most popular and widely used. There are also many other stereological tools but they are not presented in detail due to a limited volume of this chapter or limited applicability in geosciences.

Quantitative fractography covers a set of tools devoted to analysis of fracture surfaces. It is also based on analysis of sections and projections but characterizes distribution of several objects over curvilinear surfaces instead of the whole volume. These methods adapted general-purpose stereological relations and offer new parameters as well as relations suited to specific characteristics of fracture surfaces. Quantitative fractography has been extensively developed in the 1980s of the twentieth century. An overview available in ASM Handbook (Underwood and Banerji 2013) is in fact a copy of the chapter published in 1987. Some more advanced concepts have been presented by Wojnar and Kumosa (1990). Currently quantitative fractography is almost forgotten. This is a consequence of extensive use of confocal microscopy, stereoscopic methods, and, last of all, tomography. This last method is especially useful in the case of analysis of rocks being potential gas or oil deposits. Quantitative fractography could be possibly applied in analyses of land relief, glacier melting, etc.

Stereological methods are sometimes divided into two groups, namely model-based and design-based stereology (Stereology information 2020). Model-based stereology is recognized in this classification as a set of classical methods, currently a little bit outdated. Model-based methods work well only if the models truly represent the real objects. In other words, these methods are not assumption free and their applicability is limited by these assumptions.

Typical examples of model-based procedures are the methods for evaluation of particle size distributions. Among them one can list Saltykov (area analysis), Spector (chord analysis), or Lord and Willis (graphical analysis of chords). All these methods are restricted to a model shape of sphere (Underwood 1970). All these methods have the same drawback – the total number of objects and final shape of the distribution function are computed on the basis of analysis of smallest particles in experimental distribution function. Unfortunately, measurements or counts within this class of objects are coupled with the highest bias. The results can be corrected during further analysis but it requires additional assumptions.

Design-based stereology is presented as a more universal and powerful set of methods. Its main property is a lack of assumptions concerning shape, size, orientation, and distribution. Nevertheless, this division is rather artificial and not precise. Classical stereological methods were derived with only one, but relatively strong assumption: the objects analyzed should be IUR (independent, uniform, and random). If any structure investigated does not fulfill this assumption it can be analyzed if only the test system assures independent, uniform, and random sampling and measurements. The method of vertical section is a very good example. Consequently, even the oldest stereological methods can be used within the framework of design-based stereology.

One of the most difficult tasks is evaluation of the number of objects in the tested volume (N_V). This parameter can be easily estimated using CT methods. Stereological analysis requires assumptions concerning shape and is usually very approximate. The problem can be solved by the use of disector, i.e., analysis of two parallel sections, called reference and lookup, respectively. The rule of counting is very simple: objects visible only in the reference section are counted as related to the test volume between sections. The main restriction of this method is the distance between test sections. It has to be smaller than the possible smallest object size. Consequently, this method is successful in analysis of cells or cell nuclei. Unfortunately, it is practically useless in inspection of minerals, rocks, or soils where very small objects are often observed.

The next example of design-based stereology is surfactor, a method for estimation of the surface area. This method

requires isotropic sectioning. It is also applicable mainly for biological structures and in the case of rocks or soils, the use of classical or vertical sections method will be a better choice.

Another subset of stereological methods is known under the name of second-order stereology (Stereology information 2020). These techniques are devoted to relationships, for example, the distribution of particles. One should remember that these methods have been initially designed for biological research and possible applications of second-order stereology methods in geosciences is limited. Therefore, and also due to the limited space, second-order stereology is not discussed in this chapter.

Stereology in Computer Era

The unprecedented progress in computer technology enabled invention or significant improvement in numerous research methods suitable for three-dimensional analysis of various structures. Below there are listed some examples:

- CT (computed tomography) offers three-dimensional (3D) reconstruction and solves the problem of evaluation of spatial arrangement and connectivity, which cannot be characterized using stereological methods
- EBSD (electron backscatter diffraction) used in scanning electron microscopy (SEM) allows for precise evaluation of shape and orientation of single grains in polycrystalline materials
- FIB-SEM (focused ion beam) enables 3D reconstruction of micro volumes
- Confocal microscopy, which allows for quick analysis of curved surfaces without overlaps or 3D analysis of transparent tissues (in microscale) or
- Seismic tomography (suitable for a very-large-scale research)

Consequently, the knowledge of stereology is often neglected.

Simultaneously, 2D imaging still is and probably will be in the near future very popular in any research. Such situation is clearly visible in medicine. In spite of the progress in CT, MRI (magnetic resonance imaging) or PET (positron emission tomography) techniques, simple X-ray, or USG images are commonly used in renowned hospitals and the quality of these 2D techniques is gradually improved. Outside medicine an enormous flood of images comes from countless number of several cameras. Almost all these images are processed by computers and part of them is next analyzed. This analysis is performed usually with the use of computerized tools, mainly image analysis (IA) and artificial intelligence (AI).

Classical stereological analysis with manual counting or measurements is extremely laborious and time consuming, while IA and AI are very convenient as they are fast and give repeatable results. Consequently, these new methods are recognized much more attractive than stereology. There is, however, a significant risk of large errors introduced by image analysis tools. Small errors in object detection can significantly affect the results obtained and the measurement errors can be larger than in the case of manual stereological analysis. Moreover, even exact and precise results from a single image not necessarily reflect the scatter of the whole structure. And stereological tools can help to establish appropriate sampling strategy or image resolution. Further interpretation of raw data obtained from automatic measurements also is a domain of stereology. Some algorithms used in quantification in image analysis has a stereological background, for example, the Crofton formula for perimeter evaluation and techniques for unbiased evaluation of several parameters. Manual methods can be irreplaceable in the case of structures not suitable for automatic detection. Many IA software packages offer stereological plug-ins, which combine manual detection with automated measurements or calculations.

So, after decades of parallel and even concurrent development, one observes in recent years some synergy. Adding the term image analysis to the name of the International Society for Stereology has a symbolic significance. Computer technology can enormously increase effectiveness of stereological analysis and stereology can help to avoid rough errors connected with uncritical application of easy-to-use computerized tools.

Applications in Geosciences

An overview of potential development, problems, and limitations in application of stereology in geosciences has been presented many years ago by Bodziony (1993). Surprisingly, the majority of his conclusions is still valid. He stressed the role of stereological analysis in mineral liberation and quantitative evaluation of the structure of rocks and soils. Finally, Bodziony has predicted the dominating role of automatic image analysis and the need to develop image acquisition methods which ensure good contrast in the case of transparent minerals. Due to the development of new tools and progress in traditional ones one can currently analyze directly 3D data or evaluate much more complex objects than porosity only in the soil structure.

Jutzeler et al. (2012) concentrated on grain size evaluation in pyroclastic deposits created during volcanic eruptions. They use stereological analysis as an alternative to

sieving and 3D imaging. Their method of analysis, called functional stereology, is based on simulation of 2D size distribution on the basis of real or model 3D size distribution. Next, experimental data is compared to the simulated 2D distribution which allows for matching appropriate 3D size distribution. This method was validated on synthetic rock with precisely defined grain size distribution. Application of stereological analysis allows inexpensive interpretation of data obtained from image analysis and results more accurate than sieving.

An interesting concept to combine stereological methods and modern data acquisition from terrestrial lidar was presented for forest inventories (Lynch et al. 2018). Estimators of the single tree bole as well as volume and height characteristics are based on stereological principles and appropriate sampling. Forestry has not been a part of geosciences, but forests are an extremely important part of biosphere and can have some impact on classical geosciences.

In recent years stereological methods are rarely used as a main tool in data interpretation. Nevertheless, the knowledge of basic principles of stereology can play important role in correct design of the experiments, sampling, and data interpretation.

Conclusions

Classical stereological methods have been initially developed for manual analysis. In spite of this the rules and methods of stereology are suitable also for contemporary modern computerized methods of image analysis.

Possible applications of stereological methods in geosciences cover several areas, for example:

- Classical stereological analysis, i.e., evaluation of 3D parameters from 2D (two-dimensional) images
- Analysis of 2D images. Stereological methods can be helpful in unbiased estimation of the number of objects, size distributions, or shape factors
- Analysis of curvilinear surfaces, for example, fractures or topography of the Earth. Appropriate tools constitute quantitative fractography
- Evaluation of quantitative parameters of 3D objects detected using tomography or systematic sectioning
- Design of experiments, including orientation of possible sections or determination of the necessary number of objects to be analyzed, etc.

The use of stereological methods in analysis of digital images can significantly improve the results or help to

avoid erroneous quantification and therefore should be commonly applied.

Cross-References

- [Data Acquisition](#)
- [Grain Size Analysis](#)
- [Mathematical Morphology](#)
- [Morphometry](#)
- [Object Boundary Analysis](#)
- [Object-Based Image Analysis](#)
- [Point Pattern Statistics](#)
- [Porosity](#)
- [Quality Control](#)
- [Quantitative Stratigraphy](#)
- [Shape](#)
- [Stochastic Geometry in the Geosciences](#)

Bibliography

- Baddeley AJ, Gundersen HJG, Cruz-Orive LM (1987) Estimation of surface area from vertical sections. *J Microsc* 142:259–275
- Bodziony J (1993) Stereology in geosciences: achievements, difficulties and limitations. *Acta Stereol* 12(2):211–222
- Hilliard JE, Lawson LR (2003) *Stereology and stochastic geometry*. Kluwer Academic Publishers, Dordrecht/Boston/London
- ISSIA (2020) International Society for Stereology and Image Analysis. <https://www.issia.net/>. Accessed 3 Jun 2020
- Jutzeler M, Proussevitch AA, Allen SR (2012) Grain-size distribution of volcanoclastic rocks 1: a new technique based on functional stereology. *J Volcanol Geotherm Res* 239–240:1–11. <https://doi.org/10.1016/j.jvolgeores.2012.05.013>
- Lynch TB, Ståhl G, Gove JH (2018) Use of stereology in forest inventories – a brief history and prospects for the future. *Forests* 9: 251–277. <https://doi.org/10.3390/f9050251>
- Peterson DA (2020) Stereology. In: Kompoliti K, Metman LV (eds) *The encyclopedia of movement disorders*. Academic Press, 2010. <https://www.sciencedirect.com/topics/medicine-and-dentistry/stereology>. Accessed 3 Jun 2020
- Shape (2020) <https://en.wikipedia.org/wiki/Shape>. Accessed 3 Jun 2020
- Stereology (2020a) <https://en.wikipedia.org/wiki/Stereology>. Accessed 3 Jun 2020
- Stereology (2020b) <https://www.merriam-webster.com/dictionary/stereology>. Accessed 3 Jun 2020
- Stereology information for the biological sciences (2020) <http://www.stereology.info>. Accessed 3 Jun 2020
- Underwood EE (1970) *Quantitative stereology*. Addison-Wesley, Reading
- Underwood EE (1987) Stereologia – the first newsletter of the I.S.S. *Acta Stereol* 6(Suppl II):215–267
- Underwood EE, Banerji K (2013) Quantitative fractography. In: *ASM handbook, Fractography*, Seventh printing, vol 12. ASM International, Materials Park, pp 193–210
- Weibel ER (1987) Ideas and tools: the invention and development of stereology. *Acta Stereol* 6(Suppl II):23–33 ISS
- Wojnar L (2020) Analiza obrazu. Jak to działa. Politechnika Krakowska, Krakow
- Wojnar L, Kumosa M (1990) Quantitative analysis of overlapped fracture surfaces. *Eng Fract Mech* 4:597–609

Stochastic Geometry in the Geosciences

Dietrich Stoyan

Institut für Stochastik, TU Bergakademie Freiberg, Freiberg, Germany

Definition

Stochastic geometry is the branch of applied probability that studies random systems of geometrical objects scattered in 3D or 2D space. The theory considers various stochastic models, which may be cell- or pixel-based as well as object-based, and tries to determine the characteristics of these models. Modern geostatistics use successfully such model classes of stochastic geometry as marked point processes or Boolean models, and general ideas of stochastic geometry play a role in modeling and simulating geological structures. The corresponding statistical aspects (data analysis, model fitting) are treated by methods of spatial statistics and multiple-point statistics.

Overview

In geoscientific applications the geometrical objects may be of four types:

- Points as considered in the *article* ► [“Point Pattern Statistics”](#)
- Line segments, curves (1D objects), e.g., river courses/networks, centerlines of channels in fluvial reservoirs, fractures/faults seen on the Earth’s surface or in outcrops, lineaments, striations. . .
- Surface or plane pieces (2D objects), e.g., fractures, faults, joints, veins, geological horizons and boundaries, intrusion-host contacts. . .
- Bodies (3D objects), e.g., tessellation cells, specific geological bodies, grains in sedimentary, magmatic and metamorphic rocks, lobes, shales, fluvial channels. . .

These objects may range from microscopic to plate-tectonic scale.

The corresponding mathematical terms are:

- Fiber processes
- Surface processes
- Full-dimensional random sets
- Tessellations

The use of the term “process” is a mathematical convention; some readers may prefer instead the term “field”. The

objects are sets (in mathematical sense) in a 2D or 3D reference frame, usually a geographic coordinate system.

The standard reference to applied stochastic geometry in general is Chiu et al. (2013). All formulas given without reference in this text can be found in this book. Stochastic geometry has operated from the very beginning in the spirit of multiple-point statistics. It was never limited to second-order or two-point methods.

A basic difficulty of application of stochastic geometry in the geosciences consists in the fact that many of its models assume that the structures considered are stationary. That means that they (theoretically) are thought to be extendable infinitely into space and that their random fluctuations are similar everywhere in space. While this assumption is often realistic at the microscopic scale, it is often not in studies of larger geoscientific objects. Therefore, when the standard models of stochastic geometry are used, care is necessary. If, as in many situations, short-range relationships are of main interest, a clever choice of the window of observation may help; see the discussion in the article ► “Point Pattern Statistics”.

Furthermore, it is a problem that stochastic geometry has been developed as a mathematical discipline with main applications in image analysis and materials science, independently of ideas of geosciences. This has led to different terminologies. The present text follows mainly the rules of geomathematics, with some notes on the mathematical terminology.

The following presents some fundamental models of the four model classes. These models aim to help in understanding the geological/mineralogical processes of forming the observed structures and to get information about the interaction of their components, but are also the basis for characterization and simulations.

Models of Stochastic Geometry

Marked Point Processes

Marked point processes (MPP) serve as a basis for many models of stochastic geometry in the geosciences, as they describe random collectives of objects scattered in space. Formally, an MPP is a double sequence $\{[x_n, M_n]\}$ of points x_n and marks M_n . The x_n describe the locations of the objects, being constructed by the statistician (e.g., centers), while the M_n are the objects, mathematically “sets”. In the mathematical literature such an MPP is called a “germ-grain process”, germ = point and grain = mark; the set-theoretic union of all grains at their germs is called “germ-grain model”. In the geostatistical literature the term “object model” is used. Pyrcz and Deutsch (2014), section 4.4, is an excellent introduction to the statistical problems related to MPP in the geosciences.

Stationarity of MPP means invariance of the distribution under translations, which shift the points but retain the marks, i.e., $[x, M] \rightarrow [x + \delta, M]$.

The intensity λ of a stationary MPP is the mean number of points per area or volume unit. The marks are characterized by a *mark distribution*, which is a distribution of a random set, or, if the objects are parameterized, of a number or vector. Whenever possible, one tries to parameterize the marks. Simple models are balls, discs, or segments, and there are models for other objects such as channels.

Usually, there are spatial correlations between the germs and grains. A typical case is that in regions of high point density the grains tend to be small. So it is often necessary to model and simulate the point pattern $\{x_n\}$ and the marks M_n simultaneously, which makes simulations rather complicated in practice.

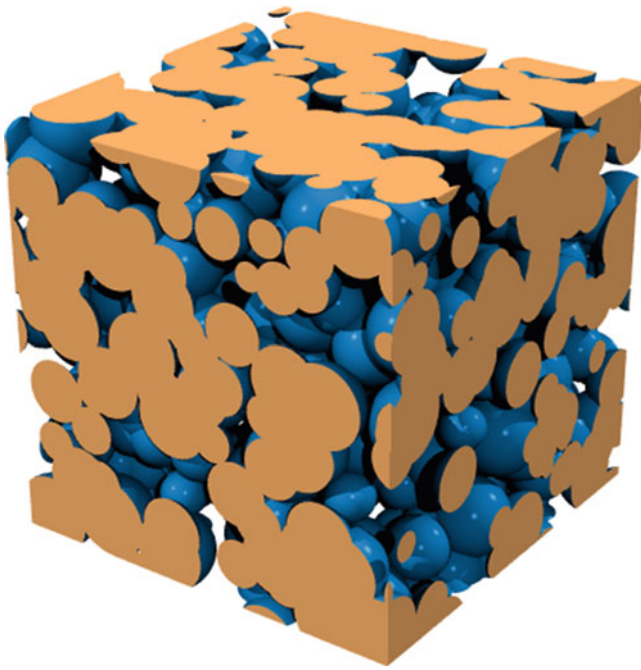
The Poisson Point Process

The Poisson point process is an important building stone of many models of stochastic geometry. Details are explained in the article ► “Point Pattern Statistics”. For the understanding here it may be sufficient to know that this point process is stationary and isotropic and completely spatially random (CSR), there is full independence and randomness of the point distribution. Its distribution is fully determined by the intensity λ : the random number $N(A)$ of points in a set A has a Poisson distribution with mean $\lambda v(A)$, where $v(A)$ is the area or volume of A .

The Poisson-Boolean Model

The Poisson-Boolean model, abbreviated here as PB model, is a universal and key model of stochastic geometry, which can be applied for modeling fiber processes, surface processes as well as structures of 3D bodies. It is a special object or germ-grain model, for two categories or phases. In the mathematical literature, the Poisson-Boolean model is called simply “Boolean model”, while in the geomathematical literature every model composed by grains or objects is called “Boolean”. The name was coined within the school of Georges Matheron in the context of indicator functions and random sets and refers to the English mathematician George Boole, who is for mathematicians closely related to “either-or” or “0–1-structures” (see the relationship between random set X and random field $\{Z(x)\}$ below).

The model construction is very simple. It starts with a homogeneous Poisson point process $\{x_n\}$ of intensity λ , the set of “germs”. Each germ is the center of an object or “grain”, a bounded set, which can be random. All the grains have the same distribution and are stochastically independent, which makes analysis and simulation easy. The set-theoretic union X of all the grains $M_n + x_n$ is called the Poisson-Boolean model.



Stochastic Geometry in the Geosciences, Fig. 1 A simulated PB model. In a cube and in its neighborhood random points are scattered. Each point is the center of a ball of random radius. The set-theoretic union of all balls and ball-pieces in the cube is a sample of a PB model. (Data courtesy of Felix Ballani)

A simple example is the case where the grains are balls with random radii. Figure 1 shows a piece of a simulated sample of a PB model with spherical grains.

The term “germ” suggests a relationship to some growth model. Indeed, the germs may be starting points of growth and with increasing time the grains may grow. The corresponding formula for the time-dependent volume fraction $V_V(t)$ of the space occupied by the grains is named the *Johnson–Mehl–Avrami–Kolmogorov crystal-growth formula*.

For the PB model there exists an elaborate mathematical theory (Chiu et al. 2013). It is stationary and, if the grain distribution is rotation-invariant, isotropic. For many distributional summary characteristics, formulas are available in the case of convex grains. Some examples are given here for the three-dimensional case:

- Volume fraction V_V = part of space covered by the grains:

$$V_V = 1 - \exp(-\lambda \bar{V}), \quad (1)$$

where $\bar{V}$ is the mean grain volume.

- Covariance $C(r)$ = probability that two points of inter-point distance r both lie within the grains of the model, in the isotropic case:

$$C(r) = 1 - \exp(-\lambda \bar{\gamma}(r)) \text{ for } r \geq 0, \quad (2)$$

where $\bar{\gamma}(r)$ is the so-called set-covariance of the grains, a statistical grain characteristic for which formulas exist.

- Spherical contact distribution function $H_s(r)$ = distribution function of the distance from a random test point outside the grains to the boundary of the PB model, in the case of spherical grains:

$$H_s(r) = 1 - \exp\left(-\frac{4}{3}\lambda r\left(3\bar{R}^2 + 3r\bar{R} + r^2\right)\right) \text{ for } r \geq 0, \quad (3)$$

where $\bar{R}$ and $\bar{R}^2$ are the first two moments of the sphere radius R .

When a line intersects a PB model, it creates a series of intervals, called chords, which are alternating within and outside of the set X . Their random lengths are independent, with distribution depending on the grain distribution for “in” and exponential distribution for “out”. The intersection of a PB model with a plane/line is a 2D/1D PB model.

In small-scale applications, e.g., of porous media, the PB model is used often for producing more or less realistic geometrical patterns, and the grain shapes and the Poisson process intensity are chosen only from the standpoint of the final result, the statistical properties of the resulting random-set pattern. Typical three-dimensional grains are balls and the so-called Poisson-polyhedra, i.e., polyhedra generated by the Poisson plane process. In this sense the porous structure of sandstones was modeled (Arns et al. 2002); of course, a PB model cannot mimic sandstone formation.

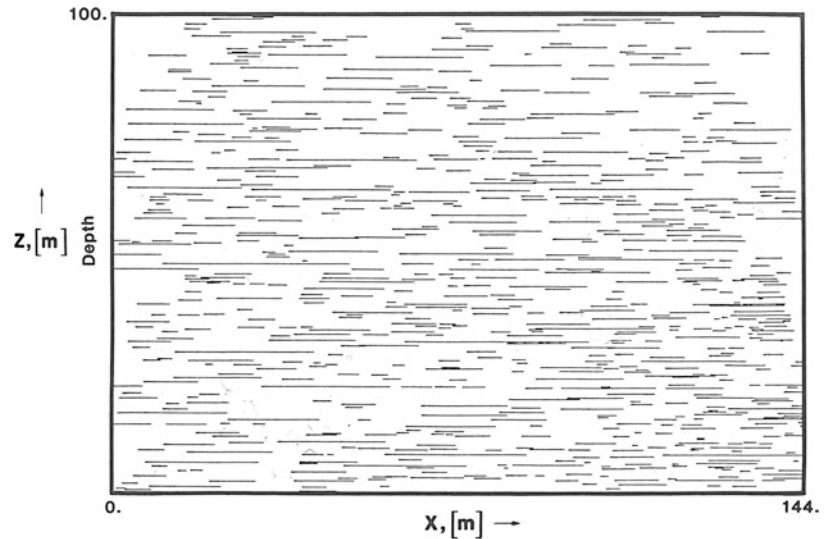
Statistical methods, including goodness-of-fit tests, for the PB model are presented in Chiu et al. (2013). Statistics is difficult, if only the union set X is given as in Fig. 1 and not the germs and grains. Figure 2 shows the Assakao sandstone data set as in Haldorsen and Chang (1986). It looks like a PB model with grains = shale streaks (considered as horizontal segments of random lengths); as germs the segment centers may be used. However, the germs show local alignments (Haldorsen and Chang 1986) and the streak lengths are spatially correlated, in regions of high streak density the lengths tend to be short, so the hypothesis that this structure belongs to a PB model must be rejected, without a formal test.

Fiber Processes

Fiber processes are mathematical models for random systems of curve pieces in 2D or 3D, if in 2D, also called “networks.” There exists some theory for such systems in the case of stationarity (Chiu et al. 2013). The corresponding mean-value parameter is L_A (in 2D) or L_V (in 3D), the mean length of fiber pieces per unit area or volume. The orientations of the fibers are described by a rose of directions. In the stationary

Stochastic Geometry in the Geosciences, Fig. 2

The Assakao Sandstone data set. The data are from Haldorsen and Chang (1986). The formation is generally sandstone (white) with occasional shale streaks (black). This cannot be seen as a sample of a *Poisson* Boolean model, since the streaks tend to form clusters if being of short length. The pattern can also be interpreted as a line-segment process



and isotropic case the spatial length density L_V can be determined by stereology. Spatial correlations in a fiber process can be described by suitable second-order characteristics (Chiu et al. 2013).

On a large scale, fiber systems appear in the planar case as fracture networks at the surface of Earth or at rock outcrops. In some cases it is important to give the fibers a random thickness.

A simple model of a fiber process (both in 2D and 3D) is the *PB segment process*. This is a PB model where the grains are line segments of random length and orientation, while the germs stand for nuclei. The corresponding PB model is a system of segments, randomly distributed in the plane or in the space. It satisfies $L_V = \lambda \cdot \bar{l}$, where $\bar{l}$ is the mean segment length. The spherical contact distribution function in 3D is

$$H_s(r) = 1 - \exp\left(-\lambda\pi r^2\left(\bar{l} + \frac{4r}{3}\right)\right) \text{ for } r \geq 0.$$

This model serves as a starting point for the construction of more realistic models, by making the fibers sinuous, thick, and interacting. There are models of fracture networks with nucleation, growth, and arresting (Chiu et al. 2013; Davy et al. 2013), where differential equations or the Weibull fracture model of materials engineering introduce the temporal element, which lead to power laws.

Surface Processes

Surface processes are mathematical models for systems of 2D sheets in 3D space. Typical cases are joint systems or systems of boundaries between distinct geological facies or discrete fracture networks (DFN). Wellmann and Caumon (2018) is an excellent review on geological subsurface detection and modeling.

There exists some mathematical theory for such systems in the case of stationarity (Chiu et al. 2013). The corresponding mean-value parameter is S_V , the mean (surface) area per unit volume. The directionality of the surfaces is described by a rose of orientations of surface normals. In the stationary and isotropic case S_V can be determined by Eq. (9). Spatial correlations in a surface process can be described by suitable second-order characteristics (Chiu et al. 2013).

A simple model of a surface process is the *PB disc process*. This is a PB model where the grains are discs of random size and orientation randomly scattered in the space; the germs may stand for faults. It implies $S_V = \lambda \cdot \bar{S}$; where $\bar{S}$ is the mean disc area. The spherical contact distribution function is

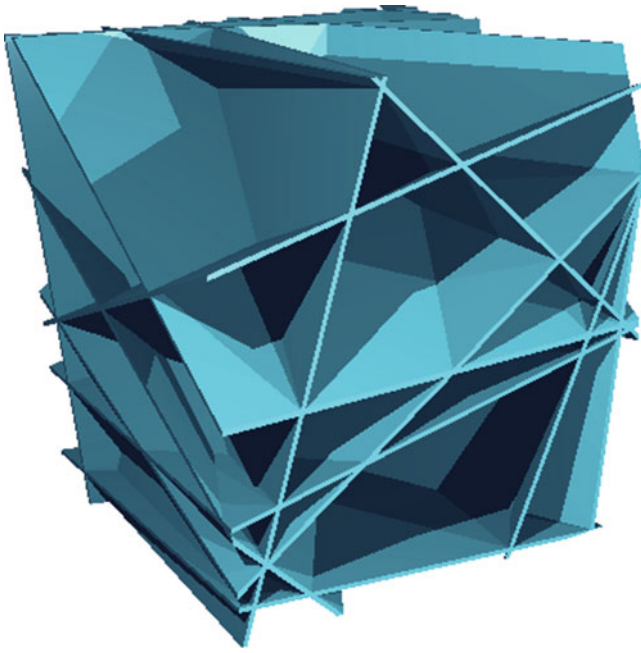
$$H_s(r) = 1 - \exp\left(-\lambda\pi r\left(2\bar{R}^2 + \pi r\bar{R} + \frac{4r^2}{3}\right)\right) \text{ for } r \geq 0,$$

where $\bar{R}$ and $\bar{R}^2$ are the first two moments of disc radius. The parameter of the exponential distribution of chord lengths on lines intersecting the disc system is $\lambda\pi\bar{R}^2/2$. This simple model was used for modeling DFNs; it was refined, for example, in (Bonneau et al. 2016) by sequential seeding further discs according to a starting PB disc process following rules that take into account the ages of discs.

A further model is the *Poisson plane process*, a random system of planes in 3D; see Fig. 3. This is not a system of germs and grains (= planes), but it is defined as a Poisson point process in a 3D representation space. This highly idealized model serves sometimes as building stone for more complicated models.

The intersection with a line yields a linear Poisson process of intensity $P_L = S_V/2$, the number of planes hitting a fixed convex set has a Poisson distribution.

The planes of a Poisson plane process generate a tessellation of the space in convex polyhedral cells, the probability



Stochastic Geometry in the Geosciences, Fig. 3 Poisson plane process. Simulated sample of a Poisson plane process. (Data courtesy of Felix Ballani)

distribution of which constitutes the so-called *Poisson polyhedron*, which is a model for crystalline grains in object models.

Modeling and Simulating DFNs

Discrete fracture networks are studied in geology for understanding tectonics and geological structures and in engineering for characterizing hydraulic and mechanical properties of rock mass in the context of studies of groundwater or hydrocarbon flow and of storage constructions. For this, plausible 3D models are indispensable.

Here three approaches for unconditional simulation of fault networks are mentioned, which are quite different but nevertheless two of them go under the same name GEOFRAC: the agglomerative approach by Koike et al. (2015) and the disruptive approach by Ivanova et al. (2014). In both approaches a single fracture or fault is a planar convex set, not a disc. Of course, the actual shape of fractures is more complicated: fault surfaces are rough surfaces that may be modeled as random fields, as also aperture or void height. Often surfaces are represented as polygonal surfaces, which can be modified by perturbing its nodes. The third, a sequential approach is described in the papers Cherpeau et al. (2010) and Cherpeau and Caumon (2015).

The Sequential Approach The algorithm sequentially simulates faults in the domain of interest. The faults are grouped into fault families according to regional tectonic knowledge and ranked chronologically. Each fault is modeled as a finite

level surface, which allows for truncation in the course of the simulation. The fault centers are randomly distributed, partly related to older and previously simulated faults; the faults are characterized by orientation, size, and sinuosity parameters (Cherpeau et al. 2010). The paper (Cherpeau and Caumon 2015) contains a flow chart that shows details of the algorithm. Important is a step which improves the initial geometry by various filtering operations. Also information from fault traces from seismic surveys or outcrops is used. Figure 4 shows some simulated fault families.

The Agglomerative Approach The model has as starting point a nonstationary PB model, the germs of which form an inhomogeneous Poisson process with intensity function piecewise constant in 3D grid cells, determined by data. The grains are discs of equal size but of spatially correlated orientation, also its distribution comes from data. These discs are connected by some rules to larger objects that are simplified geometrically by replacing them by flat figures (given by mean strike and dip) and forming the convex hull, which yields the faults.

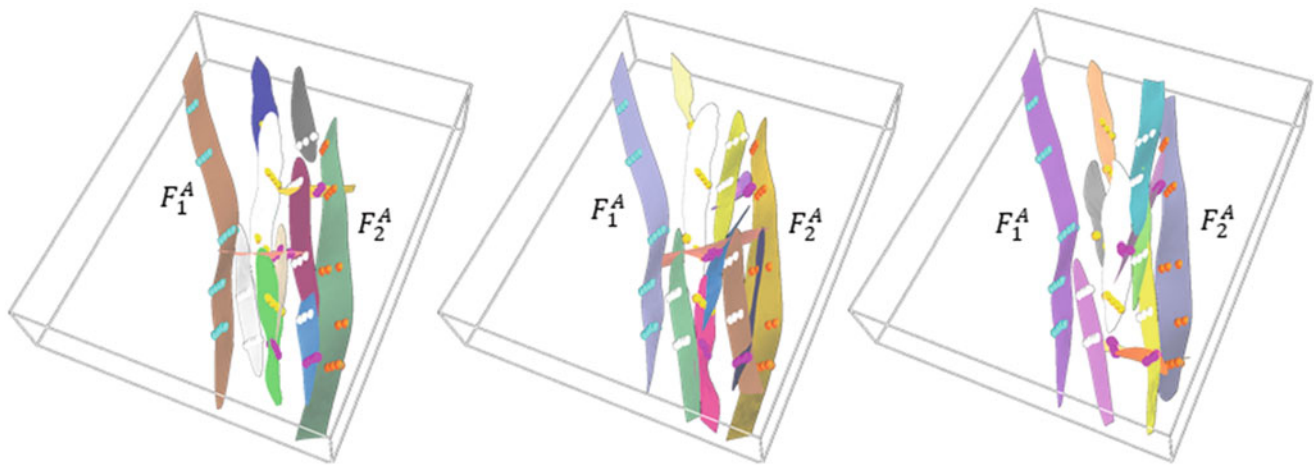
The Disruptive Approach The model has as starting point a Poisson plane process with a specified orientation distribution and area density S_V , which is also the final area density of the fracture system. On all planes of the process, points of independent Poisson processes of the same intensity ρ are scattered and the corresponding 2D Voronoi tessellations are constructed. The polygonal cells of these tessellations are the faults of the final fracture system, after changing their locations and orientations by random translations and rotations, which make the polygons to mutually intersect.

Full-Dimensional Random Sets

General

All geometric two-phase structures considered in this entry are random sets. This section focuses on sets of full spatial dimension, i.e., sets consisting of planar components if in 2D and of spatial components such as bodies if in 3D. Additionally, it is assumed that these structures, which are denoted by X , are stationary and isotropic. They have a positive *area* (if 2D) or *volume fraction* (if 3D) A_A or V_V , respectively, for which here the general symbol p is used.

Such random sets are described by various statistical summary characteristics, which are already explained for the PB model. The simplest is p = part of space covered by X = probability that a random test point hits X . The second-order behavior is characterized by the *covariance* $C(r)$, the probability that both members of a test point pair of inter-point distance r hit X . These characteristics belong to a traditional geostatistical view of random sets, which results from the fact that each random set X can be interpreted as a random field



Stochastic Geometry in the Geosciences, Fig. 4 Three fault families. Three simulated samples of fault systems by the sequential approach (Cherpeau et al. 2010)

$\{Z(x)\}$ by using the indicator function $1_X(x) = 1$ if $x \in X$ and $= 0$ otherwise and setting $Z(x) = 1_X(x)$. The mean of this field is p and the variogram is

$$\gamma_z(r) = p - C(r) \text{ for } r \geq 0. \quad (4)$$

However, there are many further summary characteristics of random sets, which are beyond geostatistical two-point thinking, sometimes used in the context of “multiple-point statistics.” For example, the empty space between components of X is characterized by the so-called *contact distribution functions*, e.g., the *spherical* $H_s(r)$ as explained for the PB model. A similar characteristic is the *linear* contact distribution function $H_l(r)$, which is the distribution of the random distance from a test point outside of X to X , measured in random direction. This $H_l(r)$ is closely related to the *chord length distribution function*. Some of these functions also make sense for point, fiber, and surface processes. Elements of statistics for random sets are presented in Chiu et al. (2013).

Statistics for single objects or grains uses typically parameterized models developed for special cases. Examples are channels (Hauge et al. 2017; Rongier et al. 2017), lobes (Pyrz and Deutsch 2014), and faults as in DFNs.

Three Model Classes for Full-Dimensional Random Sets

1. Pixel models, i.e., sets constructed by cells or pixels of two colors, where all pixels of the same color form the random set
2. Object models or germ-grain models
3. Sets constructed from random fields

A class of *pixel models* are Markov random fields, which can be equivalently seen as Gibbs fields, which are described for

the bivariate (black-white) case in Chiu et al. (2013), they are called “Gibbs discrete random sets”. The distribution of such a set is given by

$$P(X = \xi) = \frac{1}{Z} \exp(-\beta U(\xi)), \text{ where } \xi \text{ is a black pixel set,} \quad (5)$$

where β is a positive parameter called inverse temperature, and $U(\xi)$ is the “energy” of ξ and Z a normalizing constant. Pixel sets with high energy appear rarely, and so by choosing the energy the resulting random black pixel patterns can be controlled. Generalization to the multivariate case is possible.

Object models or germ-grain models are special MPP where the marks are 3D geometrical objects, i.e., the models are collections of objects of random shape and random locations. The object centers are the germs and object bodies the grains.

A simple special case is the PB model. While here simple overlaps or intersections of the grains are possible, often in geoscientific applications the grains interact spatially, do not simply intersect, or are even definable. Therefore, object models relevant for the geosciences are rather complicated, with various mutual spatial relationships between the objects, and are usually mainly built for simulation. In these simulations, the germs play only a secondary role, which justifies the name “object model”; one cannot simulate first the germs and then add the grains. A particular case of object models are random packings of hard objects, as given by heaps of sand grains or pebbles. Also these systems are typically studied by simulation, with methods that guarantee a high volume fraction V_V (Chiu et al. 2013).

The derivation of mathematical formulas for model characteristics of germ-grain models must be seen as hopeless for recent mathematics. Only the equation for volume fraction p of a stationary germ-grain model with non-intersecting grains is simple:

$$p = \lambda \bar{V}, \quad (6)$$

where $\bar{V}$ is the mean grain volume. Sometimes the chord lengths outside the objects follow approximately exponential distributions, also for models beyond the PB model.

Statistical analyses try to estimate the intensity λ of the germ point process (with germ = center) and to determine information of the grains, perhaps mean volume, other size characteristics, and, in the best case, the mark distribution.

Excursion Sets (or level sets) are a “continuous alternative” to object models; they yield models with smooth boundaries. Such a set is constructed starting from a random field $\{Z(x)\}$ by truncating, i.e., by setting

$$X = \{x : Z(x) \geq u\}, \quad (7)$$

i.e., as the set of all points of the space where the field $\{Z(x)\}$ is larger than a threshold u . Thus, excursion sets appear in the context of indicator kriging. Often the case is considered where the random field is Gaussian and homogeneous and isotropic with mean μ , variance σ^2 , and covariance function $k(r)$ ($k(0) = \sigma^2$).

Then the area or volume fraction of X are given by $p = 1 - \Phi\left(\frac{u-\mu}{\sigma}\right)$, where $\Phi(\cdot)$ is the standard normal distribution function. For the covariance $C(r)$ there exists an integral formula, which connects it with $k(r)$.

Excursion sets can be easily simulated if $\{Z(x)\}$ is given, but their statistics, i.e., the determination of the latent random field $\{Z(x)\}$, is complicated if only samples of the random set X are available (Chiu et al. 2013). The whole theory presented above can be generalized to the case of multi-phase structures.

Representative Volume Elements

In the context of statistical analyses and physical calculations (e.g., determination of permeability or mechanical properties) a big problem is the determination of the necessary window of observation and calculation, for obtaining reliable mean values. (Windows too small yield imprecise values, windows too large cause too much work.) Using a terminology that first appeared in the context of homogenization, one speaks of representative volume elements (RVE). The determination of RVEs is sketched in (Chiu et al. 2013). It is important to note that for the same material the RVEs may differ if determined for different spatial characteristics as, for example, porosity and permeability.

Random Tessellations

General

A *tessellation* or *mosaic* is a division of the plane or of the space into cells. Such patterns occur often in the geosciences, for example, in the context of polycrystalline microstructures and crack patterns, such as columnar joint patterns in basalt or other rocks or mud cracks. There is a vast literature, use Chiu et al. (2013) as starting point, where statistical methods are also presented.

Two important models of tessellations are the Voronoi and Delaunay tessellations, which are both constructed relative to a given point pattern or process.

The Voronoi Tessellation

For the cells of a Voronoi tessellation (VT), a pattern of points called generators, nuclei, or seeds is the starting point: All points of the plane/space closer to a generator than to all other generators form its cell. These cells are convex sets, which have in the planar case the mean vertex or side number 6, if the generators form a stationary (non-lattice) point process.

A VT can be also interpreted as the result of a growth process, which explains partly its use as a model in the geosciences. The generators play the role of nuclei, in which growth begins at the same instant. The speed of growth is uniform at all nuclei and uniform in all directions, such that the nuclei become discs/balls if empty space is available. When two discs/balls come in contact, growth stops at the contact point and an edge/face begins to develop, while in other places growth is continued. Finally, the whole plane/space is divided into convex cells.

This model helps to understand and to describe natural processes, for example, the formation of sorted patterned grounds in polar or alpine environments or the process of growth of crystals in aggregates of crystals within cavities, where the growth process also goes into the third dimension, orthogonal to the plane on which the crystals grow. The cells of VTs are used as model building stones for more complicated models, as in the disruptive approach of GEOFRAC. There is software for the construction of VT both in 2D and 3D.

For the VT there exist formulas for important characteristics in the case where the generators form a homogeneous Poisson process of intensity ρ (Chiu et al. 2013): the mean of cell area is ρ^{-1} and its variance is $1.280 \rho^{-2}$. The mean number of vertex points per area unit is 2ρ , the mean cell edge density (length per unit area) is $L_A = 2\sqrt{\rho}$.

The VT is frequently used as a statistical tool, in particular for interpolation in the case of irregularly distributed points of measurement (Chiu et al. 2013). Merland et al. (2014) consider VTs in the context of constructing simulation grids, where the tessellation cells are grid cells. The generator positions are optimized so that the cell facets honor geological

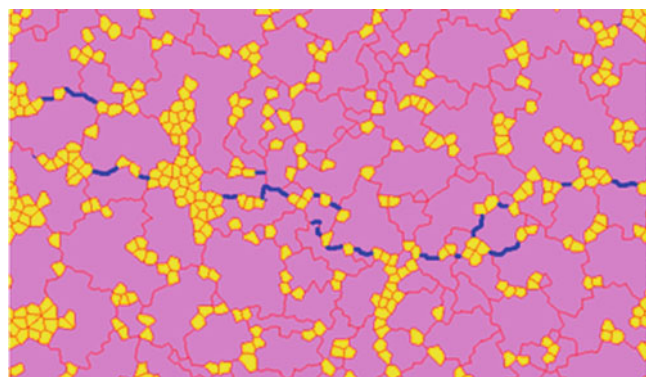
structural facets. Also, VTs help in approximating the microstructure of rocks (e.g., sandstone) by two-component cell systems as starting point for modeling compression and tensile tests (Li et al. 2017), see Fig. 5.

Further Tessellation Models

Closely related to VTs are *Delaunay tessellations*, which constitute triangulations of the plane. Given a VT, the Delaunay tessellation is constructed out of the triangles/tetrahedra formed by the generators whose cells share the same vertex. Figure 6 shows a VT and some triangles of the corresponding Delaunay tessellation.

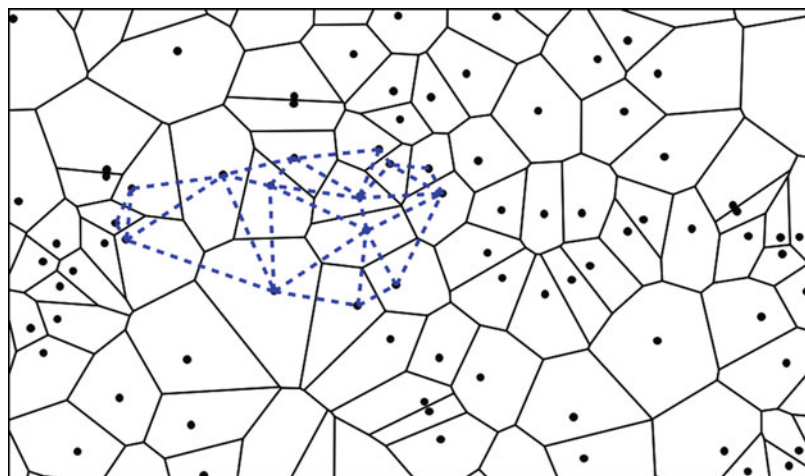
Triangulations based on Delaunay tessellations serve as statistical tools for interpolation and as means for the approximation of spatial surfaces. It makes sense to consider points of a point pattern as “neighbors” if they are connected directly by Delaunay edges.

There are many other models (Chiu et al. 2013), some of them starting from systems of edges, which may be cracks in the geoscientific context. Also tessellations on sphere



Stochastic Geometry in the Geosciences, Fig. 5 Voronoi-based tension test. Schematic result of a numerical tension test with fractures connecting neighboring pores (Li et al. 2017). *Magenta* solid phase, *yellow* pore phase, *blue* tensile cracks

Stochastic Geometry in the Geosciences, Fig. 6 A Voronoi tessellation and some Delaunay triangles. Planar Voronoi cells and their generators. In blue some Delaunay triangles are shown. (Data courtesy of Sung Nok Chiu)



surfaces appear, in particular the largest tessellation on Earth, the system of plates of the Earth’s lithosphere.

Tools of Stochastic Geometry

Regionalization

A regularized variable is a quantity $Z(x)$ defined for every location x of space. Examples are porosity at point scale at x (either 0 or 1) or ash content of coal at x (a nonnegative number, being the ash content of a small sample (ball or cube) centered at x). The statistics of regionalized variables is statistics for random fields and sometimes this goes under the name “geostatistics.” Though the basic structures of stochastic geometry are not regionalized variables, sometimes it makes sense to regionalize such structures, in order then to apply the strong methods of geostatistics. Two approaches are:

1. *Smooth geometrical structures.* For example, consider the number of points $Z(x)$ of a point process in the ball $b(x, R)$ for every point x of the space for some chosen radius R . Or, for a fiber process, consider the total length $Z(x)$ of all fiber pieces in $b(x, R)$, perhaps divided by the ball volume. For the case of statistically homogeneous and isotropic structures this approach is systematically considered in (Chiu et al. 2013), Section 6.4.4.

For inhomogeneous structures smoothing regionalization leads to estimates of the intensity function $\lambda(x)$ of a point process or spatially variable fracture density $S_V(x)$ of a fracture system. The choice of the smoothing parameter R presents a statistical problem.

2. *Generate the distance function.* Consider the distance $d(x)$ from the point x to the nearest point in the geometrical structure X . Set $Z(x) = d(x)$. If X is stationary and isotropic then so is also the random field $\{Z(x)\}$. An example for this approach is presented in Baddeley et al. (2016). There X is

a pattern of geological lineaments and the distance field is used to study correlations between X and a point pattern of copper deposits.

Simulation

Typically, all realistic structures of stochastic geometry relevant for the geosciences are too complicated for analytical-mathematical treatment. Therefore, they are studied by simulation. Simulation is used in deriving statistical summary characteristics for models, for generating samples for investigations of structural properties such as permeability, for generating test images for statistical investigations and multiple-point statistics and for generating realistic images of deposits starting from geological exploration, and for demonstrating the uncertainty of geological modeling by simulating repeatedly the same model.

Various methods are in use, depending on the type of model. There are simple direct methods, as for the PB model (see Fig. 1), for Poisson planes (see Fig. 2), or for Voronoi tessellations with respect to some point processes (using the nontrivial algorithms for the tessellation construction), see Chiu et al. (2013). For more complicated cases Markov chain Monte Carlo (MCMC) methods may be suitable for generating structures with inner interaction, e.g., for Gibbs fields.

A classical introduction to simulation is Lantuejoul (2002). For the generation of *pixel patterns* with prescribed statistical properties, geomathematics uses *multiple-point simulation*, which is of a sequential nature, see article ► “Multiple Point Statistics”. Here two examples are briefly described. In both cases, the pixels are visited on a random path through the lattice and the values at the pixels on the paths are changed. The first method, which was introduced by Strebelle (2002), works locally, uses conditional distributions (c.d.) for the new value at a pixel x given the values at n pixels in the neighborhood of x . The c.d.’s come often from *training images*. The method can be seen as working with an “empirical Gibbs sampler”; if a suitable model would be available, the c.d. were given by the model.

The other method, called “reconstruction,” works globally. It uses summary characteristics as targets for the whole pattern, which come from training images or parts of the structure that has to be reconstructed, e.g., planar sections or outcrops. For this purpose not only is the variogram used but also other characteristics as, e.g., various contact distribution functions. The algorithm minimizes the sum of squared differences between the calculated and empirical targets via stochastic optimization techniques such as simulated annealing by changing the value of one pixel per step, see article ► “Simulated Annealing” and Arns et al. (2002); Ballani and Stoyan (2015); Chiu et al. (2013); and Peredo and Ortiz (2011). Both approaches can be made conditional, using

simply known values at some pixels, known from exploration or planar sections (outcrops) and fixed during the simulation.

Object models or germ-grain models are simulated by quite different methods (Hauge et al. 2017; Pyrcz and Deutsch 2014), see Holden et al. (1998) for mathematical details. Starting from some initial configuration, the objects (lobes, channels, etc.) are removed and/or added to the system, following some rules for the mutual spatial relationships of the objects. Also these simulations can be made conditional, which is rather complicated (Hauge et al. 2017).

Stereology

Stereology is the art of obtaining 3D information from planar or linear sections, in the larger scale by investigations in the geosciences from rock outcrops, geophysical measurements, or drill holes, at the small scale from polished sections. See the *article Stereology*. There is a series of mean-value formulas, from which two are presented here.

Volume Fraction

$$V_V = A_A = L_L = P_P \quad (8)$$

A stationary 3D structure with one phase of interest is considered, V_V denotes the part of space occupied by this phase. This volume fraction can be obtained by measuring the *area fraction* A_A in a planar section, or the *linear fraction* L_L in a linear section, or the *point fraction* P_P on a lattice of test points.

Specific Surface

$$S_V = \frac{4}{\pi} L_A = 2P_L \quad (9)$$

For a 3D stationary and isotropic surface process S_V denotes the specific surface, mean surface area per unit volume. It can be obtained by measuring the *line density* L_A in a planar section (the mean length of section profiles from surface pieces per unit area) or the *specific number* P_L of intersection points of test line and surface pieces per length unit.

Summary

Geomathematics largely studies random geometrical structures. Thus stochastic geometry is potentially an important discipline in this area. It offers ideas that can aid in systematizing geomathematical knowledge and methods. At present stochastic geometry is dominated by applications in materials science, so systematic work is needed to adapt it for use in the geosciences.

It may be fruitful to add stochastic geometry to the general body of theory for geomathematics. The fact that most models

of stochastic geometry assume stationarity and isotropy is a difficult problem (as also in classical geostatistics). Real geological structures can rarely be modeled this way; this is a general problem of spatial statistics.

Cross-References

- [Geostatistics](#)
- [Markov Random Fields](#)
- [Multiple Point Statistics](#)
- [Porous Medium](#)
- [Simulated Annealing](#)
- [Simulation](#)
- [Stereology](#)
- [Three-Dimensional Geologic Modeling](#)

Bibliography

- Arns CH, Knackstedt MA, Mecke KR (2002) Characterising the morphology of disordered materials. In: Mecke K, Stoyan D (eds) *Morphology of condensed matter. Physics and geometry of spatially complex systems*. Springer-Verlag, Berlin, Heidelberg/New York, pp 37–74
- Baddeley A, Rubak E, Turner R (2016) *Spatial point patterns. Methodology and applications with R*. CRC Press, Boca Raton
- Ballani F, Stoyan D (2015) Reconstruction of random heterogeneous media. *J Microsc* 258:173–178
- Bonneau F, Caumon G, Renard P (2016) Impact of a stochastic sequential initiation of fractures on the spatial correlations and connectivity of discrete fracture networks. *J Geophys Res Solid Earth* 121:5641–5658
- Cherpeau N, Caumon G (2015) Stochastic structural modelling in sparse data situations. *Pet Geosci* 21:233–247
- Cherpeau N, Caumon G, Lévy B (2010) Stochastic simulations of fault networks from 2D seismic lines. *SEG Expanded Abstracts* 29:2366–2370
- Chiu SN, Stoyan D, Kendall WS, Mecke J (2013) *Stochastic geometry and its applications*. J. Wiley & Sons, Chichester
- Davy P, Le Goc R, Darcel C (2013) A model of fracture nucleation, growth and arrest, and consequences for fracture density and scaling. *J Geophys Res Solid Earth* 118:1393–1407
- Haldorsen HH, Chang DM (1986) Notes on stochastic shales; from outcrop to simulation model. In: Lake LW, Carroll HB Jr (eds) *Reservoir characterization*. Academic Press, Orlando, pp 445–485
- Hauge R, Vigsnæs M, Fjellvoll B, Vevle ML, Skorstad A (2017) Object-based modelling with dense well data. In: Gómez-Hernández JJ, Rodrigo-Ilari J, Rodrigo-Clavero ME, Cassiraga E, Vargas-Guzmán JA (eds) *Geostatistics Valencia*. Springer International Publ. AG, Cham, pp 557–572
- Holden L, Hauge R, Skare Ø, Skorstad A (1998) Modelling of fluvial reservoirs with object models. *Math Geol* 30:473–496
- Ivanova VM, Sousa R, Murihy B, Einstein HH (2014) Mathematical algorithm development and parametric studies with the GEOFRAC three-dimensional stochastic model of natural rock fracture systems. *Comput Geosci* 67:100–109
- Koike K, Kubo T, Liu C, Masoud A, Amano K, Kurihara A, Matsuoka T, Lanyon B (2015) 3D geostatistical modeling of fracture system in a granitic massif to characterize hydraulic properties and fracture distribution. *Tectonophysics* 660:1–16
- Lantuejoul C (2002) *Geostatistical simulation*. Springer-Verlag, Berlin, Heidelberg
- Li J, Konietzky H, Frühwirth T (2017) Voronoi-based DEM simulation approach for sandstone considering grain structure and pore size. *Rock Mech Rock Eng* 50:2749–2761
- Merland R, Caumon G, Lévy B, Collon-Drouaillet P (2014) Voronoi grids conforming to 3D structural features. *Comput Geosci* 18:373–383
- Peredo O, Ortiz JM (2011) Parallel implementation of simulated annealing to reproduce multiple-point statistics. *Comput Geosci* 37:1110–1121
- Pyrz MJ, Deutsch CV (2014) *Geostatistical reservoir modeling*, 2nd edn. Oxford University Press, Oxford/New York
- Rongier G, Collon P, Renard P (2017) Stochastic simulation of channelized sedimentary bodies using a constrained L-system. *Comput Geosci* 105:158–169
- Strebel S (2002) Conditional simulation of complex geological structures using multiple-point statistics. *Math Geol* 34:1–22
- Wellmann F, Caumon G (2018) 3-D structural geological models: concepts, methods, and uncertainties. *Adv Geophys* 59:1–121

Stoyan, Dietrich

Jan Harff

Institute of Marine and Environmental Sciences, University of Szczecin, Szczecin, Poland



Fig. 1 Dietrich Stoyan, 2009 (Photo by Helga Stoyan)

Biography

Dietrich Stoyan is a German mathematician who has worked mainly in geostatistics, spatial statistics, and stochastic geometry. He was born in Berlin in 1940 and grew up in the Halberstadt region in East Germany. Stoyan studied mathematics from 1959 to 1964 at the Technical University of Dresden, and then went to Freiberg in Saxony, where he has

lived since then. Initially, he worked in an institute for industrial research, where he applied geostatistical models to problems of lignite exploration. He received his PhD in 1967 (Dr.-Ing.) at the Bergakademie Freiberg (BAF, the worldwide oldest higher mining school) and completed his habilitation in mathematics at BAF in 1975. In 1976, he became a lecturer (Dozent) for mathematics at BAF. After the political changes in Europe and the German reunification, Dietrich Stoyan served as President (Rector) of the BAF from 1991 to 1997, which was renamed “Technische Universität Bergakademie Freiberg” at that time.

Dietrich Stoyan’s field of work is applied stochastics. He started in queueing theory and later turned to stochastic geometry and spatial statistics. He is an author or coauthor of books on stochastic geometry and point process statistics which are standard references. These books and many of his papers describe the geometry of geological phenomena by mathematical models from stochastic geometry, such as tectonic faults along oceanic rift zones or karst structures. At the first World Congress of the Bernoulli Society held at Tashkent in 1986, Dietrich Stoyan gave an invited lecture on “Stochastic Geometry in Geology”. He created a series of models for point processes and random sets and developed various methods in spatial statistics, especially for marked point processes. He initiated the use of the pair correlation function in point process statistics, and together with his PhD student, André Tscheschel, he introduced the reconstruction method for point processes. He has performed statistical analyses of basalt columns’ geometry and the evolution of taxa in paleontology. In addition, he was able to clarify a difficult text by Georgius Agricola from the sixteenth century on mineral classification, using basic ideas of combinatorics. Four of Stoyan’s most important scientific publications are listed in the references (Stoyan and Stoyan 1994; Stoyan and Penttinen 2000; Illian et al. 2008; Chiu et al. 2013).

Stoyan holds honorary doctorates from the universities of Jyväskylä (Finland) and Technical University of Dresden (Germany). He is a member of the Academia Europaea, of the Leopoldina (National German Academy of Sciences), and of the Berlin-Brandenburg Academy of Sciences and is a Fellow of the Institute for Mathematical Statistics.

Bibliography

- Chiu SN, Stoyan D, Kendall WS, Mecke J (2013) *Stochastic geometry and its applications*. Wiley, Chichester
- Illian J, Penttinen AK, Stoyan H, Stoyan D (2008) *Statistical modelling of spatial point patterns*. Wiley, Chichester
- Stoyan D, Penttinen AK (2000) Recent applications of point process methods in forestry statistics. *Stat Sci* 15:61–78
- Stoyan D, Stoyan H (1994) *Fractals, random shapes and point fields*. Wiley, Chichester

Stratigraphic Sequence

Katherine L. Silversides

Australian Centre for Field Robotics, Rio Tinto Centre for Mining Automation, The University of Sydney, Sydney, NSW, Australia

Definition

A stratigraphic sequence is a series of sedimentary rocks that are classified based on their stratal stacking patterns and stratigraphic relationships (Catuneanu 2017).

Introduction

A stratigraphic sequence is a series of rock layers formed in sedimentary depositional environments. The stacking patterns of the layers and their relationships are used in the interpretation and classification of individual sedimentary units. The boundaries are defined by contacts that mark distinct changes in the stratal stacking pattern between two units. Sequence stratigraphy can be used as a generic, process-based method for mapping and correlation within a sedimentary basin that is independent of the geological setting and data used. However, multiple different approaches to stratigraphic sequencing exist (Catuneanu 2017; Mitchum et al. 2002). Studying the stratigraphic sequence of a sedimentary basin can provide information about the distribution of source rocks, depositional environment, migration pathways and traps, and some information about therefore the properties of the individual units (Bjørlykke 2015; Dalrymple and Choi 2007).

In a stratigraphic sequence the units are collated into genetically related groups, with chronostratigraphically significant surfaces bounding each group. These surfaces are generally depositional discontinuities where there are breaks in the deposition of new material, and often erosion. They normally occur as nonconformities or angular unconformities and can cut through multiple units within the group. Igneous layers can also occur, as volcanic layers or intrusions. In comparison, the surfaces separating units within a group are generally conformable and mark changes in the material being deposited. The rate of deposition of individual units is dependent on both the rate of sediment accommodation (i.e., relative fall in sea level) and the sediment supply rate (Bjørlykke 2015; Diessel 2007; Mitchum et al. 2002; Posamentier 2002).

The type of sequence deposited also depends on the interplay between accommodation space and sediment supply (Fig. 1). Progradational deposition occurs when the sediment

Banded Iron Formation

Another example of stratigraphic sequences is banded iron formations (BIF). These are rock units that contain deposited thin layers that alternate between a chert or silica-based layer and an iron oxide layer, often magnetite dominated. These layers form on a large range of scales, ranging from submillimeter-scale bands to the macrobands that can be meters thick. BIF typically contain 20–40% iron and 40–50% silica. Other layers of materials such as shales or volcanic rock may also be present. BIF are commonly found in Precambrian sedimentary successions (3800–543 Ma); however, they are no longer deposited and therefore their formation cannot be directly observed. Additionally, BIF lack the textural features that are common in other stratigraphic sequences. Therefore, their method of formation is still debated. Their deposition is linked to significant changes in the composition of the Earth's atmosphere and oceans. As unaltered BIF can retain the chemical signatures of the seawater they precipitated from, they can be used to study ancient seawater composition. Where sections of the BIF have been enriched through supergene or hypogene processes, they form high-grade iron ore deposits (Mloszewska et al. 2015; Trendall 2005).

Summary

Stratigraphy sequences are layers of sedimentary rocks formed in various underwater depositional environments. They are interpreted and classified based on their relative relationships and contacts. The formation of each layer depends on multiple factors such as sediment source, sediment supply, sediment accommodation, and depositional environment. Examples of stratigraphic sequences include coal deposits, petroleum fields, and BIF.

Cross-References

► [Sedimentation and Diagenesis](#)

Bibliography

- Bjørlykke K (2015) Petroleum geoscience: from sedimentary environments to rock physics, 2nd edn. Springer, Berlin Heidelberg, p 662p
- Catuneanu O (2006) Principles of sequence stratigraphy. Elsevier, Amsterdam, p 375p
- Catuneanu O (2017) Sequence stratigraphy: guidelines for a standard methodology. In: Montenari M (ed) Stratigraphy and timescales, vol 2. Academic Press, pp 1–57
- Dalrymple W, Choi K (2007) Morphologic and facies trends through the fluvial–marine transition in tide-dominated depositional systems: a

- schematic framework for environmental and sequence-stratigraphic interpretation. *Earth Sci Rev* 81:135–174
- Diessel CFK (2007) Utility of coal petrology for sequence-stratigraphic analysis. *Int J Coal Geol* 70:3–34
- Mitchum RM, Vail PR, Sangree JB (2002) Sequence stratigraphy: evolution and effects. In: Armentrout JM, Rosen NC (eds) Sequence stratigraphic models for exploration and production: evolving methodology, emerging models, and application histories. SEPM Society for Sedimentary Geology
- Mloszewska AM, Haugaard RN, Pecoits E, Konhauser KO (2015) Banded iron formation. In: Gargaud M et al (eds) Encyclopedia of astrobiology. Springer, Berlin, Heidelberg, pp 229–238
- Posamentier HW (2002) Sequence stratigraphy past, present and future, and the role of 3-D seismic data. In: Armentrout JM, Rosen NC (eds) Sequence stratigraphic models for exploration and production: evolving methodology, emerging models, and application histories. SEPM Society for Sedimentary Geology
- Trendall A (2005) Sedimentary rocks: banded iron formations. In: Selley RC, Cocks LRM, Plimer IR (eds) Encyclopedia of geology. Elsevier, pp 37–42
- Warwick PD (2005) Coal systems analysis: a new approach to the understanding of coal formation, coal quality and environmental considerations, and coal as a source rock for hydrocarbons. In: Warwick PD (ed) Coal systems analysis. The Geological Society of America, Colorado, pp 1–8

Structuring Element

B. S. Daya Sagar¹ and D. Arun Kumar²

¹Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, Karnataka, India

²Department of Electronics and Communication Engineering and Center for Research and Innovation, KSRM College of Engineering, Kadapa, Andhra Pradesh, India

Definition

Structuring element (SE) is the part of an image that operates on the image itself to obtain the shape/size information of the objects in the image.

Introduction

The utilization of earth's resources such as water, minerals, crops, and forest by human beings is increasing day by day (Campbell 1987). Due to this reason, scientists are motivated to monitor the resources on a temporal basis. The advantages of remote sensing such as spatial resolution, spectral resolution, radiometric resolution, and temporal resolution provide the information about the resources periodically (Lillesand et al. 2014). The information of resources is obtained in an image format by the remote sensing sensors.

An image consists of finite elements called pixels/patterns/pels (Campbell 1987). The pixel is a digital number (DN) with a value between 0 and 255 (0 to 2^8-1) for an 8-bit display system (Lillesand et al. 2014). The DNs in an image replicate the objects present in the area of interest. There exist various mathematical methods to operate on the pixels of an image to extract the information about the objects in the images. These methodologies contribute to important remote sensing image analysis approaches such as image segmentation, image classification, change detection, etc. (Lillesand et al. 2014). There exists a good literature survey in the domain of image segmentation, image classification, change detection, etc.

In the initial stage, the raw images are preprocessed using various preprocessing techniques for radiometric/geometric corrections (Campbell 1987). The preprocessed image is used to extract the information by applying the mathematical models. The images are processed through various approaches such as spatial domain, frequency domain, etc. Spatial domain is the most preferable approach in image analysis to extract the information from an image.

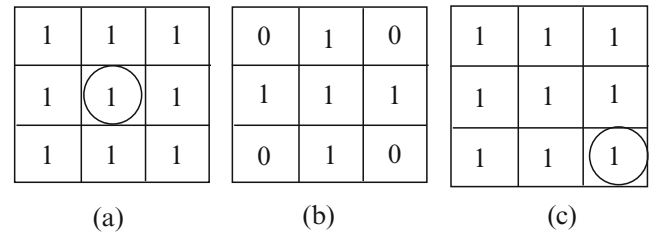
In the spatial domain, a filter/window/kernel is operated on the input image to get the information related to edges, boundaries, lines, and closed objects (Lillesand et al. 2014). The filtering techniques contribute to segment the information about the objects in the image to a finite level. In the direction of spatial domain analysis of images, a structuring element-based filtering was suggested to obtain the shape/size information of objects in digital images. A set theory-based approach is implemented in structuring element-based spatial filtering and the field of study was named mathematical morphology (MM) (Matheron 1975; Serra 1982).

Structuring Element

The structuring element is a microstructure of the set employed to perform morphological operations. The role of the structuring element (B) is to unravel the hidden morphological properties of the set (X) (Serra 1982). In general, set X is considered as the input image and B is a micro part of X and B operates on X .

The SE elements are categorized as flat, non-flat, and graph structures (Soille 1999). The main characteristics of structuring elements include shape, size, orientation, and symmetry as shown in Fig. 1.

A square SE of size 3×3 with the center element associated by eight of its neighbors (also called eight connectivity) is presented in Fig. 1a. Similarly, the rhombical SE with 4-connectivity is given in Fig. 1b. The square and rhombical SEs are symmetric about the center or origin, that is, the rotation of the SE on its origin will result in SE with change in values. The SE which is asymmetric about the origin (bottom right) is given in Fig. 1c. The structuring element (B) in a 2D plane (Z^2) can be a 3×3 square given as $B = \{(-1,-1), (-1,0), (-1,1), (0,-1), (0,0), (0,1), (1,-1),$



Structuring Element, Fig. 1 (a) Square (disc in eight connectivity grid), (b) rhombical (disc in 4-connectivity grid) in shapes of primitive size of 3×3 , symmetric about the origin, and (c) structuring element that is asymmetric about the origin (shown in parenthesis)

$(1,0), (1,1)\}$, the elements at the coordinates given in B are binary 1s. Similarly, the rhombic structuring element B in (Z^2) is given as $B = \{(-1,0), (0,-1), (0,0), (0,1), (1,0)\}$.

Characteristic Information of Structuring Element

SE possesses the characteristics like shape, size, orientation, and origin to perform the morphologic transformations on set X . Broadly, the SE is categorized as symmetric and asymmetric (Maragos and Schafer 1986). A symmetric template is defined as $B^s = [-b : b \in B]$, where B^s is obtained by rotating B by 180° on the plane and b is an element in B (Maragos 1989). In Fig. 1a, and 1b B is symmetric with respect to the origin with a size of 3×3 . In asymmetric SE, the symmetric property of B is not satisfied indicating that a 180° shift in B will not produce B . The asymmetric SE of size 3×3 is given as,

$$B = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad (1)$$

Shape of Structuring Element

Different shapes of SEs are considered to extract the shape/size information of objects in an image (Maragos and Ziff 1990). The important shapes of SEs are square, rhombic, ball, line, etc. The square and rhombic SEs are given in Fig. 1a, b.

Decomposition of Structuring Element

The structuring element of higher connectivity can be decomposed to lower connectivity to obtain the finite level of details from the input image (Sagar 2013). In Fig. 2a, the square SE of 8-connectivity (Fig. 1a) is decomposed to line SE of 2-connectivity to obtain the lines located in 45° orientation from the input image. Similarly, the square SE with 8 connectivity is decomposed to SE with 1-connectivity to obtain the upper line details located at 135° (Fig. 2b) orientation in an image. The square SE is decomposed to rhombic SE (Fig. 2c) to obtain the finite details of the objects in an image.

to process the input image to obtain the information about the objects of different sizes and shapes.

Cross-References

- [Grayscale Mathematical Morphology](#)
- [Mathematical Morphology](#)
- [Morphological Closing](#)
- [Morphological Dilation](#)
- [Morphological Erosion](#)
- [Morphological Opening](#)
- [Morphological Pruning](#)
- [Serra, Jean](#)
- [Thickening](#)

Bibliography

- Campbell JB (1987) Introduction to remote sensing. The Guilford Press
- Lillesand T, Kiefer RW, Chipman J (2014) Remote sensing and image interpretation. Wiley
- Maragos PA (1989) Pattern spectrum and shape representation. *IEEE Trans Pattern Anal Mach Intell* 11:701–716
- Maragos PA, Schafer RW (1986) Morphological skeleton representation and coding of binary images. *IEEE Trans Acoust Speech Signal Process* 34(5):1228
- Maragos P, Ziff RD (1990) Threshold superposition in morphological image analysis systems. *IEEE Trans Pattern Anal Mach Intell* 12(5):498–504
- Matheron G (1975) Random sets and integral geometry. Wiley, New York
- Sagar BSD (2013) Mathematical morphology in geomorphology and GISci. CRC Press (Taylor & Francis Group), Boca Raton, p 546
- Serra J (1982) Image analysis and mathematical morphology. Academic, London
- Soille P (1999) Morphological image analysis: principles and applications. Springer, Heidelberg

Sums of Geologic Random Variables, Properties

G. M. Kaufman
Sloan School of Management, MIT, Cambridge, MA, USA

Definition

Politicians, policy analysts, and business representatives often ask earth scientists to provide personal probability judgments about uncertain geological quantities. Assessment of oil and gas in unexplored petroleum plays and basins is a recurring example. Few geologists can provide sharp coherent judgments about probabilistic dependencies and often rely

on Pearson type correlations, flawed measures of dependency in many settings. More sophisticated techniques are needed that assure coherence of probabilistic judgments about uncertain geological random quantities and that reduce geologists' cognitive load.

For economists, bureaucrats, and politicians, right-tail probabilities are often the most interesting feature of a probabilistic oil and gas assessment. What, for example, is the probability of finding at least one more giant field in a mature petroleum province? Lillestøl and Sinding-Larsen's (2017) study of giant fields based on 182 North Sea discoveries highlights the importance of accurate modeling of tail probabilities. Blondes et al. (2013a, b) offer a rationale for careful attention to dependencies:

In the Circum-Arctic aggregation of the 48 AUs, the 90-percent uncertainty interval for recoverable gas is 1,471, 2,009, or 3,515 tcf for assumptions of independence, assessor specified dependency (correlation), or total dependence respectively. Clearly, decision makers who rely on assessment results need accurate interval projections. Too broad an interval provides little information; too narrow an interval gives a false sense of precision.

A rich probabilistic finance and actuarial risk analysis literature explores properties of sums of random variables when only marginal distributions are available (Hoeffding 1940; Fréchet 1951). It offers methods for assuring coherence of geologists' personal probability judgements about uncertain geological quantities in the absence of empirical data.

Probabilistic assessments of oil and gas in unexplored petroleum plays and basins are recurring examples. In the absence of hard data geologists' personal judgments about marginal distributions of geologic attributes are, in the large, reasonably well calibrated. However, their personal judgments about dependencies among them are problematic. According to Chen et al. (2012) "...variable correlations are so common among geologic variables that ignoring their interdependence may lead to serious bias, affecting both the resulting resource potential estimation."

Section "[Preliminaries](#)" describes co-monotonicity and its role in computing upper and lower bounds on sums of random variables. Section "[Thumbnail Case Studies](#)" presents thumbnail sketches of three geologic case studies. Concluding remarks appear in section "[Concluding Remarks](#)."

Preliminaries

Define F_X to be the distribution function of a random vector $\mathbf{X} = (X_1, \dots, X_n)^t$ with domain $\mathbf{R}^n$ and marginal distributions F_i , $i = 1, \dots, n$. Set $F_X(\mathbf{x}) = \text{Prob}\{X_1 \leq x_1, \dots, X_n \leq x_n\}$.

Divide difficulties, and assume that each F_i is continuous and possesses a one-to-one inverse. Define the p^{th} fractile of X_i as the value in the domain of X_i such that $\text{Prob}\{X_i \leq x_p\} = p$ and its inverse as $F_i^{-1}(p) = x_i(p)$. In turn the p^{th} fractile of the sum

$S_n = X_1 + \dots + X_n$ is s_p such that $\text{Prob}\{S_n \leq s_p\} = p$ or $F_{S_n}^{-1}(p) = s_p$.

What conditions guarantee that fractiles are strictly additive? That is, for all $p \in (0, 1)$ $s_p = x_1(p) + \dots + x_n(p)$? Imposition of functional dependencies among $X_1, \dots, X_n$ is one route to sufficient conditions. Suppose that $X_1, \dots, X_n$ share a common domain D_X and consider n continuous invertible functions h_i , each with domain D_X . Suppose that $x_i = h_i(x_1)$ for all $x_i \in D_X, i = 2, \dots, n$. Then $\text{Prob}\{X_1 + h_2(X_1) + \dots + h_n(X_1) < s\}$. The omnibus function $g(x_1) = x_1 + h_2(x_1) + \dots + h_n(x_1), x_1 \in D_X$ is continuous and invertible so $\text{Prob}\{g(X_1) < s\} = \text{Prob}\{X_1 < g^{-1}(s)\}$. The p^{th} fractile of S_n is s_p such that $\text{Prob}\{g(X_1) < s_p\} = p$ or $\text{Prob}\{X_1 < g^{-1}(s_p)\} = p$ leading to $x_1(p) = g^{-1}(s_p)$. Equivalently $g(x_1(p)) = s_p$. Functional dependencies of this type are too strong to model most geological data. Co-monotonicity is a more flexible approach to modeling joint behavior of dependent random variables.

Definition The random vector $\mathbf{X} = (X_1, \dots, X_n)^T$ is co-monotonic if and only if $(X_1, \dots, X_n) =_d (F_1^{-1}(U), \dots, F_n^{-1}(U))$, U a uniform random variable with domain $(0, 1)$.

Here $=_d$ means agreement in distribution. Each element of a co-monotonic random vector is a function of a single random variable U so all elements of $\mathbf{X}$ exhibit strong positive dependency. McNeil et al. (2005) provide a more general definition: $\mathbf{X}$ is co-monotonic if and only if it agrees in distribution with a random vector, each of whose components is a nondecreasing function of a single random variable. If elements of $\mathbf{X}$ are co-monotonic increasing one element of $\mathbf{X}$ increases all others. Research Report 9914 Goovaerts et al. (2000) provides a readable account of properties of sums of co-monotonic random variables in an actuarial context. Deelstra et al. (2009) review co-monotonicity in financial economics.

Foreshadowing geologists' critique that some elements of $\mathbf{X}$ may be independent or possibly negatively dependent (rather rare), co-monotonicity provides upper and lower bounds on a sum of random variables with specified marginal distributions that embrace a wide range of dependence structures. When these bounds are tight, reasonable projections of probability distributions of aggregates can be computed using marginal distributions and certain conditional expectations (see section "Bounds" (3)).

Bounds

A random variable X precedes a random variable Y in convex order, denoted by $X \geq_{cx} Y$ if and only if $E(g(X)) \geq E(g(Y))$ for all real convex functions g for which expectations are finite. Kaas et al. (2009) use convex order to show that fractiles of co-monotonic random variables additive in the following sense: for any random vector $\mathbf{X} = (X_1, \dots, X_n)$

possessing marginal cumulative distribution functions $F_1, \dots, F_n$ and U a uniform $(0, 1)$ random variable

$$(X_1 + \dots + X_n) \leq_{cx} S_u \equiv F_1^{-1}(U) + \dots + F_n^{-1}(U). \quad (1)$$

If $S_u =_d F_1^{-1}(U) + \dots + F_n^{-1}(U)$ it follows immediately that the p^{th} fractile of S_u is $F_{S_u}^{-1}(p) = F_1^{-1}(p) + \dots + F_n^{-1}(p)$, all $p \in (0, 1)$. They point out that (1) is a supremum in convex order and a best bound for marginal distributions in a Fréchet space. If a random vector $\mathbf{X}$ with marginal distributions $F_1, \dots, F_n$ belong to a Fréchet space $\mathcal{F}_n$ the joint cumulative distribution function $\text{Prob}\{X_1 \leq x_1, \dots, X_n \leq x_n\}$ of $\mathbf{X}$ is bounded from above by $M_n \equiv \min\{F_1(x_1), \dots, F_n(x_n)\}$. Goovaerts et al. note that M_n is reachable in $\mathcal{F}_n$. For sums of elements of $\mathbf{X}$ introduction of a random Z such that distribution of each X_i given Z is known with certainty leads to refined upper and lower bounds. In a geologic context Z is interpretable as a latent (background) variable describing gross geologic characteristics of, for example, a petroleum assessment unit. The conditioning variable Z might be regression dependent on geologic attributes of an assessment unit and need not be scalar. These authors define $F_{X_i|Z}^{-1}(U)$ to be a random variable $f_i(U, Z)$ that for $(U, Z) = (u, z)$ assumes value $F_{X_i|z}^{-1}(u)$ and prove that for U uniform $(0, 1)$ and Z independent of U

$$(X_1 + \dots + X_n) \leq_{cx} S_u^* \equiv F_{X_1|Z}^{-1}(U) + \dots + F_{X_n|Z}^{-1}(U). \quad (2)$$

Jensen's inequality leads to a lower bound

$$E(X_1|Z) + \dots + E(X_n|Z) \leq_{cx} (X_1 + \dots + X_n). \quad (3)$$

Kaas et al. (2009) point out that (a) the random vector $E(X_1|Z) + \dots + E(X_n|Z)$ will not in general have marginal distributions $F_1, \dots, F_n$ and (b) if $E(X_1|Z), \dots, E(X_n|Z)$ are either jointly nonincreasing or nondecreasing functions of Z the LHS in (3) is a sum of co-monotonous random variables and $\text{Var}(E(X_i|Z)) < \text{Var}(X_i)$ unless $\text{Var}(E(X_i|Z)) = 0$. In order to compute the cdf of the LHS of (3), suppose that (b) obtains and each of the random variables $E(X_1|Z), \dots, E(X_n|Z)$ are nondecreasing functions of increasing $Z = z$. Write the lower bound as, $E(X_1|Z) + \dots + E(X_n|Z) = E(S|Z)$, and define $F_{E(X_i|Z)}(x) = \text{Prob}\{E(X_i|Z) \leq x\}$. They show that if the cdf of $E(X_i|Z)$ is continuous and increasing:

$$F_{E(X_1|Z)}^{-1}(F_{E(S|Z)}(x)) + \dots + F_{E(X_n|Z)}^{-1}(F_{E(S|Z)}(x)) = x, \quad (4)$$

a prescription for calculating a lower bound. The quality of (3) depends on the choice of model for Z . Kaas et al. (2002) and Goovaerts et al. (2000) demonstrate that (1) and (3) provide reasonable bounds on the cumulative distribution function of sums of dependent lognormal random variables. Comparison

of predictive distributions of undiscovered mineral resources derived by conventional methods with co-monotonic bounds on them is a promising avenue of research.

Thumbnail Case Studies

USGS Oil and Gas Resource Projections

Chen et al. (2013) describe the 1980 USGS assessment system called FASP (fast appraisal system for petroleum resources). FASP incorporates perfect positive correlation between micro-level reservoir attributes but allows specification of any positive correlation among play resources. Nevertheless, the USGS 2000 World Petroleum Assessment aggregates undiscovered resource volumes from assessment unit level to regional level using perfect correlation to add fractiles and arrive at regional level aggregates. Recognizing that at the global-level dependencies among large regional aggregates of resources are unlikely to be perfectly correlated they assign pairwise correlation 0.5 to each pair of eight regions (Klett et al. 2000). No sensitivity analysis of aggregate projections is given.

Many USGS assessment studies present tables of fractiles of individual assessment units and add fractiles to arrive at total resource fractiles. Fractile addition is justified by stating that “Fractiles are additive under assumption of perfect positive correlation.” Table 2 in “Assessment of Undiscovered Continuous Oil and Gas Resources in the Monterey Formation, San Joaquin Basin Province, California” USGS Fact Sheet 2015–3058 September 2015 and Table 2 in USGS Fact Sheet 2014–3082 “Assessment of Potential Shale-Oil and Shale-Gas Resources in Silurian shales of Jordan” September 2014 are examples. Chen et al. (2013) cite additional examples (Klett et al. 2000, 2004, 2005). It is easy to show that “perfect correlation” is not robust to variations in specification of functional forms of marginal distributions. Worse, addition of fractiles without careful attention to properties of the joint distribution of geological uncertain quantities leads to incoherence. On the other hand, mutual independence allows specification of arbitrary marginal probability distributions without affecting coherence but leads to unacceptably narrow probability projections of sums of oil and gas magnitudes.

A salient feature of Pearson’s correlation coefficient is that random variables X and Y possess correlation 1.0 or -1.0 only if X and Y are linearly dependent. Denuit and Dehaene (2003) point to a limiting case, a pair of normal random variables for which the variance of one member of the pair is zero. Then, if X and Y are jointly lognormal and $\log X$ is a linear function of $\log Y$, the Pearson correlation of $\log X$ and $\log Y$ is either 1.0 or -1.0 . However, the Pearson correlation of

X and Y is less than 1.0. These authors provide a more nuanced treatment. Suppose F_1 and F_2 are marginal cumulative distribution functions of X and Y , respectively, each concentrated on $(0, \infty)$ and U is a uniform random variable independent of X and Y . Using super-modularity, they prove that if F_1 and F_2 lie in a Fréchet space, the Pearson correlation coefficient $r(X, Y)$ of X and Y is bounded by

$$\frac{\text{Cov}(F_1^{-1}(U), F_2^{-1}(1-U))}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} \leq r(X, Y) \leq \frac{\text{Cov}(F_1^{-1}(U), F_2^{-1}(U))}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} \quad (5)$$

so perfect correlation is not achievable. They also prove that it is possible for a pair of co-monotonic lognormal random variables to have pairwise correlation close to zero, contradicting the intuitive notion that small correlation implies weak dependence.

These authors document several important features of Kendall’s τ and Spearman’s ρ . (Spearman’s ρ is at the center of Iman and Conover’s method deployed in the USGS (2013) CO_2 sequestration study to compute predictive probability distributions of aggregates). First, both are invariant with respect to strictly monotone transformations. Second, when one variable is a nondecreasing (nonincreasing) transformation of the other they equal 1 (or -1) at the Fréchet upper (resp. lower) bound. They note that at a value of 1.0 or -1.0 , Kendall’s τ and Spearman’s ρ achieve Fréchet bounds. Kendall’s τ and Spearman’s ρ are more desirable measures of association for non-normal multivariate distributions than Spearman’s r because the latter does not share Kendall and Spearman’s correlation invariance properties which come into play in Iman and Conover’s method. Denuit and Dehaene prove the nonobvious fact that if positively or negatively quadrant-dependent random pairs are jointly uncorrelated, they are mutually independent.

“Perfect correlation” as an omnibus argument for adding fractiles has many pitfalls. Co-monotonic bounds on random sums are a conceptually satisfactory alternative.

USGS Probabilistic Assessment of CO_2 Storage Capacity

The USGS probabilistic assessment of CO_2 sequestration in mature petroleum reservoirs (Blondes et al. 2013a, b) is based on both micro- and macro-assessments. Their macro-assessment aggregates storage assessment units (SAUs) at basin, regional, and national levels. An objective was to take into account dependencies among assessment units arising from “overlap of geologic analogs, assessment methods and assessors” using individual SAU marginal probability distributions and “. . . a correlation matrix obtained by expert

elicitation describing interdependencies between pairs of SAUs.” The correlation matrix dimension is 192×192 . Because a menagerie of marginal distributions – Beta-PERT, lognormal, truncated lognormal – were deployed at the micro-level, the use of standard multivariate distribution theory is not appropriate. Dependencies among storage capacity magnitudes rest on an innovative distribution free method developed by Iman and Conover (1982) allowing marginal distributions and dependencies to be estimated from distinct data sets.

How to aggregate from basin, to region, and then to a national scale? A single stage using the correlation matrix for all SAUs in the study? Or successively aggregate subsets of SAUs in multiple stages? Blondes et al. (2013b) conclude that:

Although the single-stage approach requires determination of significantly more correlation coefficients, it captures geologic dependencies among similar units in different basins and it is less sensitive to fluctuations in low correlation coefficients than the multiple stage approach.

Successive aggregation in multiple stages drastically reduces the number of pairwise correlations elicited from geologists at the expense of requiring each assessor to appraise pairwise correlations of sums of assessment unit magnitudes. Appraisal of resource volumes in an individual SAU are more likely to be well calibrated than direct appraisal of a sum of SAU volumes (Kaufman et al. 2017).

Cupolas and Oil and Gas Resource Assessment

Chen et al. (2013) emphasize that at an assessment micro-level, reservoir attributes such as porosity, permeability, pressure, and temperature are often decisively dependent and that empirical data suggest dependencies are present among more aggregate assessment units in mature provinces. They argue that a basin’s tectonic framework exerts “strong geographic control” over many geological features and leads to geographic and spatial dependencies and that because plays in a given basin share “. . . petroleum system elements, such as source rocks, regional top seal, migration fairways, timing, regional tectonics for trap formation, and accumulation preservation factors.” They are the first to use copulas in this setting.

Sklar (1959) proved that subject to mild restrictions a multivariate cumulative distribution can be mapped to a joint cumulative distribution of uniform random variables called a cupola. As with Iman and Conover’s method, cupolas allow marginal distribution shapes to be computed from data sets distinct from those used to estimate dependency structure.

Suppose as in section “Thumbnail Case Studies” above that F_X is the distribution function of a random vector $\mathbf{X} = (X_1, \dots, X_n)^t$ with domain $\mathbf{R}^n$ and marginal cumulative

distributions F_i , $i = 1, \dots, n$. Let $\mathbf{U}_n = (U_1, \dots, U_n)$ be a vector of independent uniform $(0, 1)$ random variables and $\mathbf{u}_n = (u_1, \dots, u_n)$ be a realization of $\mathbf{U}_n$. Then with $u_i = F_i(x_i)$, $i = 1, \dots, n$ $\text{Prob}\{X_1 \leq x_1, \dots, X_n \leq x_n\} = \text{Prob}\{U_1 \leq u_1, \dots, U_n \leq u_n\}$.

Definition $C(u_1, \dots, u_n) = \text{Prob}\{U_1 \leq u_1, \dots, U_n \leq u_n\}$ is the cupola of F_X .

Set $dF_i = f_i$, $i = 1, \dots, n$ and $dC(u_1, \dots, u_n) = c(u_1, \dots, u_n)du_1 \dots du_n$. The joint density of $\mathbf{X}$ can be written as $c(u_1, \dots, u_n) \times f_1(x_1) \times \dots \times f_n(x_n)$. The term c in the joint density captures the dependency structure of elements of $\mathbf{X}$. Because $\text{Prob}\{X_1 \leq x_1, \dots, X_n \leq x_n\} = \text{Prob}\{U_1 \leq u_1, \dots, U_n \leq u_n\}$ a procedure for generating samples from C produces samples of $\mathbf{X}$ by inversion of $u_i = F_i(x_i)$, $i = 1, \dots, n$.

Computation requires choice of a cupola functional form. Chen et al. chose the bivariate normal cupola, a popular choice closely tied to standard multivariate normal distribution theory.

Their regional resource assessment of the Canadian Arctic’s Beaufort-McKenzie Basin analyzes 48 “significant” oil and gas discoveries containing 53 distinct accumulations in three major petroleum systems. Empirical data allows estimation of pairwise correlations among reservoir attributes – area, porosity, oil saturation, and net pay. The authors treat geologic risk factors as probabilistically independent because the data does not allow empirical estimation of dependencies among them. Dependencies are restricted to play reservoir volume attributes. In turn these dependencies impact of the distribution of total resource volumes.

Four plays illustrate how to incorporate dependencies among individual play resources. No method for eliciting geologists’ judgments about dependencies between plays is described. However, the authors assert that between play correlations are large because all four plays share the same source rock and petroleum system. They call attention to substantial differences between total ultimate oil resource medians under the assumption of independence and under the assumption of within and between play correlations: the latter is 1.6 times the former.

In sum, to be realistic, probabilistic appraisal of oil and gas resources in unexplored and partially explored regions must account for multiple sources of dependencies. Cupolas are useful for doing so.

Concluding Remarks

In the absence of empirical data that allows resolution of the vexing problem of how to address probabilistic dependencies

among and between elements of large sets of geologic random variables data, we need methods that refocus and streamline expert geological judgment inputs as well as analytical methods for modeling dependencies that go beyond pairwise correlation and its cousins.

Bibliography

- Blondes MS, Schuenemeyer JH, Olea RA, Drew LJ (2013a) Aggregation of carbon dioxide sequestration storage assessment units. *Stoch Environ Res Risk Assess* 27:1839–1859
- Blondes MS, Scheunemeyer JH, Drew LJ, Warwick PD (2013b) Probabilistic aggregation of individual assessment units in the U.S. Geological Survey national CO₂ sequestration assessment. *Energy Procedia* 37:5110–5117
- Chen Z, Osadetz KG, Dixon J, Deitrich J (2012) Using copulas to implement variable dependencies in petroleum resource assessment: example from Beaufort-Mackenzie Basin. *Canada AAPG Bull* 96 30: 439–457
- Deelstra G, Jan Dhaene J, Vanmaelez M (2009) An overview of comonotonicity and its applications in finance and insurance. Working paper
- Denuit, Dehaene (2003) Simple characterizations of co-monotonicity and counter-monotonicity by extremal correlations. Univ. of Louvain working paper
- Fréchet M (1951) Sur les tableaux de corrélation dont les marges sont donnés. *Ann Univ Lyon Sect A, Ser 3* 14:53–77
- Goovaerts et al (2000) Research Report 9914 U. of Leuven
- Hoeffding W (1940) Masstabinvariante Korrelationstheorie. *Schriften des mathematischen Instituts und des Instituts für angewandte Mathematik des I.hii-versität Berlin* 5:179–233
- Iman RL, Conover WJ (1982) A distribution-free approach to inducing rank correlation among input variables. *Commun Statist – Simul Comput* 11(3):311–334
- Kaas R, Goovaerts M, Dhaene J, Denuit M (2001) *Modern Actuarial Risk Theory*. Kluwer Academic Publishers, Boston.
- Kaufman GM (2017) *Properties of Sums of Geological Random Variables Handbook of Mathematical Geosciences* B.S. Daya Sagar, Quiming Cheng, Fritz Agterberg eds. Springer ISBN 978-3-319-78998-9
- Klett TR (2004) Future petroleum supply: exploration or development?: 2nd U.S. Geological Survey conference on reserve growth Denver Federal Center, Lakewood, Colorado, May 6, 2004
- Klett TR, Charpentier RR, Schmoker JW (2000) Assessment operational procedures: US Geological Survey digital data series 60. <http://energy.cr.usgs.gov/WEcont/chaps/OP.pdf>. Accessed 1 Aug 2011
- Klett TR, Gautier DL, Ahlbrandt TS (2005) An evaluation of the U.S. Geological Survey world petroleum assessment 2000. *AAPG Bull* 89(8):1033–1042
- Lillestøl, Jostein; Sinding-Larsen, Richard, 2017. Creaming and the depletion of resources: A Bayesian data analysis, Discussion Papers 2017/16, Norwegian School of Economics, Department of Business and Management Science.
- McNeil AJ, Frey R, Embrechts P (2005) *Quantitative risk management. Concepts, techniques and tools*. Princeton series in finance. Princeton University Press, Princeton. ISBN 0-691-12255-5, MR 2175089, Zbl 1089.91037
- Sklar M (1959) Fonctions de répartition à n dimensions et leur marges. *Publ Inst Statist Univ Paris* 8:229–231

Support Vector Machines

Rogério G. Negri

São Paulo State University (UNESP), Institute of Science and Technology (ICT), São José dos Campos, São Paulo, Brazil

Synonyms

Support vector network

Definition

Support vector machine (SVM) is a supervised machine learning algorithm applied to data classification and regression. Starting from the 1990s early formalization presented by Cortes and Vapnik (1995), support vector machine (SVM) has been highlighted as a machine learning method widely applied in several research areas. Some reasons for all attention and focus received by the SVM method, especially over the two first decades of the twenty-first century, are its solid theoretical basis and characteristics like data distribution independence, simple algorithm architecture, moderate computational complexity, and excellent generalization ability.

A Formalization for Linearly Separable Data

The SVM method comprises to define a linear discriminant function $g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$ able to distinguish a set of vectors linearly separable $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i = 1, \dots, m\}$ between the classes ω_1 , when $y_i = +1$, and ω_2 , if $y_i = -1$. The sets $\mathcal{X} \subseteq \mathbb{R}^n$ and $\mathcal{Y}$ represent the attribute (vector) space of the data and a set of class indicators, respectively. Moreover, the function g delivers the largest separation margin between the pair of classes.

Behind the process of determining the maximum margin separation hyperplane lies the simple concept of distance from a point to line (Theodoridis and Koutroumbas 2008; Webb and Copsey 2011). Knowing that the distance from a given vector (point) $\mathbf{x}_i$ to the separation hyperplane (line) $g(\mathbf{x}) = 0$ is computed through $\frac{|g(\mathbf{x}_i)|}{\|\mathbf{w}\|}$, when rescaling $\mathbf{w}$ to make $\mathbf{x}_u \in \omega_1$ and $\mathbf{x}_v \in \omega_2$ from $\mathcal{D}$, identified as the closest points to $g(\mathbf{x}) = 0$, has distance equal to one from such hyperplane, it is modeled the following optimization problem which solution leads to $\mathbf{w}$ and b parameters assigned to the largest separation margin hyperplane:

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \mathbf{w}^T \mathbf{w} \\ \text{subject to : } & y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1, i = 1, \dots, m \end{aligned} \quad (1)$$

Figure 1a depicts the geometric relations of the maximum margin hyperplane. Vectors located at the limits of separation margin (i.e., $g(\mathbf{x}) = \pm 1$) are the ones responsible to define the optimum hyperplane, the so-called support vectors.

The optimization problem that allows reaching the maximum margin hyperplane is a convex function with linear constraints. Consequently, the Lagrangian multipliers approach allows solving such optimization problems more conveniently. Adopting this approach, solve (1) is equivalent to optimize the following:

$$\begin{aligned} \max_{\lambda} L_D(\lambda) &= \sum_{i=1}^m \lambda_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \lambda_i \lambda_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ \text{subject to : } &\begin{cases} \lambda_i \geq 0, & i = 1, \dots, m \\ \sum_{i=1}^m \lambda_i y_i = 0 \end{cases} \end{aligned} \quad (2)$$

where $\lambda_i \geq 0$; $i = 1, \dots, m$ are Lagrange multipliers.

The Lagrange multipliers that come after solving (2) allow computing the maximum margin parameters. While the vector $\mathbf{w}$ is equivalent to $\sum_{i=1}^m \lambda_i y_i \mathbf{x}_i$, the scalar b is computed after substitute $\mathbf{w}$ in $g(\mathbf{x}_i)$, where $\mathbf{x}_i$ is a support vector with $y_i = +1$, and consequently, $\mathbf{w}^T \mathbf{x}_i + b = 1$.

Additionally, beyond providing a simpler way to compute the optimum parameters, the inner product in Eq. 2 (i.e., $\mathbf{x}_i^T \mathbf{x}_j$), may also be replaced by a *kernel function*.

As a consequence of the obtained maximum margin hyperplane, rises the following discrimination rule:

$$g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b \begin{cases} \geq 0 \Rightarrow \mathbf{x} \in \omega_1 \\ < 0 \Rightarrow \mathbf{x} \in \omega_2 \end{cases} \quad (3)$$

Formalization for Nonlinearly Separable Data

The previous formulation assumes a classification problem where the vectors/data are linearly separable. However, it is usual that the classification problems are not restricted to such situations. In order to deal with nonlinearly separable data, slack variables $\xi_i \geq 0$, $i = 1 \dots, m$ are included to make valid the statements:

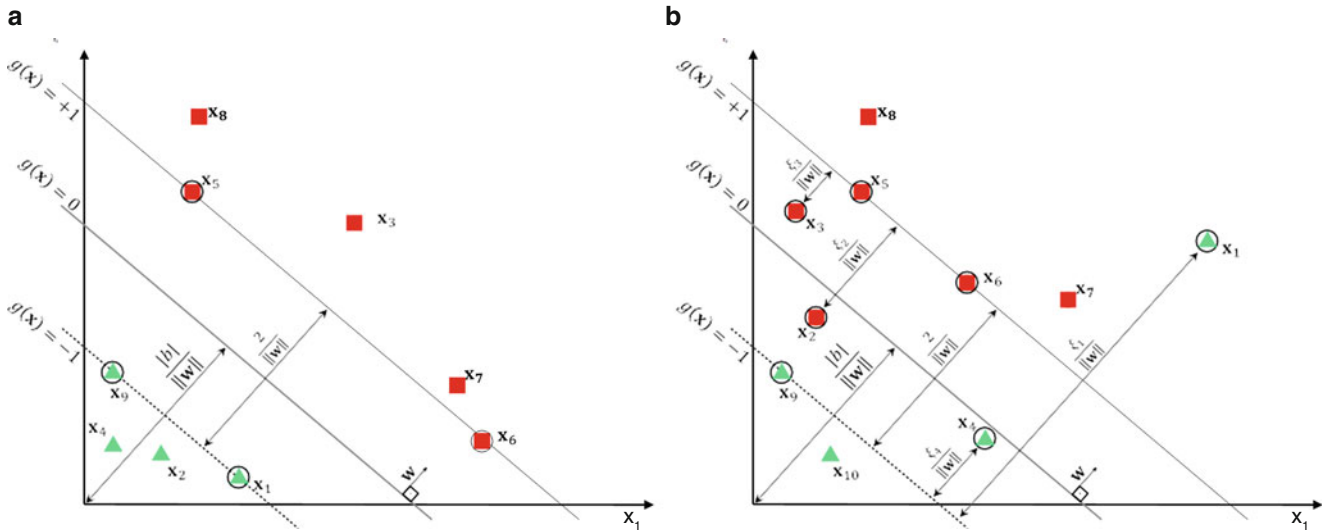
- (i) If $\mathbf{x}_i \in \omega_1$, then $g(\mathbf{x}_i) = \mathbf{w}^T \mathbf{x}_i + b \geq +1 - \xi_i$
- (ii) If $\mathbf{x}_i \in \omega_2$, then $g(\mathbf{x}_i) = \mathbf{w}^T \mathbf{x}_i + b \leq -1 + \xi_i$

Geometrically, the slack variables represent the displacement observed over the wrongly classified data concerning its respective separation margin limit, as shown in Fig. 1b.

The term $C \sum_{i=1}^m \xi_i$ is included in the objective function of Eq. (1) to embrace a cost due to the nonlinear separable data. Such term counts and penalizes the wrong classifications by the separation hyperplane. $C > 0$ is a regularization parameter. Consequently, the optimization problem formulated to deal with nonseparable data differs from Eq. (2) only regarding its constraints; this is:

$$\begin{aligned} \max_{\lambda} & \sum_{i=1}^m \lambda_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \lambda_i \lambda_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ \text{subject to : } &\begin{cases} 0 \leq \lambda_i \leq C, & i = 1, \dots, m \\ \sum_{i=1}^m \lambda_i y_i = 0 \end{cases} \end{aligned} \quad (4)$$

After solving Eq. (4), the parameter $\mathbf{w}$ is computed similarly as made for the linearly separable case. b may be calculated by substituting a support vector with null slack variable into the relations (i) or (ii).



Support Vector Machines, Fig. 1 Representation, in both linearly and nonlinearly separable cases, for the separation hyperplane parameters, margin, and support vectors (circulated). (a) Linearly separable data. (b) Nonlinearly separable data

Kernel Functions

The previous formulations show how the separation hyperplane is determined in the SVM method. Nonetheless, when the data is highly nonlinear, the use of linear surfaces to separate such data do not make sense. Alternatively, the nonlinear data may be remapped from the original attribute space to a new *feature space* where the separation may be better.

Such remapping process may occur implicitly after substitute the inner products $\mathbf{w}^T \mathbf{x}_i$ and $\mathbf{x}_i^T \mathbf{x}_j$ in (3) and (4) by a function $K(\cdot, \cdot)$ which attains the conditions of *Mercer's theorem* (Theodoridis and Koutroumbas 2008):

Mercer's theorem Let $\mathbf{u}, \mathbf{v} \in \mathcal{X} \subseteq \mathbb{R}^n$ and $\phi(\cdot)$ a mapping $\mathbf{u} \mapsto \phi(\mathbf{u}) \in \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space. The inner product $\phi(\mathbf{u})^T \phi(\mathbf{v})$ it is equivalent to a continuous and symmetric function $K(\mathbf{u}, \mathbf{v})$ that satisfies:

$$\int_{\mathcal{X}} \int_{\mathcal{X}} K(\mathbf{u}, \mathbf{v}) h(\mathbf{u}) h(\mathbf{v}) d\mathbf{u} d\mathbf{v} \geq 0 \quad (5)$$

for a given $h(\mathbf{u})$, such that:

$$\int_{\mathcal{X}} h(\mathbf{u})^2 d\mathbf{u} < +\infty \quad (6)$$

As consequence of this theorem, for any function $K(\cdot, \cdot)$ that satisfies (5) and (6), it is possible to affirm there is a vector space where $K(\cdot, \cdot)$, a so-called kernel function, represents an inner product.

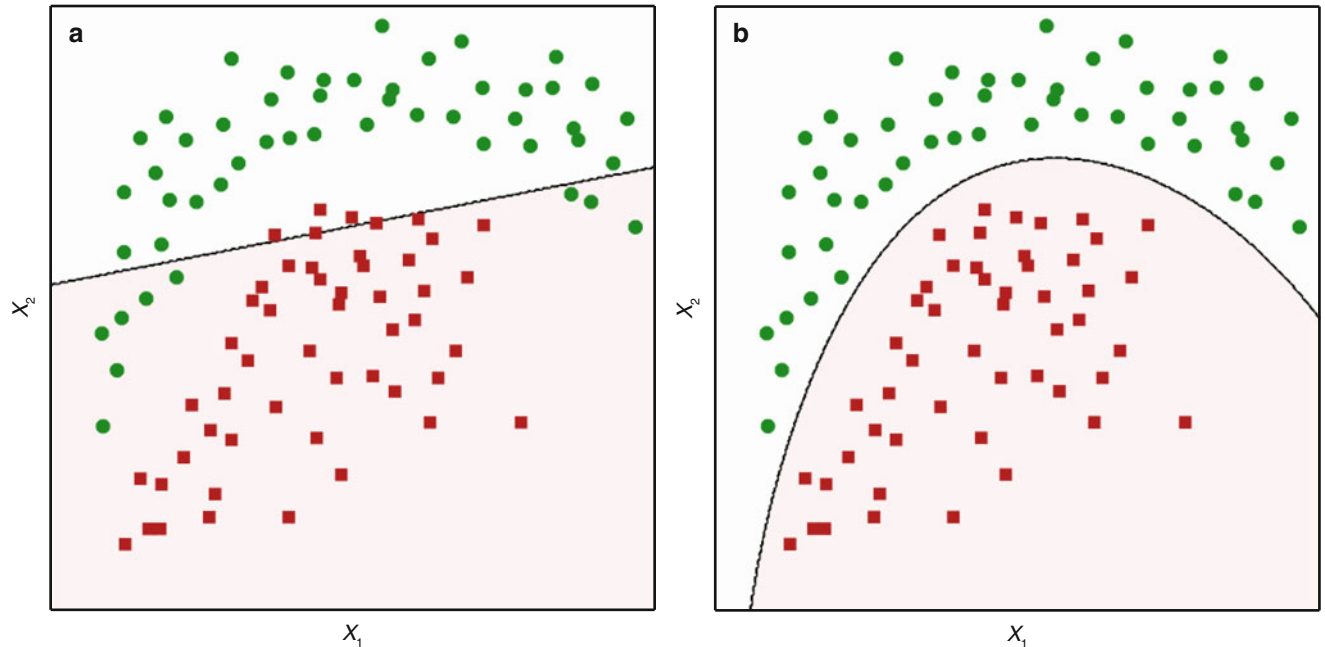
Although the Mercer's theorem gives sufficient condition to define functions that stands for an inner product into a vector space, this space is unknown. Consequently, it is not possible to define $\phi(\cdot)$ from $K(\cdot, \cdot)$. Moreover, check if a given function $K(\cdot, \cdot)$ attains the conditions of Mercer's theorem may comprise a nontrivial task (Schölkopf and Smola 2002).

Examples of frequently adopted kernels are the radial basis function (RBF) ($K_{\text{RBF}}(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2}$), Polynomial ($K_{\text{Pol}}(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i^T \mathbf{x}_j + \alpha)^q$) and Sigmoid ($K_{\text{sigm}}(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma \mathbf{x}_i^T \mathbf{x}_j + \alpha)$) functions, where $q \in \mathbb{N}^*$ and $\gamma, \alpha \in \mathbb{R}_+^*$ are parameters. The usual inner product between vectors is referred as linear kernel function ($K_{\text{Linear}}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$).

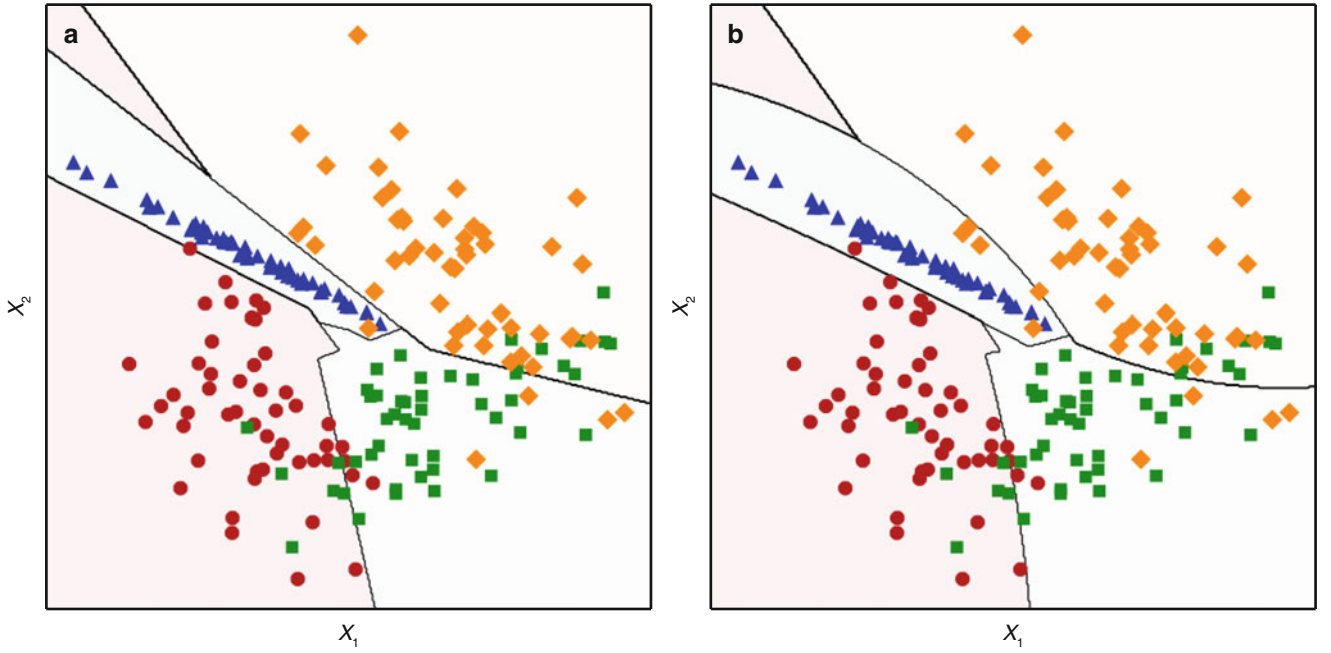
As a reference, Fig. 2 depicts the separation surfaces defined by the Linear and RBF kernel functions when applied to a nonlinear separable training set. The flexibility provided by the RBF kernel delivers a better adherence to the data.

Multiclass Strategies

According to the previous discussions, the classification process was restricted to a pair of classes. However, practical problems may involve a larger number of classes. Multiclass strategies are usually employed to deal with such a restriction.



Support Vector Machines, Fig. 2 Separation hyperplanes using linear and RBF kernels. (a) Linear. (b) RBF



Support Vector Machines, Fig. 3 Decision regions using different multiclass strategies and kernel functions. (a) Linear ($C = 100$) – OVO. (b) RBF ($C = 100, \gamma = 0.5$) – OVR

Among several proposes found in the literature, the one-versus-rest (OVR) and one-versus-one (OVO) are the most commons multiclass strategies.

The OVR strategy, supposing a classification problem involving c classes, defines c separation hyperplanes. Each hyperplane is designed to separate one “reference” class from the other classes. Consequently, the final decision about the class of a given vector is made according to the hyperplane the delivers a classification concerning its reference class with the biggest vector-to-hyperplane distance.

On the other hand, for the OVO strategy, a problem with c classes unfolds into $c(c - 1)/2$ hyperplanes (i.e., the number of combinations of c elements taken 2 at a time) responsible for separating a pair of classes. Since the vectors may be assigned to the same class several times, the final decision is made according to the most assigned class (i.e., a majority voting scheme).

Figure 3 depicts the decision regions defined through a training set that simulates a multiclass classification problem. Distinct strategies and kernels tend to imply different decision regions.

Regression with SVM

Intuitive changes in the initial formulation may adapt the SVM method for data regression purpose. Given the set of observations $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\} \subset \mathcal{X} \times \mathbb{R}$, the following relations are stated:

$$\begin{aligned} \text{(iii)} \quad & (\mathbf{w}^T \mathbf{x}_i + b) - y_i \leq \epsilon + \xi_i, \\ \text{(iv)} \quad & y_i - (\mathbf{w}^T \mathbf{x}_i + b) \leq \epsilon + \hat{\xi}_i, \end{aligned}$$

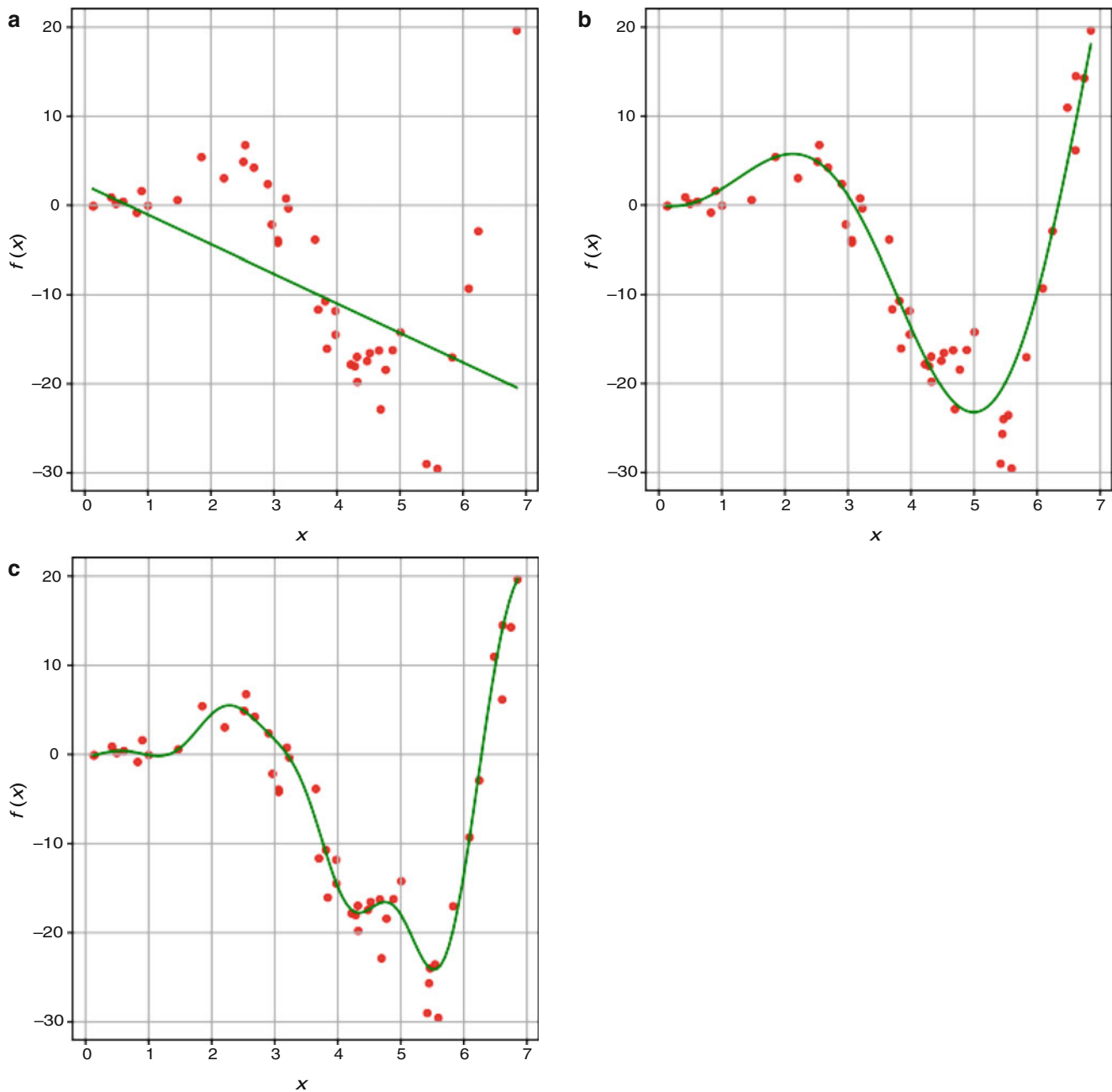
where $\xi_i, \hat{\xi}_i \geq 0$ are slack variables that control the positive and negative divergences with amplitude above to $\epsilon \in \mathbb{R}_+$.

As consequence of (iii) and (iv), $g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$ comprises a regression model that when applied to $\mathbf{x}_i$ should reveal a deviation from the expected result y_i . The computing of ξ_i and $\hat{\xi}_i$, for $i = 1, \dots, m$, that provides the smallest deviation between $g(\mathbf{x}_i)$ and y_i is obtained by:

$$\begin{aligned} \max_{\lambda_i, \hat{\lambda}_i} \quad & \left\{ -\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m (\lambda_i - \hat{\lambda}_i) (\lambda_j - \hat{\lambda}_j) \mathbf{x}_i^T \mathbf{x}_j + \right. \\ & \left. \sum_{i=1}^m (\hat{\lambda}_i - \lambda_i) y_i - \epsilon \sum_{i=1}^m (\hat{\lambda}_i + \lambda_i) \right\} \\ \text{subject to:} \quad & \begin{cases} \sum_{i=1}^m (\lambda_i - \hat{\lambda}_i) = 0 \\ 0 \leq \lambda_i, \hat{\lambda}_i \leq C/m; i = 1, \dots, m \end{cases} \end{aligned} \quad (7)$$

where λ_i and $\hat{\lambda}_i$ are Lagrange multipliers.

The solution from Eq. (7) allows defining $\mathbf{w} = \sum_{i=1}^m (\hat{\lambda}_i - \lambda_i) \mathbf{x}_i$. Consequently, the regression model g it is rewritten as $g(\mathbf{x}) = \sum_{i=1}^m (\hat{\lambda}_i - \lambda_i) \mathbf{x}_i^T \mathbf{x} + b$, where b may be estimated as the average of $|g(\mathbf{x}_i - y_i)| < \epsilon$ considering only the pairs $(\mathbf{x}_i, y_i)$ such that $0 < \lambda_i < C$ or $0 < \hat{\lambda}_i < C$.



Support Vector Machines, Fig. 4 Regression models provided by SVM and distinct kernel functions and parameters. (a) Linear ($C = 100$). (b) RBF ($C = 100, \gamma = 0.1$). (c) RBF ($C = 100, \gamma = 1.0$)

Furthermore, it is worth to mention that the product $\mathbf{x}_i^T \mathbf{x}$ in g may be substituted by a convenient kernel function.

Considering a data set simulated by $f(x) = x^2 \sin(x) + x\zeta$, with $\zeta \sim \mathcal{N}(0, 1)$ and $x \in \mathbb{R}$, Fig. 4 exhibits regression models achieved using different kernels and parameters. A linear approximation is obtained using the Linear kernel. More adherent adjusts are provided using the RBF kernel, where it is also possible identifying the influence of γ parameter in the final models.

Quadratic Optimization, Parametrization, and the Computational Cost

The SVM method shows the advantages in comparison to other methods like those based on the gradient descent or adjust of probability distributions. In such examples, the learning process may be sensitive to initialization parameters that lead to suboptimal results. Conversely, the learning process of the SVM method is characterized by the optimization

of convex (quadratic) functions, which implies unique and explicit solutions Burges and Crisp (2000).

However, weakness comes from the aforementioned optimization problem, specifically regarding the number of terms in the objective function that increases in quadratic proportion to the number of training vectors, consequently demanding a high computational cost. Such peculiarity may be critical in the face of large training data sets. Over the years, this fact motivated the development and improvement of new algorithms to make the SVM training faster and light. Examples of algorithms designed for this purpose are the *sequential minimal optimization* (SMO) (Platt 1999), *SVM^{Light}* (Joachims 1999), *chunking* (Vapnik and Kotz 2006), and LibSVM (Chang and Lin 2011).

Other factors that influence the computational cost are the adopted penalty (C) and kernel parameters. The use of parameters that yields a low computational cost may impair the learning and generalization abilities. In this context, we may conclude that although the optimization algorithm is a relevant element in the SVM training, and sufficient training data and the selected parameters are also important in the learning process.

Summary

SVM is a theoretically attractive, versatile, and powerful supervised method for data classification and regression.

Cross-References

- Machine Learning
- Pattern Classification
- Regression

References

- Burges CJC, Crisp DJ (2000) Uniqueness of the svm solution. In: Solla SA, Leen TK, Müller K (eds) *Advances in neural information processing systems*, vol 12. MIT Press, pp 223–229. <http://papers.nips.cc/paper/1735-uniqueness-of-the-svm-solution.pdf>
- Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol* 2:27:1–27:27, software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- Cortes C, Vapnik V (1995) Support-vector networks. In: *Machine learning*, Springer, pp 273–297. <https://link.springer.com/article/10.1007/BF00994018#citeas>
- Joachims T (1999) Making large-scale support vector machine learning practical. In: *Advances in kernel methods*. MIT Press, Cambridge, pp 169–184
- Platt JC (1999) Fast training of support vector machines using sequential minimal optimization. MIT Press, Cambridge, pp 185–208
- Schölkopf B, Smola AJ (2002) *Learning with kernels: support vector machines, regularization, optimization, and beyond*, Adaptive computation and machine learning. MIT Press, Cambridge, Massachusetts, USA
- Theodoridis S, Koutroumbas K (2008) *Pattern recognition*, 4th edn. Academic Press, San Diego
- Vapnik V, Kotz S (2006) *Estimation of dependences based on empirical data: empirical inference science (information science and statistics)*. Springer-Verlag, Berlin, Heidelberg
- Webb AR, Copsey KD (2011) *Statistical pattern recognition*, 3rd edn. John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781119952954>

t-Distributed Stochastic Neighbor Embedding

Mehala Balamurali

Australian Centre for Field Robotics, University of Sydney,
Sydney, NSW, Australia

Definition

t-Distributed stochastic neighbor embedding (t-SNE) method is an unsupervised machine learning technique for nonlinear dimensionality reduction to visualize high-dimensional data into low dimensional space. This method was introduced by Van der Maaten and Hinton in 2008. Among the many other dimensionality reduction approaches that are nonlinear methods which project the high-dimensional data into low-dimensional space by keeping nonsimilar data points far apart but are not capable for visualizing the high-dimensional real data, the nonlinear mapping by t-SNE groups the similar datapoints closer and therefore preserves the structure of the high dimensional in the low-dimensional space. Therefore, t-SNE is able to visualize the natural clusters in low dimensional space. Hence t-SNE is considered as a popular visualization tool that supports cluster analysis of high-dimensional data. t-SNE method has been used in many research fields.

Implementation

The t-SNE algorithm was proposed as a variation of stochastic neighbor embedding (SNE) (Hinton and Roweis 2002) to overcome the problem of optimizing the cost function. SNE converts the distance between the similar data points in high dimensional into conditional probabilities.

Given the data points x_i and x_j , the conditional probability, $p_{j|i}$, is given by Eq. 1 such that x_i would pick x_j as its neighbor if neighbors were picked in proportion to their probability

density under a Gaussian centered at x_i (Fig. 1). In Eq. 1, the distance between x_i and x_j is calculated in Euclidean distance, and the σ_i is the variance of the Gaussian that is centered on datapoint x_i . $p_{j|i} = 0$

$$p_{j|i} = \frac{\exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(\frac{-\|x_i - x_k\|^2}{2\sigma_i^2}\right)} \quad (1)$$

The variance, σ^2 , of the Gaussian kernel is calculated using a binary search such that the entropy, $(P_i) = -\sum_j P_{j|i} \log_2 P_{j|i}$, of the probability distribution over all the data points is equal to $\log_2(Perp)$, where *Perp* is the user-specified perplexity (Van der Maaten and Hinton 2008).

In the low-dimensional space, SNE computes a similar conditional probability for map points y_i and y_j (corresponding to the two data points, x_i and x_j) using another Gaussian kernel with variance equaling 1/2

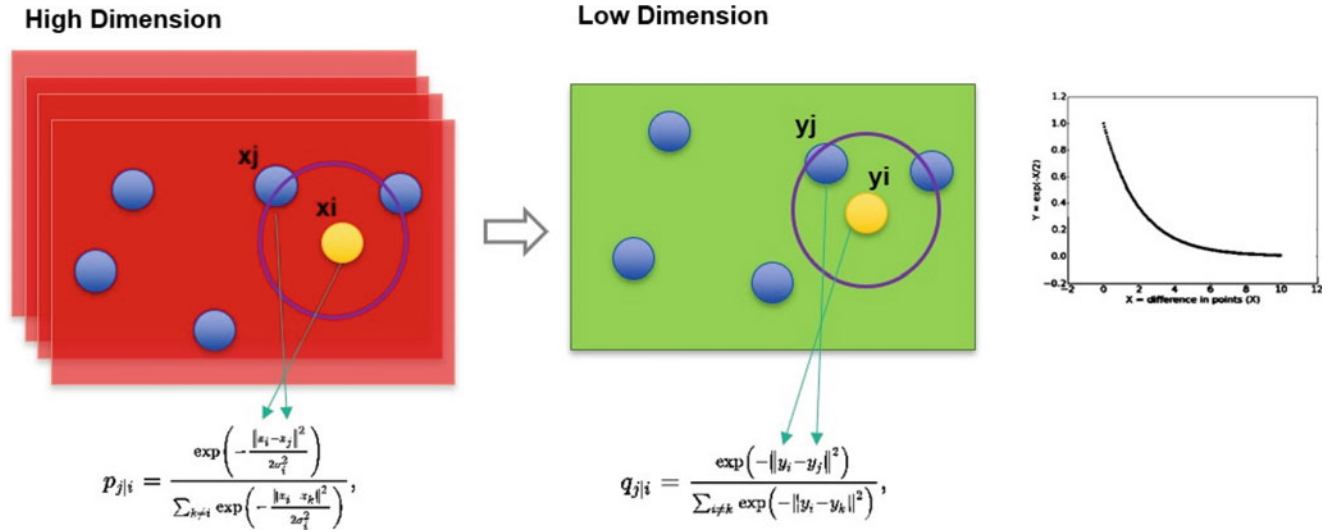
$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)} \quad (2)$$

and $q_{i|i} = 0$ by definition. The conditional probabilities, $p_{j|i}$ and $q_{j|i}$, should be equal; if the map points, y_i and y_j , exactly represent the similarity between the high-dimensional data points, x_i and x_j .

SNE thus arranges map points in the low-dimensional space to minimize the discrepancy between $p_{j|i}$ and $q_{j|i}$ measured by the Kullback-Leibler (KL) divergence considering all data points. The cost function to be minimized is

$$C = \sum_i KL(P_i | Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \quad (3)$$

where P_i is the conditional probability distribution of data point x_i with respect to all other data points and Q_i is the



t-Distributed Stochastic Neighbor Embedding, Fig. 1 Conditional probability of data points in high and low dimensional space

conditional probability distribution of map point y_i over all other map points. As the KL is asymmetric, Eq. 3 uses a large cost to place nearer data points in high dimensional to keep them furthest in low dimensional and use low cost to place widely separated data points in high dimensional to nearer in low dimensional map. The cost function is minimized through various optimization methods such as gradient descent. The gradient of the cost function of SNE is

$$\frac{\delta C}{\delta y_i} = 2 \sum_j (P_{ji} - q_{ji} + P_{ij} - q_{ij}) (y_i - y_j) \quad (4)$$

Furthermore, the momentum used in SNE drives the optimization faster and evade local optima.

$$y^{(t)} = y^{(t-1)} + \eta \frac{\delta C}{\delta y} + \alpha(t) (y^{(t-1)} - y^{(t-2)}) \quad (5)$$

where (t) is the momentum at iteration t ; $y(t)$ is the solution at iteration t ; and η is the learning rate.

t-SNE is introduced to overcome the downsides of SNE; cost function is difficult to optimize and the crowding problem, by symmetrizing the cost function and incorporating student t -distribution to calculate the resemblances of data points in low dimension. Hence the t-SNE is described by the dataspace, map space, and cost function by Eqs. 6, 7, and 8:

$$p_{ij} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_k - x_i\|^2 / 2\sigma_i^2)} \quad (6)$$

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i} (1 + \|y_k - y_i\|^2)^{-1}} \quad (7)$$

$$C = KL(P \parallel Q) = \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (8)$$

The gradients descent is given by

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (P_{ij} - q_{ij}) (y_i - y_j) (1 + \|y_i - y_j\|^2)^{-1} \quad (9)$$

Perplexity, learning rate and number of iterations are some of the hyper parameters that control the clusters in the t-SEN reduced coordinates. Among them perplexity is the most influential factor, and therefore the results of t-SNE are fairly robust to perplexity change (Fig. 2). In Eq. 3, small value and larger values of σ^2 determine the pairs x and x with small distance and large distance, respectively. Though a recommended range for perplexity is 5–50, a frequently used perplexity is 30.

Algorithm 1: Simple version of t-Distributed Stochastic Neighbor Embedding

Data: data set = $\{x_1, x_2, \dots, x_{n1}\}$,

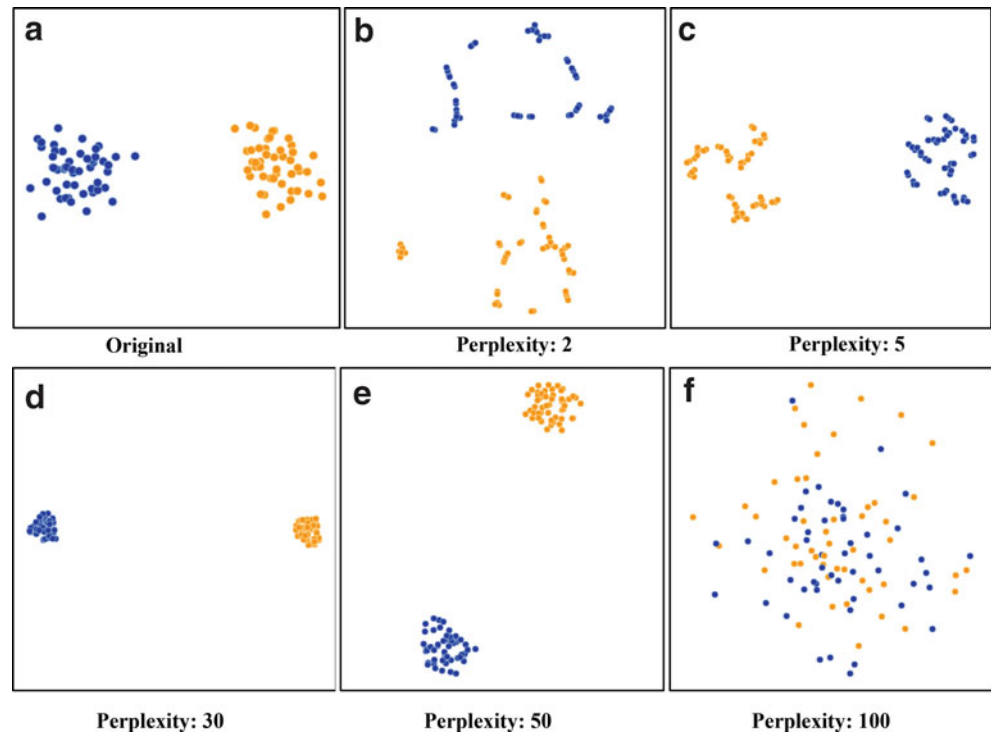
cost function parameters: perplexity $Perp$,

optimization parameters: number of iterations T , learning rate η , momentum $\alpha(t)$.

Result: low-dimensional data representation $Y^{(T)} = \{y_1, y_2, \dots, y_n\}$.

t-Distributed Stochastic Neighbor Embedding,

Fig. 2 (a) The original clusters and plots (b–f) show the impact of perplexity on resulting embedding with the chosen perplexity values 2, 5, 30, 50, and 100, respectively. Each plot was produced by 5000 runs with epsilon 10 (Wattenberg et al. 2016)



Begin
 compute pairwise affinities p_{ij} with perplexity $Perp$ (using Eq. 1)
 set $p_{ij} = \frac{p_{ij} + p_{ji}}{2n}$
 sample initial solution $Y^{(0)} = \{y_1, y_2, \dots, y_n\}$ from $N(0, 10^{-4}I)$
for $t=1$ **to** T **do**
 compute low-dimensional affinities q_{ij} (using Eq. 7)
 compute gradient $\frac{\partial C}{\partial y}$ (using Eq. 9)
 set $y^{(t)} = y^{(t-1)} + \eta \frac{\partial C}{\partial y} + \alpha(t)(y^{(t-1)} - y^{(t-2)})$
end
end

Algorithm Vander Maaten and Hinton (2008)

Application of t-SNE in Cluster Delineation

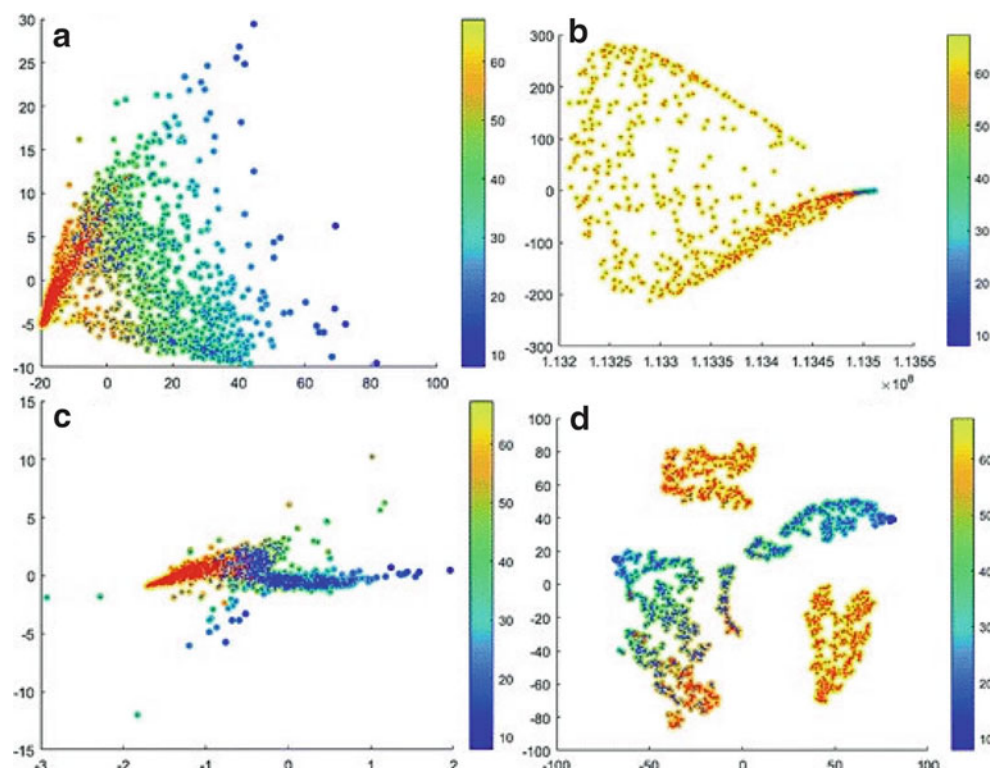
Application of t-SNE was widely studied in biology, and the results showed a more accurate representation of multidimensional data when it is embedding into low dimensional t-SNE coordinates. It is the key technique used in two tools: viSNE and automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE), for analyzing mass cytometry data in

immunology (Amir et al. 2013; Shekhar et al. 2014). t-SNE-based algorithm is found as one of the techniques that can better represent the similarities between the cell type clusters (Saeys et al. 2016). In recent years t-SNE has become a popular method for visual exploration of scRNA-seq data because t-SNE can create well separated clusters in 2D coordinates of the high-dimensional data that can be well clustered (McInnes et al. 2018; Becht et al. 2019). In 2019, Aizarani et al. (2019) and Dobie et al. (2019) used t-SNE to highlight the important liver cell partitions by mapping the single-cell transcriptomes from liver tissue. Kobak and Berens (2019) showed that the initialization with PCA and larger perplexity preserves the macroscopic structure and mesoscopic structure of the gene data, respectively. Hyper-spectral images with the thousands of spatial-spectral dimensionalities often suffer from unwanted feature spectra that correspond to the same ground truth. t-SNE was successfully applied in this domain to map a nonlinear representation of spectral features in a lower 2D space (Pouyet et al. 2018, Zhang et al. 2018).

In 2016, t-SNE was introduced to embed mineral and waste exploration datasets into 2D t-SNE coordinate (Balamurali and Melkumyan 2006). The study showed that t-SNE can preserve more information in the data compared to other dimensionality methods PCA, LLE, and kPCA (Fig. 3). In addition, the authors applied spectral clustering on t-SNE

t-Distributed Stochastic Neighbor Embedding,

Fig. 3 Two-dimensional visualization on mineral and waste exploration dataset. The (d) 2D t-SNE embedding shows more distinct clusters than the (a) PCA, (b) kPCA and (c) LLE. Color bars show the Fe distribution. Blue and red dots show the waste and



coordinates to extract clusters using the t-SNE coordinates and the chemical distribution within each cluster showed meaningful geological domains (Fig. 4).

Later, this method was compared with other clustering methods, self-organizing map (SOM) and spanning-tree progression analysis of density-normalized events (SPADE), to study the material grouping (Balamurali et al. 2019) and to group potential distinct material domains (Figs. 5 and 6). Horrocks et al. in 2019 used t-SNE to study the process of rock hydration using geochemical data. In this study the authors compared the embedding results by t-SNE using all elements and subset of elements (selected features) and demonstrated that t-SNE using subset of features outperformed in differentiating altered and unaltered specimens. In 2021 t-SNE was employed for understanding spatial and temporal patterns of groundwater geochemistry (Liu et al. 2021). The results obtained using t-SNE on chemical data were encouraging. Overall, t-SNE was found to be capable of producing geochemical embeddings wherein external geological properties were reflected in the cluster or intra-cluster structure.

A novel technique EC-tSNE (ensemble clustering based t-distributed stochastic neighbor embedding) was proposed to minimize stochastic variation in the standard

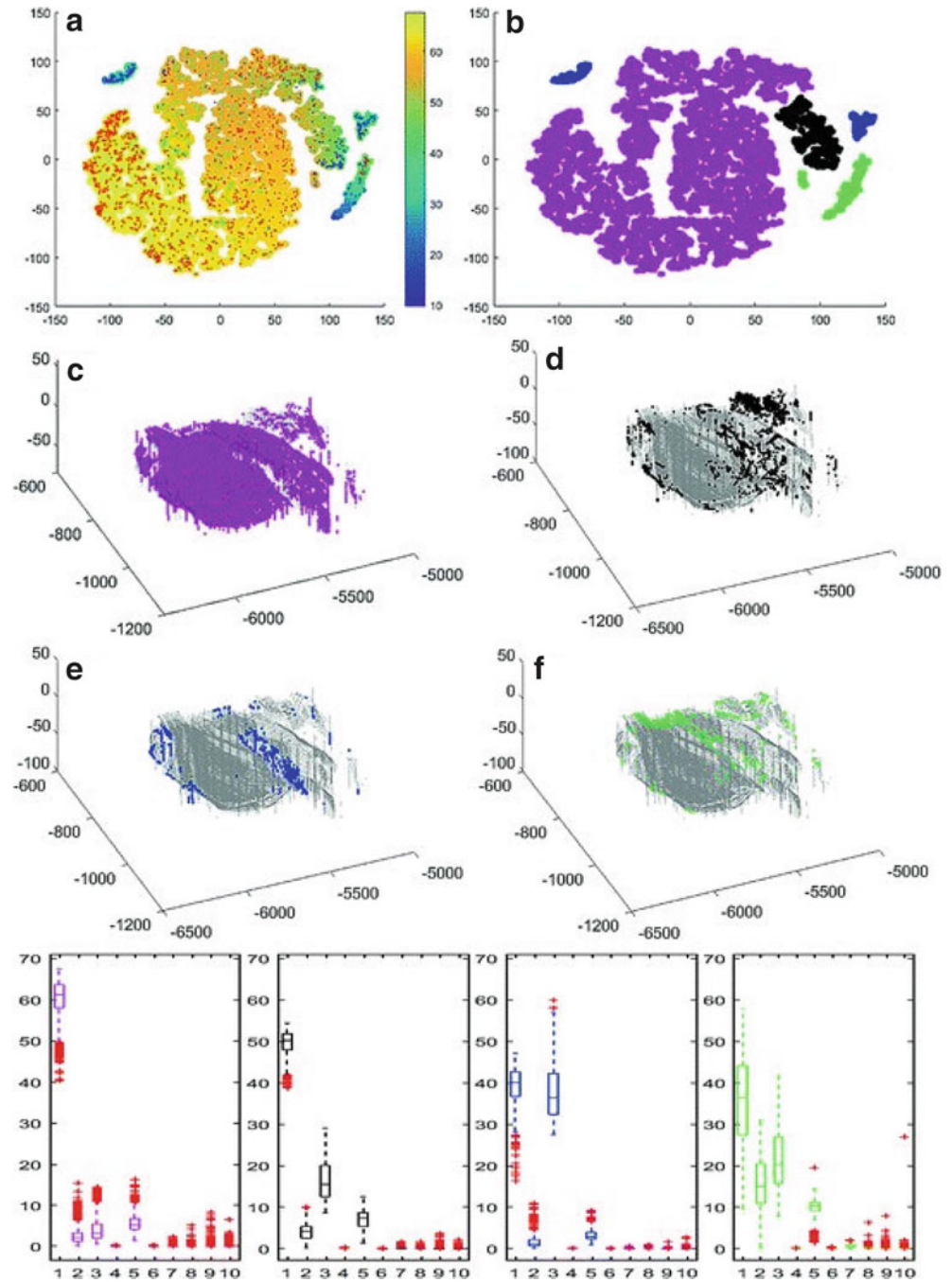
t-SNE approach (Balamurali and Melkumyan 2020) and therefore consistently identified sub-geological regions that they were not previously known (Fig. 7). This study used cluster number optimization which measures the consistency of samples within cluster. The proposed approach also used the agglomerative clustering which increase the reliability of the results obtained from t-SNE and spectral clustering by reducing randomness. Hence the complete system seeks to find the most stable or dominant cluster. A recent study compared EC-tSNE with the robust estimator approach called MCD (minimum covariance determinant) in terms of their efficacy for outlier removal in multivariate geochemical data (Leung et al. 2019).

Summary

t-SNE is popularly used in high-parameter gene expression studies to understand the diversity of cell population and visualize the subpopulations in low-dimensional space. Recently this has emerged as a state-of-the-art method that has been adopted in other domains such as spatial zone delineation in geology and hydrology, biomedical signal

t-Distributed Stochastic Neighbor Embedding,

Fig. 4 Visualization of assay data on t-SNE coordinates. Color bar shows the Fe distribution in plot (a). Blue, red, and green dots show the waste, mineral, and hydrated samples, respectively. (b) Spectral clustering results on the reduced 2D using t-SNE. (c), (d), (e) and (f) Corresponding clustering on their transformed spatial coordinates of production and exploration data. Box plots show the chemical distribution corresponding to each cluster in the order 1–10: Fe, Al₂O₃, SiO₂, P, LOI, S, TiO₂, MgO, CaO, and Mn (Balamurali and Melkumyan 2016)

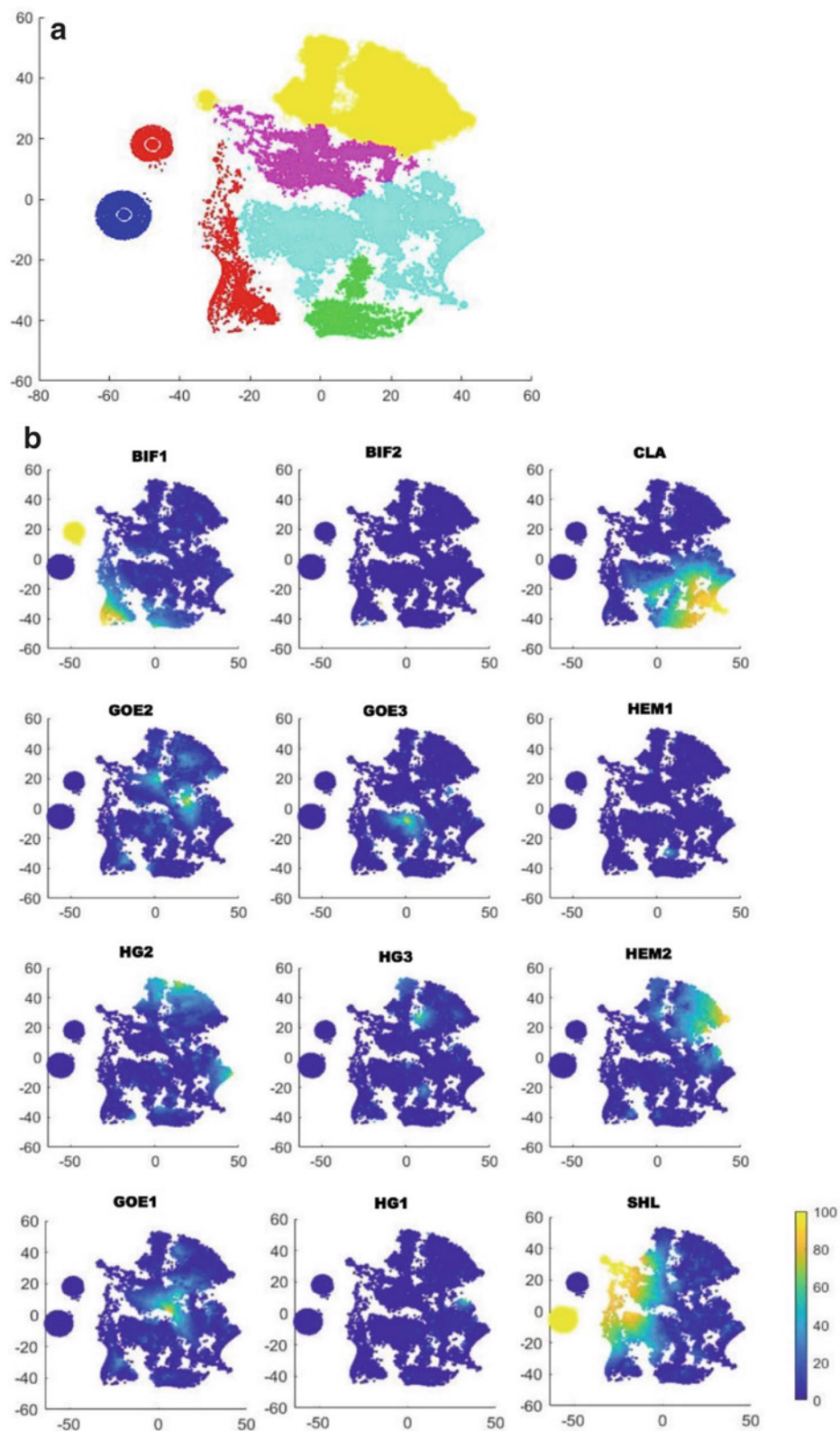


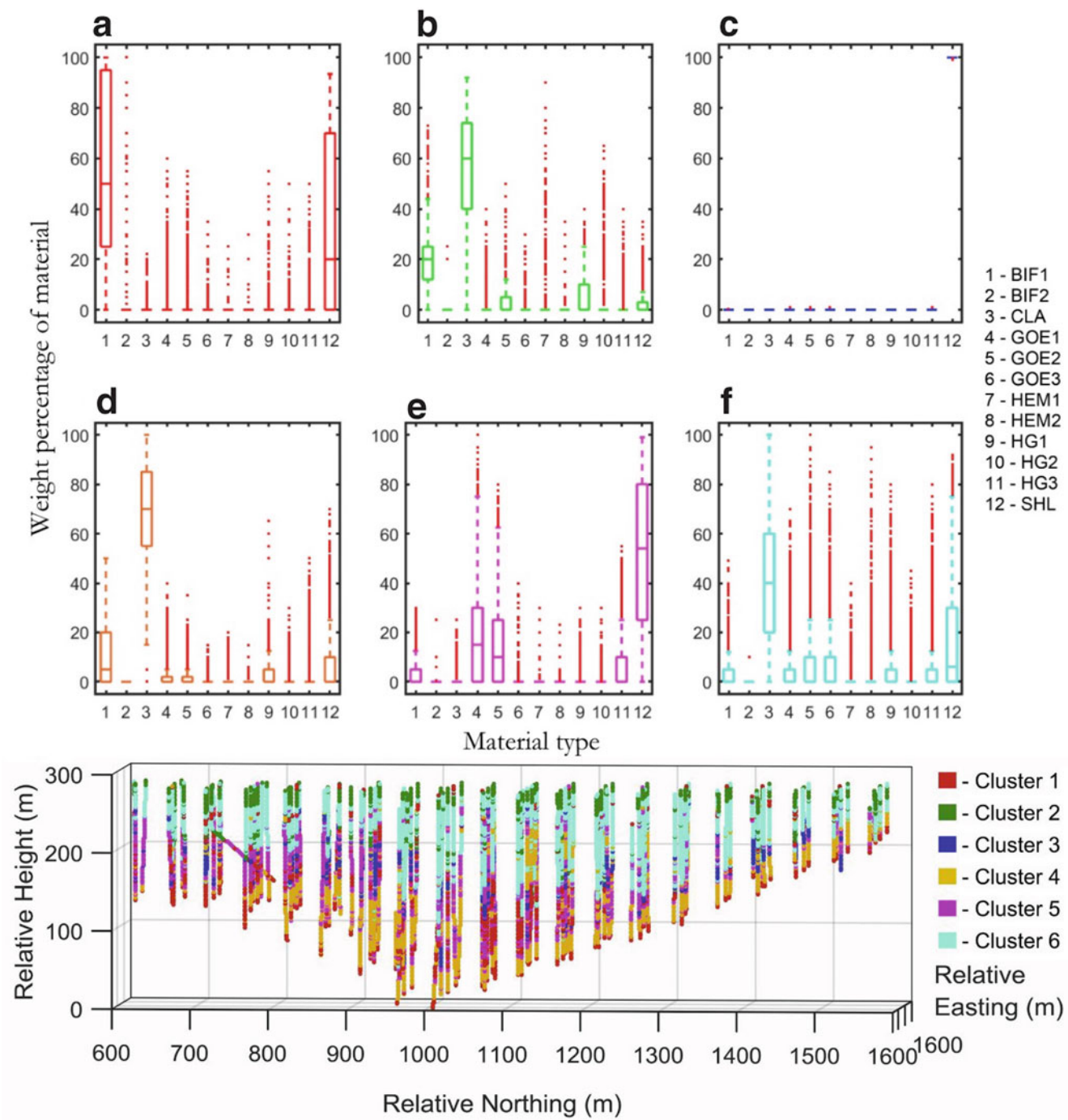
analysis, natural language processing, and music analysis. While t-SNE can be used for visual interpretation of the clusters presents in the high-dimensional data in the low dimension, the method cannot be used as a standalone method

for cluster extraction. Recent studies have proposed how this can be achieved as an ensemble approach by incorporating other machine learning techniques. Further studies are needed to optimize the t-SNE results.

t-Distributed Stochastic Neighbor Embedding,

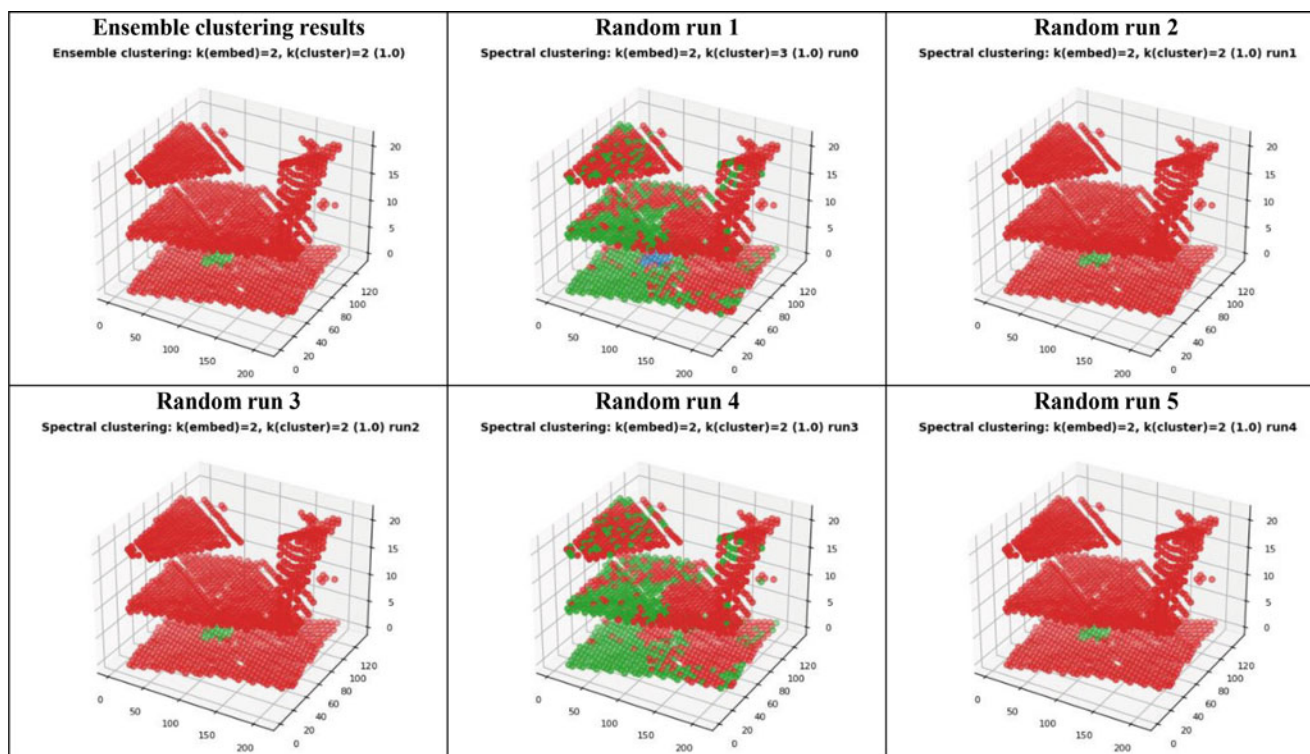
Fig. 5 (a) t-SNE 2D plot showing six clusters corresponding to potentially distinct material domains on t-SNE1 and t-SNE2 coordinates, (b) subplots show the contributions of different features to the clusters seen on the top 2D map. The color scale represents the percentage of the material type located at that location (Balamurali et al. 2019)





t-Distributed Stochastic Neighbor Embedding, Fig. 6 The box plots (a–f) show the distribution of material types within the clusters from the t-SNE. The color represents the corresponding cluster color in

the 2D t-SNE (Fig. 5a). (g) The 3D plot is a cross-section of the deposit, showing points along the drill holes, grouping according to t-SNE grouping. (Balamurali et al. 2019)



t-Distributed Stochastic Neighbor Embedding, Fig. 7 Results of EC-tSNE: ability to extract stable clusters despite random variability from t-SNE spectral clustering. The plots show the clusters results

colored in the spatial coordinates (Picture credit: Raymond Leung) (Balamurali and Melkumyan 2020)

Cross-References

- [Compositional Data](#)
- [Data Visualization](#)
- [Exploratory Data Analysis](#)
- [Machine Learning](#)
- [Mine Planning](#)
- [Multivariate Analysis](#)
- [Statistical Outliers](#)

Bibliography

- Aizarani N, Saviano A, Sagar M, L., Durand, S., Herman, J.S., Pessaux, P., Baumert, T.F., Grun, D. (2019) A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature* 572(7768):199–204
- Amir AD, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, Shenfeld DK, Krishnaswamy S, Nolan GP, D. Pe'er. (2013) viSNE enables visualization of high dimensionalsingle-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol* 31:545–552
- Balamurali M, Melkumyan A (2016) t-SNE based visualisation and clustering of geological domain. In: Hirose A, Ozawa S, Doya K, Ikeda K, Lee M, Liu D (eds) *Neural information processing. ICONIP 2016. Lecture notes in computer science*, vol 9950. Springer, Cham. https://doi.org/10.1007/978-3-319-46681-1_67
- Balamurali M, Melkumyan A (2020) Computer Aided sub-domain detection using t-SNE incorporating cluster ensemble for improved mine modelling. *Math Geosci* (under review)
- Balamurali M, Silversides KL, Melkumyan A (2019) A comparison of t-SNE, SOM and SPADE for identifying material type domains in geological data. *Comput Geosci*. <https://doi.org/10.1016/j.cageo.2019.01.011>
- Becht E et al (2019) Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* 37:38
- Dobie R, Wilson-Kanamori JR, Henderson BEP, Smith JR, Matchett KP, Portman JR, Wallenborg K, Picelli S, Zagorska A, Pendem SV, Hudson TE, Wu MM, Budas GR, Breckenridge DG, Harrison EM, Mole DJ, Wigmore SJ, Ramachandran P, Ponting CP, Teichmann SA, Marioni JC, Henderson NC (2019) Single-cell transcriptomics uncovers zonation of function in the mesenchyme during liver fibrosis. *Cell Rep* 29(7):1832–1847
- Hinton GE, Roweis ST (2002) *Stochastic neighbor embedding, advances in neural information processing systems*. The MIT Press, Cambridge, MA, pp 833–840
- Horrocks T, Holden EJ, Wedge D, Wijns C, Fiorentini M (2019) Geochemical characterisation of rock hydration processes using t-SNE. *Comput Geosci* 124:46–57
- Kobak K, Berens P (2019) The art of using t-SNE for single-cell transcriptomics. *Nat Commun* 10(1):5416
- Leung R, Balamurali M, Melkumyan A (2019) Sample truncation strategies for outlier removal in geochemical data: the MCD robust distance approach versus t-SNE ensemble clustering. *Math Geosci*. <https://doi.org/10.1007/s11004-019-09839-z>
- Liu H, Yang J, Ye M, James SC, Tang Z, Dong J, Xing T (2021) Using t-distributed Stochastic Neighbor Embedding (t-SNE) for cluster

- analysis and spatial zone delineation of groundwater geochemistry data. *J Hydrol* 597:126146., ISSN 0022-1694. <https://doi.org/10.1016/j.jhydrol.2021.126146>
- McInnes L, Healy J, Melville J (2018) UMAP: uniform manifold approximation and projection for dimension reduction. <https://arxiv.org/abs/1802.03426>. Return to ref 10 in article
- Pouyet E, Rohani N, Katsaggelos AK, Cossairt O, Walton M (2018) Innovative data reduction and visualization strategy for hyperspectral imaging datasets using t-SNE approach. *Pure Appl Chem* 90(3): 493–506. <https://doi.org/10.1515/pac-2017-0907>
- Saeyns Y, Van Gassen S, Lambrecht BN (2016) Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nat Rev Immunol* 16:449–462. <https://doi.org/10.1038/nri.2016.56>
- Shekhar K, Brodin P, Davis MM, Chakraborty AK (2014) Automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE). *Proc Natl Acad Sci U S A* 111:202–207
- Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9(11):2579–2605
- Wattenberg et al (2016) How to use t-SNE effectively. *Distill*. <https://doi.org/10.23915/distill.00002>
- Zhang J, Chen L, Zhuo L, Liang X, Li J (2018) An efficient hyperspectral image retrieval method: deep spectral-spatial feature extraction with DCGAN and dimensionality reduction using t-SNE-based NM hashing. *Remote Sens* 10:271. <https://doi.org/10.3390/rs10020271>

Text Mining

Chengbin Wang¹ and Xiaogang Ma²

¹State Key Laboratory of Geological Processes and Mineral Resources & School of Earth Resources, China University of Geosciences, Wuhan, China

²Department of Computer Science, University of Idaho, Moscow, ID, USA

Definition

Text mining in geoscience is an emerging research area that uses methods of natural language processing and machine learning to extract structured information from the unstructured geoscience literature and makes semantic reasoning for knowledge discovery in geosciences. Research topics of text mining include, but are not limited to, text clustering, word segmentation, document summarization, entity and semantic relation extraction, knowledge base, and graph.

Text Mining and Mathematical Geosciences

Geoscience is a knowledge-intensive discipline. It contains large amounts of numeric spatial dataset (e.g., remote sensing, geochemistry, and geophysics) and map dataset. The rapid development of detection techniques in the micro- and macroscales in the past decades has been increasing both the

volume and quality of geoscience data. The mathematical geoscience is to apply mathematical methods and computers for processing these spatial geoscience data for knowledge discovery, such as mantle convection (Iwamori et al. 2010), mineral predication, discovery of new minerals (Morrison et al. 2017), geologic modeling, and lithofacies identification.

Moreover, the numeric dataset and map dataset are not the only outputs of geoscience research. Important information and knowledge produced by geoscience researches are recorded in unstructured textual form and thus hidden in the big data of geological literature (Wang and Ma 2019). Gil et al. (2018) proposed a research agenda of intelligent systems for geosciences that will result in fundamental new capabilities for understanding the Earth system. Automated information extraction and integration from published literature is listed as a key research direction in the agenda. In the era of big data, geoscience literature has become a big “mineral resource” for data mining. The improved computing performance including software and hardware enables an opportunity to understand the evolution of Earth system using text mining from geoscience literature (Wang and Ma 2019). The text mining for geoscience literature has become a new research direction and hotspot in the mathematical geosciences.

Text Mining for Geoscience Literature

Geoscience literature not only shares some common characteristics with other types of literature, but also is characterized by the spatial information and multidisciplinary terminology. Text mining for geoscience literature must depend on the geoscience knowledge. Therefore, building domain ontology and vocabulary of geosciences is a basic step for text mining. The ontology model determines which entities and semantic relationship information to extract from the geoscience literature. In the field of geoinformatics, a series of geological ontology models have been designed for different purpose, such as geologic time (Cox and Richard 2015; Ma et al. 2011, 2012; Wang et al. 2018a), geological modeling (Perrin et al. 2005), geological map (Mantovani et al. 2020), SWEET (Raskin 2006), mineral exploration (Ma et al. 2010), and geological structure (Zhong et al. 2009).

The first step of text mining is collecting and preprocessing text data. There are many forms for users to get the text data. For example, API permission from an online database, document scanning, and data extracting from Web. The obtained data from different sources may be recoded in diverse formats, such as text files and scanned images. It is necessary to transform the data into an organized, computer-readable format (Wang and Ma 2019). Due to the ambiguity, polysemy, and irregular input, it is difficult for computers to read and

understand the text data of natural language. Therefore, it is necessary to segment a piece of geological text into semantic word sequences for further computer processing (Wang and Ma 2019).

After geoscience literature data is structured, different topics including text clustering, document summarization, entity and semantic relation extraction, knowledge base, and graph can be studied. However, many geoscientists prefer to extract entities and their semantic links to build domain knowledge base and knowledge graph (Qi et al. 2020). Knowledge graph is a semantic network with directed graph structure. Each node in a graph represents an entity, and an edge represents the linking relationship between two entities (Wang et al. 2018b). The graph visualizes the nodes and the links between nodes to represent the information network and semantic relationships in a specific domain. The knowledge base is data pool to store knowledge graph in the form of triples. An informative geoscience knowledge base and knowledge graph can provide an infrastructure for geoscience scientists to change scientific research paradigm and be used to explore knowledge discovery for some big science questions such as origin of life and coevolution by the way of semantic reasoning.

Traditional statistical methods are unable to process geoscience literature. The information extraction from geoscience literature depends on the machine learning method. The machine learning methods of CNN (Convolutional Neural Networks), LSTM (Long short-term memory), and CRF (Conditional random field) have been applied in information extraction from geological literature (Qiu et al. 2018; Li et al. 2018; Wang et al. 2018b). The development in interpreted programming language and the wide-spreading open-source libraries enable scholars in various disciplines to quickly learn the latest algorithms and apply them to their domain-specific researches (Wang and Ma 2019). There are many widely used open and free libraries in text mining, such as TensorFlow (Abadi et al. 2016), DeepDive (De Sa et al. 2016), Caffe (Jia et al. 2014), CNTK (Seide and Agarwal, 2016), and MXnet (Chen et al. 2015). Even if a researcher has only the basic skills in programming, he or she will be able to make a good application of text mining using these libraries.

Text mining is still a new research direction in geoscience and has been used to assist to database construction and knowledge discovery. PaleoDeepDive, a machine reading system, was developed to extract fossil information from literature. The extracted results have been used to update the PBDB database (Peters et al. 2014). The GeoDeepDive, the updated version of PaleoDeepDive, is still ongoing and extracts geological unit information in North America from the geological literature and build the Macrostrata (<https://macrostrat.org/>).

Summary

Although text mining is an emerging research direction for geosciences, the research outputs have proven it is a powerful tool to transform earth science by providing new research tool, new data sources, and research views in the era of big data and artificial intelligence. This research direction requires the attention and involvement of more computer scientists and earth scientists.

Cross-References

- [Artificial Intelligence in the Earth Sciences](#)
- [Data Mining](#)
- [Machine Learning](#)

Bibliography

- Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M (2016) Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:160304467
- Chen T, Li M, Li Y, Lin M, Wang N, Wang M, Xiao T, Xu B, Zhang C, Zhang Z (2015) Mxnet: a flexible and efficient machine learning library for heterogeneous distributed systems. arXiv preprint arXiv:151201274
- Cox SJ, Richard SM (2015) A geologic timescale ontology and service. *Earth Sci Inf* 8(1):5–19
- De Sa C, Ratner A, Ré C, Shin J, Wang F, Wu S, Zhang C (2016) Deepdive: declarative knowledge base construction. *ACM SIGMOD Rec* 45(1):60–67
- Gil Y, Pierce SA, Babaie H, Banerjee A, Borne K, Bust G, Cheatham M, Ebert-Uphoff I, Gomes C, Hill M (2018) Intelligent systems for geosciences: an essential research agenda. *Commun ACM* 62(1): 76–84
- Iwamori H, Albarède F, Nakamura H (2010) Global structure of mantle isotopic heterogeneity and its implications for mantle differentiation and convection. *Earth Planet Sci Lett* 299(3–4):339–351
- Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: convolutional architecture for fast feature embedding. In: *Proceedings of the 22nd ACM international conference on multimedia*, pp 675–678
- Li S, Chen J, Xiang J (2018) Prospecting information extraction by text mining based on convolutional neural networks—a case study of the Lala copper deposit, China. *IEEE Access* 6:52286–52297
- Ma X, Wu C, Carranza EJM, Schetselaar EM, van der Meer FD, Liu G, Wang X, Zhang X (2010) Development of a controlled vocabulary for semantic interoperability of mineral exploration geodata for mining projects. *Comput Geosci* 36(12):1512–1522
- Ma X, Carranza EJM, Wu C, Van Der Meer FD, Liu G (2011) A SKOS-based multilingual thesaurus of geological time scale for interoperability of online geological maps. *Comput Geosci* 37(10):1602–1615
- Ma X, Carranza EJM, Wu C, van der Meer FD (2012) Ontology-aided annotation, visualization, and generalization of geological time-scale information from online geological map services. *Comput Geosci* 40: 107–119
- Mantovani A, Lombardo V, Piana F (2020) Ontology-driven representation of knowledge for geological maps. *Comput Geosci* 139: 104446

- Morrison SM, Liu C, Eleish A, Prabhu A, Li C, Ralph J, Downs RT, Golden JJ, Fox P, Hummer DR (2017) Network analysis of mineralogical systems. Mineralogical Society of America, Washington, DC
- Normile D (2019) Earth scientists plan a geological Google China is backing an international effort to meld geoscience databases into a one-stop data shop. *Science* 363(6430):917–917
- Perrin M, Zhu B, Rainaud J-F, Schneider S (2005) Knowledge-driven applications for geological modeling. *J Pet Sci Eng* 47(1–2):89–104
- Peters SE, Zhang C, Livny M, Ré C (2014) A machine reading system for assembling synthetic paleontological databases. *PLoS One* 9(12)
- Qi H, Dong S, Zhang L, Hu H, Fan H (2020) Construction of earth science knowledge graph and its future perspectives. *Geol J China Univ* 26(1):02–10
- Qiu Q, Xie Z, Wu L, Li W (2018) DGeoSegmenter: a dictionary-based Chinese word segmenter for the geoscience domain. *Comput Geosci* 121:1–11
- Raskin R (2006) Development of ontologies for earth system science. *Spec Pap-Geol Soc Am* 397:195
- Seide F, Agarwal A (2016) CNTK: Microsoft's open-source deep-learning toolkit. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp 2135–2135
- Wang C, Ma X (2019) Text mining to facilitate domain knowledge discovery. In: *Text mining-analysis, programming and application*. IntechOpen. <https://www.intechopen.com/online-first/text-mining-to-facilitate-domain-knowledge-discovery>. Accessed 8 April 2019. © 2019 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution 3.0 License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited
- Wang C, Ma X, Chen J (2018a) Ontology-driven data integration and visualization for exploring regional geologic time and paleontological information. *Comput Geosci* 115:12–19
- Wang C, Ma X, Chen J, Chen J (2018b) Information extraction and knowledge graph construction from geoscience literature. *Comput Geosci* 112:112–120
- Zhong J, Aydina A, McGuinness DL (2009) Ontology of fractures. *J Struct Geol* 31(3):251–259

Thickening

B. S. Daya Sagar¹ and D. Arun Kumar²

¹Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, Karnataka, India

²Department of Electronics and Communication Engineering and Center for Research and Innovation, KSRM College of Engineering, Kadapa, Andhra Pradesh, India

Definition

Thickening is a mathematical morphology (MM) based operation which is used to *expand/grow* selected regions of image. The MM-based operations such as dilation and closing are used to grow the selected regions in an image. The thickened images (TIs) are the images with enhanced regions and detailed information.

Introduction

Remote sensing images provide the spatial information related to the objects in an area of interest in the form of digital number (DNs) (Jensen 2004). The images consist of finite elements called pixels/pels or patterns. A pixel is associated with features. Each pixel is characterized by features (Richards and Jia 2006).

There exist various approaches to pre-process and process the digital images acquired from sensors (Campbell 1987). The basic approach is the application of mathematical/statistical models on the input images to implement the tasks such as image segmentation, image classification, and change detection (Lillesand and Kiefer 1994).

The remote sensing image analysis is implemented in important domains like spatial, frequency, and wavelets. In spatial domain, the kernel/window/filter is applied on the image to extract the shape/size information of the objects in area of interest. The important operations such as convolution, filtering, and structuring elements (SEs) are applied on input images in spatial domain.

The study related to the application of SE to extract the information from input image is discussed in mathematical morphology (MM). MM deals with the extraction of shape/size features of the objects in an image. MM is mathematically implemented using set theory (Matheron 1975).

There are various types of operations to extract the information content from images using MM. The operations are erosion, dilation, opening, closing, and hit-miss (Serra 1982). The operators such as erosion-dilation, opening-closing, thickening-thinning are dual operators. The detailed study of these operators is given in basic literature related to MM (Serra 1982). In the present study, a detailed discussion of thickening operation is given in the next section.

Thickening

The TIs are obtained by applying morphological operators with structuring element (SE). The SE is a morphological filter/kernel/window used to extract/enhance the information content in the input image (Soille 1999). In thickening, the SE is applied on the input image to obtain the enhanced images. The thickening operation on the input image using the SE (B) is given as:

$$\text{Thicken}(X, B) = (X) \cup (HM(X, B)), \quad (1)$$

where, X is an input image, B is structuring element (part of X). $HM(X, B)$ is hit and miss transform of B :

$$HM(X, B) = (X \ominus B_1) \cap (X^C \ominus B_2), \quad (2)$$



Thickening, Fig. 1 Input binary image with black background and white objects in foreground (Fisher et al. 1996)



Thickening, Fig. 2 Thickened image generated by applying thickening operator on input image (given in Fig. 1) (Fisher et al. 1996)

where B_1 and B_2 are SEs related to elements of B , associated with object and background, $\ominus$ is erosion operator, X^c denote the complement of image X , and X^c is generated by replacing 1 with 0 and vice versa at each pixel of X .

Thickening is the dual of thinning. Thinning is equivalent to thickening the foreground. Thickening operation is implemented in important applications like skeleton by zone of influence (SKIZ). SKIZ is the process of dividing the objects in the image into sub regions. SKIZ is similar to Voronoi diagrams that are used to get the zone of influences in an area of interest (Sagar 2013).

Example

In the present study, the morphological thickening is applied on input image shown in Fig. 1. The input image is binary image with black background and white foreground. The black background is represented with binary 0s, and the white foreground is represented with binary 1s. A morphological thickening operation is performed on input image to generate a thickened image as given in Fig. 2. In Fig. 2, the input image is thickened by 45° filter, and so the objects in Fig. 1 are closed as given Fig. 2.

Thickening increases the area of objects in the image. In real-time applications, thickening is applied to create the buffer zones to the themes like water, traffic, land water extent.

Application of Thickening in Remote Sensing Image Analysis

In the present study, the thickening of objects like vegetation, built-up land, water bodies, etc, is obtained by operating structuring element of size 5×5 on the input image acquired



Thickening, Fig. 3 (a) Input image – sentinel multispectral imager (MSI) of Mumbai City, India, and (b) thickened image obtained by operating a 5×5 structuring element on input image

from sentinel-multispectral imager (MSI). The input image and the thickened images are given in Fig. 3a and Fig. 3b respectively. The input image of Mumbai City, India, with three major classes, namely, water, vegetation, and built-up land, as shown in Fig. 3a. The classes such as water, vegetation, and built-up land are thickened to generate the buffer zones as shown in Fig. 3b. Furthermore, the classes in input image and thickened image are mapped into vector layers to generate thematic layers which can be used for various applications.

Conclusions

In the present study, the MM-based operation named as thickening is discussed with an example. Thickening is implemented to enhance the information content in input image by thinning the information content of background

and increasing the information content of object in foreground. Thickening provides the shape/size information of objects in an image. The real-time applications of thickening include creating river buffer, road buffer zones, etc.

Cross-References

- [Morphological Dilation](#)
- [Morphological Erosion](#)
- [Structuring Element](#)

References

- Campbell JB (1987) Introduction to remote sensing. The Guilford Press
- Fisher R, Perkins S, Walker A, Wolfart E (1996) Hypermedia image processing reference. Wiley, England, pp 118–130
- Jensen JR (2004) Introductory digital image processing: a remote sensing perspective, 3rd edn. Prentice Hall, Upper Saddle River
- Lillesand TM, Kiefer RW (1994) Remote sensing and image interpretation. Wiley, New York
- Matheron G (1975) Random sets and integral geometry. Wiley, New York
- Richards JA, Jia X (2006) Remote sensing digital image analysis: an introduction, 4th edn. Springer Verlag, Germany
- Sagar BSD (2013) Mathematical morphology in geomorphology and GISci. CRC Press (Taylor & Francis Group), Boca Raton
- Serra J (1982) Image analysis and mathematical morphology. Academic Press, London
- Soille P (1999) Morphological image analysis: principles and applications. Springer-Verlag, Heidelberg

Thiergärtner, Hannes

Heinz Burger

Geoscience, Free University Berlin, Berlin, Germany



Fig. 1 Hannes Thiergärtner, courtesy of Prof. Thiergärtner

Biography

Hannes Thiergärtner belongs to the few geoscientists who introduced the development and application of mathematical procedures within the German area. His domain is the testing, adoption and use of deterministic, stochastic and heuristic models in geological research required for the discovery of mineral deposits: the important subject of statistically based mineral resource estimation. Above all, the development of new geoscientific fields of application always were and still are the central issue of his work.

Hannes was born in Dessau (Germany) in 1938 and educated in petrography, geochemistry, mineral deposit geology, mathematical statistics and programming languages at the Mining Academy of Freiberg (former GDR) where he obtained his PhD for the statistical processing of geochemical data (1967) and postdoctoral habilitation for contributions to a system of computer-aided geological data processing (1971).

For two decades, Hannes tested numerous mathematical approaches applied to geoscientific subjects and processes as senior staff member in the Central Geological Survey Berlin. Proven algorithms and methods, for example, for multivariate profile correlation in deep drilling projects, lithological classification of East German geological units, and spatial modeling of geochemical data. Outline of quasi-homogenous areas for predictive results were applied by him to explore Atlantic deep-sea mud, Baltic Sea sand, and gravel, in order to support geochemical surface prospection or reserve estimation of mines. Later he extended the scope of his work to ecological problems. His methodological progress was communicated in dozens of lectures within the German Geological Society which acknowledged these pioneer activities by awarding him the Abraham-Gottlob-Werner-Medal. Hannes was also concerned with the development of a multilingual Geoinformation System and transnational exchange of geoscientific information in collaboration with the Council for Mutual Economic Assistance, where he represented the former GDR in its subsystem GEOINFORM. He was appointed a member of the regional group of European countries in the Commission COGEODATA of the International Union for Geological Sciences from 1987 to 1990 but because of political restrictions he could communicate only with scientists from Eastern Europe. (Before 1990 very few meetings with Western colleagues were possible – under more or less conspiratorial circumstances – during conferences held in Eastern Europe).

Following additional ecologic training, he became a leading evaluator of contaminated subsoil and groundwater in the Governmental Agency for the re-privatization of industrial

companies from 1990 to 2000. Later he worked in this field as a freelancer.

Hannes was university lecturer for mathematical geosciences at the Mining Academy Freiberg, the Greifswald University, and the Free University Berlin which appointed him as extraordinary professor in 1999. He published books on the statistical processing of geochemical data (Thiergärtner 1968) and the geology of his home town and surrounding areas (Thiergärtner 2011) as well as more than 100 contributions mainly covering the application of mathematical methods to geoscientific and ecological issues. Additionally, he acted as a guest editor for several issues of geoscientific journals, one with eight contributions by IAMG founding members (Zt. Deutsche Geologische Gesellschaft, Stuttgart 155(2–4):195–285, 2005). In 1996, he established the journal *Mathematische Geologie* published at CPress Publ. (Dresden) with international contributions. He edited this journal until 2003.

Hannes Thiergärtner belongs to the founding members of the International Association for Mathematical Geosciences at the International Geological Congress 1968 in Prague. He could reactivate his IAMG membership after the German Reunification in 1990 and subsequently contributed to numerous IAMG annual conferences.

Hannes is acknowledged as a humble and creative worker not made for public showmanship but for staying primarily a geoscientist. He is married and has two adult children. In 2013 he finished his sporting career as a long-distance rower after completing exactly the mathematically exact number of 55,555 training kilometers for which he obtained the Equator Award of the German Rowing Union.

Bibliography

- Thiergärtner H (1968) Basic problems of statistical analysis of geochemical data. Grundstoffverlag, Leipzig. (in German)
- Thiergärtner H (ed) (1977) Mathematical methods and information search systems in geology. Central Geological Institute, Berlin. (in Russian)
- Thiergärtner H (ed) (1998) Mathematical groundwater modelling – unconventional solutions and limitations, vol 2. CPress Publishers, pp 1–131
- Thiergärtner H (ed) (2006) Practical application of mathematical models in geology from the 32nd Intern. Geological Congress, Florence, 2004. Math. Geology, Springer Sci., New York 38: 659–780
- Thiergärtner H (2011) City of Dessau before human settlement – a contribution to the geology of the former Duchy Anhalt. Funk Verlag, Dessau. (in German)
- Thiergärtner H, Burger H (eds) (2012) New ideas and solutions in mathematical geology. Zt. Geol. Wiss., Berlin 40: 197–415

Three-Dimensional Geologic Modeling

Qiyu Chen², Gang Liu^{1,2}, Xiaogang Ma³ and Junqiang Zhang²

¹State Key Laboratory of Biogeology and Environmental Geology, Wuhan, Hubei, China

²School of Computer Science, China University of Geosciences, Wuhan, Hubei, China

³Department of Computer Science, University of Idaho, Moscow, ID, USA

Definition

Three-dimensional geologic modeling (three-dimensional geomodeling) is the three-dimensional digital characterization, reconstruction, and reproduction of geological objects, phenomena, and processes through the comprehensive use of modern cutting-edge theory and technology, such as computer science, spatial information theory, scientific visualization technology, etc. From a broad sense, three-dimensional geomodeling is an integrated workflow of geological data and information processing including data organization and management, spatial information processing, geological interpretation, three-dimensional visualization, geological spatial analysis, statistical analysis, prediction, and so on. The purpose of three-dimensional geomodeling is to provide a visualization platform integrating scientific research, aided design, and decision support for geoscientists to deeply understand and utilize the essence and laws of geological bodies, phenomena, and processes. Building three-dimensional geological models from field and subsurface data has been a typical task in geological studies involving basic geological survey, natural resource evaluation, and hazard assessment.

Cutting-Edge Technologies of Three-dimensional Geologic Modeling

Three-dimensional geomodeling is about modeling the topology, the geometry, and the properties of a particular region (Mallet 2002). It has been rapidly developed in the past two decades through the promotion of powerful application requirements and the development of computer hardware/software, spatial information theory, three-dimensional visualization technology, and other related disciplines. The three-dimensional visualization of geological structures and heterogeneous properties can provide a more realistic and intuitive characterization of subsurface geological phenomena and

structures. Three-dimensional geological models are able to show the subsurface geological structures and physicochemical properties in a virtual three-dimensional space. Three-dimensional geomodeling has been a basic task in geological studies, geological survey, and engineering applications as well as the development of geological information system.

Three-dimensional geomodeling technology began in the 1980s (Houlding 1994), and has now formed a relatively mature modeling method system. It combines spatial information management, geological interpretation, spatial analysis, geostatistics, prediction, and visualization in a three-dimensional environment. The main contents of three-dimensional geomodeling include the study of three-dimensional spatial data model and the development of three-dimensional geomodeling methods. There are many methods for constructing three-dimensional geological models, which can be summarized as the methods for different data sets, data- and/or knowledge-driven methods, explicit and implicit methods, and geostatistics-based stochastic simulation methods. For different modeling objects, different data sources, or different uses of the models, the spatial data model and the geomodeling method are also different.

Object: Geological Structures and Properties

The objects of three-dimensional geomodeling include all subsurface structures and properties, such as geological structure, strata, rock type, various physicochemical properties, etc. For different objects, various three-dimensional geomodeling methods have been proposed, such as methods of modeling sedimentary stratigraphic framework, three-dimensional geomodeling methods of complex geological structures such as faults, folds, lenses and so on, and methods of modeling various physical and chemical properties, etc.

Data Sources

Three-dimensional geomodeling generally starts with spatial points, lines, or surfaces interpreted from geological observations, surveys, explorations, measurements, and analyses (Wellmann and Caumon 2018). The data used in geomodeling is very heterogeneous and includes all direct and indirect multiple-source data obtained during whole geological surveys and explorations. More precisely, three-

dimensional geomodeling applications typically use several data types:(1) different sources: data from different exploration departments, with different geological periods, and with different software formats; (2) different dimensions: one-dimensional observation points, two-dimensional geological maps/sections, three-dimensional DEM, etc.; (3) different types: remote sensing data, field observation data, drilling data, logging experiment data, artificial seismic data, gravity/magnetic data, physicochemical properties, etc.; (4) different precisions and scales, such as remote sensing data with different precisions, geological maps at different scales, etc. One of the goals of three-dimensional geomodeling is to continuously research and develop new technologies and methods, and use the above-mentioned data to perform three-dimensional visual characterization of subsurface geological structures and properties. So far, a series of three-dimensional geomodeling methods for different type data have been proposed, such as the automatic construction method of three-dimensional stratigraphic framework by using borehole data, the three-dimensional reconstruction method of regional geological structures through geological planes, and the interactive three-dimensional geomodeling method based on topological reasoning of geological sections, three-dimensional automatic reconstruction method by fusing multiple-source data (e.g., drilling data, remote sensing data, seismic data, and so on), etc. Therefore, how to effectively gather and use the above-mentioned multiple-source heterogeneous data to construct more reliable three-dimensional geological models is one of the most important challenges in this field.

Spatial Data Model

Three-dimensional geological modeling technology is accompanied by the development of spatial data model. As shown in Table 1, more than 20 data models have been developed (Wu 2004). They can be divided into surface-based models and voxels-based models. Generally, when modeling the spatial geometry of geological structures or geological objects, surface-based models will be used. However, voxel-based models are more advantageous for performing quantitative evaluation, analysis and prediction of geological resources, and dynamics simulation of

Three-Dimensional Geologic Modeling, Table 1 Categories of three-dimensional spatial data models (Wu 2004)

Surface-based model	Voxel-based model		Mixed model
	Regular volume	Irregular volume	
Irregular triangular network (TIN)	CSG-tree	Tetrahedral network (TEN)	TIN-CSG
Grid	Voxel	Pyramid	TIN-Octree
Boundary representation (B-Rep)	Octree	Tri-prism (TP)	Wire Framework-Block
Nonuniform rational B-splines (NURBS)	Needle	Geocellular	Octree-TEN
Wire-frame	Regular block	Irregular block	B-Rep-Voxel
Serial sections		Solid	
Sections-TIN mixed		Three-dimensional voronoi volume	
Multiple-DEMs		Generalized tri-prism (GTP)	

modeling. The formation and evolution of geological structures processes are extremely complex. It is difficult to approximate and characterize the reality only by limited data. The Both paths of the two viewpoints are linked in multiple ways, but share several commonalities. Geological knowledge is used in almost all steps such as data processing, conceptual and mathematical model construction, parameterization, and inversion. As a core step, fundamental principles and geological knowledge combine geological and geophysical information as well as conceptual and mathematical model in a principled way.

The data-driven approach may lead to the conflict between subsurface realizations and geological concepts or observations. Conversely, a common criticism for the geomodel-based approach is that (a) it starts with subjective information, which may induce bias in the model forecasts, (b) it contains interpretative elements that cannot be verified, and (c) it introduces a level of complexity that is not supported by data. Therefore, an integrated geomodeling viewpoint that fuses various data, models, and geological knowledge will be expected.

Explicit and Implicit Methods

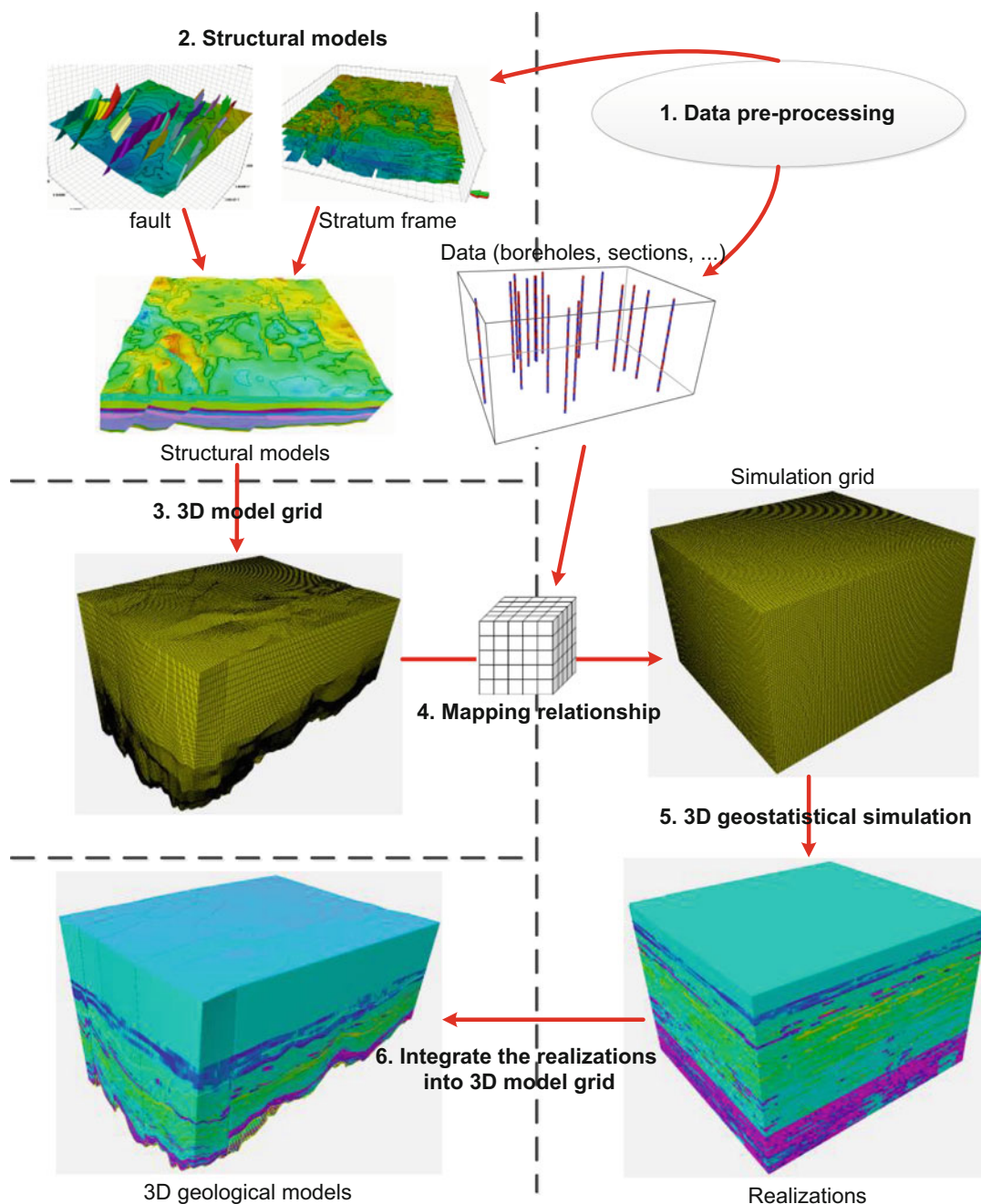
Three-dimensional explicit modeling methods refer to determining the interfaces between different geological objects with three-dimensional surface meshes by using a human-machine interactive or semi-automatic way in a three-dimensional domain. Surface meshes consist of mainly triangular irregular networks (TINs) or parametric surfaces such as NURBS.

The surface mesh consists of a set of three-dimensional nodes, so the main task is to effectively organize these three-dimensional points through human-computer interaction or interpolation on the basis of sparse data extracted from such as boreholes, cross-sections, maps, etc. In order to obtain enough three-dimensional points to express geological interfaces, various implementation variants such as splines, inverse distance weight, kriging, and NURBS have been developed and utilized in this process. As all of these methods connect points by linear segments, they tend to create very angular geological surfaces. Smoothing methods such as discrete smooth interpolation (DSI) method (Mallet 1989) allow to obtain more realistic shapes while honoring subsurface data. Although it is easy to integrate expert experience into three-dimensional geological models, it also brings uncertainty that cannot be assessed. Model construction can be automated in relatively simple structural cases, but their application in complex domains generally involves significant manual editing. For complex geological structures, the limitations of the low efficiency have gradually become prominent. Once the modeling data changes or new data is obtained, the modeler needs to reinterpret the data and rebuild the models, which causes great difficulty in model updating.

Implicit methods are more recent and they have opened new ways to automatically construct geological models with a high degree of complexity. The idea of implicit methods is to compute the scalar field $s(x)$ so that all data points sampling the surface have the same scalar value (Mallet 2002). Two main numerical methods have been proposed to create individual implicit surfaces from sparse data points: meshless methods and mesh-based methods. The general form of meshless methods includes radial basis function (RBF) interpolation and dual kriging. These meshless methods are very convenient but they involve solving a large and dense linear system. Mesh-based methods have also been proposed to compute the scalar field (Caumon et al. 2009). In this case, the domain is covered by a linear tetrahedral mesh conformable to discontinuities. Implicit methods directly consider volumes and are overall more robust and automatic than surface-based models. Complex geological structures can be obtained automatically without need for expert input and user interaction. Another capability of implicit methods is to honor several types of structural data without any projection. In terms of model updating, implicit methods also showed a significant advantage due to the automation of interpolation used in them (Wellmann and Caumon 2018).

Geostatistics-Based three-dimensional Simulation and Modeling

Geostatistical simulation is one of the most important tools for building an ensemble of probable realizations in earth science (Journel and Huijbregts 1978). Three-dimensional characterization and modeling of geological structures have been investigated for several years via geostatistical simulation. Several simulation methods have been proposed that can produce various realizations. Methods such as sequential Gaussian simulation (SGSIM) and sequential indicator simulation (SISIM) have become popular among different fields of earth sciences. These methods give a number of realizations or interpolation scenarios, which allow assessing the uncertainty and quantifying it more accurately. Due to relying on variogram, kriging-based geostatistical simulations are not able to reproduce complex patterns. Clearly, considering only two points is not sufficient for reproducing complex and heterogeneous models. Thus, several attempts in the recent years in the context of multiple-point geostatistics (MPS) have been made (Mariethoz and Caers 2014) that can capture high-order (more than two points) statistics simultaneously from a reference model called training image (TI). MPS-based simulation has recently gone under a remarkable progress in handling complex and more realistic phenomenon that can produce large amount of the expected uncertainty and variability. Several three-dimensional applications have demonstrated its applicability in three-dimensional characterization and modeling of complex geological structures (Mariethoz and Caers 2014).



Three-Dimensional Geologic Modeling, Fig. 2 Three-dimensional geological structure-property integrated modeling framework based on geostatistical simulation (Chen et al. 2020)

Because geostatistics-based simulations are usually performed on a regular Cartesian grid, the realizations cannot realistically reflect the exact morphology features of geological structures. In addition, geostatistical simulation should be done under the constraints of fundamental principles and geological knowledge. Therefore, it is an effective way to develop a three-dimensional structure-property integrated modeling method based on geostatistical simulation under

the constraints of surface-based models. Figure 2 shows a three-dimensional stochastic modeling framework for complex subsurface structures and properties by using conditional simulation technique (Chen et al. 2020). Under the framework, the realizations and various properties can be integrated into a unified model grid resulting in integrated management and analysis of these multiple three-dimensional models in a unified three-dimensional space.

Conclusions and Outlook

The fundamental goal for the theory of geological modeling is to obtain methods that allow a reasonable description of geometric features and physicochemical properties in full three-dimensional space. Geological modeling methods have been rapidly developing over the last years. Various heterogeneous data have been used to construct three-dimensional geological models. A series of three-dimensional geomodeling methods for different type data have been proposed. Arguably, two different viewpoints, data-driven and geomodel-driven, are both complementary of each other to promote the development of three-dimensional geomodeling methods, and act on every step of the three-dimensional geomodeling process. They are curly divided into explicit and implicit methods. They have their own characteristics and have been applied in various application fields of three-dimensional geomodeling. However, implicit methods are more recent and they allow to automatically construct complex geological models. Geostatistical simulation is able to automatically reproduce a number of realizations, and their ability for characterizing and modeling geological structures in three-dimensional domain have also been investigated. In the end, several interesting perspectives and recommendations for future research in the field are outlined as follows.

- **Integration:** In addition to integrating various types of heterogeneous data, geological knowledge, and mathematical models into three-dimensional geological models, the integrated characterization of geological structures and properties is still a challenge. Therefore, new unified spatial data models and integrated modeling methods are expected. In addition, the integrated expression, scheduling, and management of three-dimensional models with different scales is also a problem. In the big data era, three-dimensional geological models should be given more functions and capabilities. Three-dimensional geological models are not only used to the present subsurface geological structures and properties, but also are the platforms and portals of the integration, management, mining, visual analysis, and sharing of geological big data.
- **Visualization:** In order to further enhance the visual effect and reality of three-dimensional geological models, new visualization technologies such as virtual reality (VR) and augmented reality (AR) should be used in three-dimensional geomodeling. Moreover, advanced scientific visualization and visual analysis technologies can help data mining and knowledge discovery from three-dimensional geological models.
- **Efficiency:** In practice, the computational efficiency of modeling and simulation methods remains limited with regard to the ambition of creating reasonably accurate models explaining all observations. The development of

efficient exploiting modern parallel computer architectures is an essential area for future advances.

- **Machine learning:** The advent of efficient machine learning methods opens very interesting perspectives in the field of joint geological modeling and geostatistical simulation. Many effective approaches have been used to improve the automation and intelligence of three-dimensional geomodeling, and their effectiveness has been investigated. It shows great potential in the field of three-dimensional geomodeling in the future.

Cross-References

- [Geostatistics](#)
- [Interpolation](#)
- [Multiple Point Statistics](#)
- [Simulation](#)
- [Virtual Globe](#)

Bibliography

- Caumon G, Collon-Drouaillet P, De Veslud CLC, Viseur S, Sausse J (2009) Surface-based 3-D modeling of geological structures. *Math Geosci* 41(8):927–945
- Chen Q, Liu G, Ma X, Li X, He Z (2020) 3D stochastic modeling framework for quaternary sediments using multiple-point statistics: a case study in Minjiang estuary area, Southeast China. *Comput Geosci* 136:104404
- Houlding SW (1994) 3D geoscience modeling computer techniques for geological characterization. Springer
- Journel AG, Huijbregts CJ (1978) Mining geostatistics. Academic
- Mallet J-L (1989) Discrete smooth interpolation. *ACM Trans Graph* 8(2):121–144
- Mallet J-L (2002) *Geomodeling*. Oxford University Press
- Mariethoz G, Caers J (2014) Multiple-point geostatistics: stochastic modeling with training images. Wiley, Washington, DC
- Wellmann F, Caumon G (2018) 3-D structural geological models: concepts, methods, and uncertainties. In: *Advances in geophysics*, vol 59. Elsevier, Waltham, pp 1–121
- Wu L (2004) Topological relations embodied in a generalized tri-prism (GTP) model for a 3D geoscience modeling system. *Comput Geosci* 30(4):405–418

Time Series Analysis

Abraão D. C. Nascimento
Universidade Federal de Pernambuco, Recife, Brazil

Synonyms

Dynamic modeling; Signal processing

Definition

Some of the interesting geoscience applications to the time series (TS) users can be understood by data extracted from daily mean values of solar radiation obtained at the Earth's surface (Gilgen 2006), earthquake and explosion occurrences (Shumway and Stoffer 2011), intensity returns from synthetic aperture radar imagery (Almeida-Junior and Nascimento 2021), among others. The main goals of the time series analysis (TSA) are to propose mathematical models to describe the dependence structure of real phenomena and to predict the associated future values.

A time series can be understood as a collection of random variables with time-ordered indices a subset $\mathcal{T}$ of $\mathbb{Z}$ or $\mathbb{R}$. Mathematically, the sequence $\{X_t; t \in \mathcal{T}\}$ is a stochastic process (SP). The observed values of a SP are called the outcomes of the respective SP. In this chapter, it will adopt $t \in \mathcal{T} \subset \mathbb{Z}$. From now on, the term time series will refer to both the SP and one of its particular realizations.

The remainder of this chapter is outlined as follows. First, the TSA main concepts are presented. Second, it is approached a smoothing process by the state-space model (SSM). Third, the seasonal autoregressive integrated moving average (SARIMA) model is discussed. Finally, to illustrate the presented concepts, an application is made to time series of vegetation indices obtained with the Moderate Resolution Imaging Spectroradiometer (MODIS). All computational operations are made using the software R, cf. Shumway and Stoffer (2011), Appendix R. In particular the package `dtwSat` (Maus et al. 2019) is used for extracting the data and `forecast`, `seastests`, and `fpp` for doing the statistical operations with time series.

This chapter is intended to provide the reader with knowledge of the TSA. Basic courses in statistics and calculus theories are assumed as prerequisites.

Initial Concepts

The Time Domain Approach and Stationarity

Let $\{X_t; t \in \mathbb{Z}\}$ be a time series. X_t is said to be stationary (also known as weak stationarity, covariance stationarity, stationarity in the wide sense or second-order stationarity), if

- (i) $E|X_t|^2 < \infty$, for all $t \in \mathbb{Z}$;
- (ii) $EX_t = m$, for all $t \in \mathbb{Z}$;
- (iii) $\gamma_X(r, s) = \gamma_X(t + r, t + s)$, for all $r, s, t \in \mathbb{Z}$,

where $E(\cdot)$ is the expected value, $\text{Cov}(\cdot, \cdot)$ is the covariance operator, and $\gamma_X(r, s) \triangleq \text{Cov}(X_r, X_s)$. If $\{X_t; t \in \mathbb{Z}\}$ is a stationary time series, $\gamma_X(r, s) = \gamma_X(r - s, 0)$ for all $r, s \in \mathbb{Z}$. Thus, the autocovariance function (acvf) of a stationary process is defined as the function of one variable:

$$\gamma_X(h) \triangleq \gamma_X(h, 0) = \text{Cov}(X_{t+h}, X_t), \text{ for all } t, h \in \mathbb{Z}.$$

The term “ $\gamma_X(h)$ ” indicates an acvf of X_t at the lag h . The acvf satisfies the properties:

$$(i) \gamma_X(0) \geq 0; (ii) |\gamma_X(h)| \leq \gamma_X(0) \text{ for all } h \in \mathbb{Z},$$

$$\text{and } (iii) \gamma_X(h) = \gamma_X(-h),$$

i.e., it is limited and even. Other important function defined from $\gamma_X(h)$ is the autocorrelation function (acf) of $\{X_t\}$, say $\rho_X(h)$, defined as:

$$\rho_X(h) \triangleq \text{Cor}(X_{t+h}, X_t) = \frac{\text{Cov}(X_{t+h}, X_t)}{\sqrt{\text{Var}(X_{t+h})\text{Var}(X_t)}} = \frac{\gamma_X(h)}{\gamma_X(0)},$$

for all $t, h \in \mathbb{Z}$, where $\gamma_X(0) = \text{Var}(X_t)$ and $\text{Cor}(X_{t+h}, X_t)$ is the correlation operator between the variables X_t and X_{t+h} . For $h \in \mathbb{Z}$, the acf $\rho_X(h)$ satisfies the properties:

$$(i) \rho_X(0) = 1, (ii) |\rho_X(h)| \leq 1, \text{ and } (iii) \rho_X(h) = \rho_X(-h).$$

Other important measure is the partial acf (pacf) that can be defined in terms of $\rho_X(h)$ as

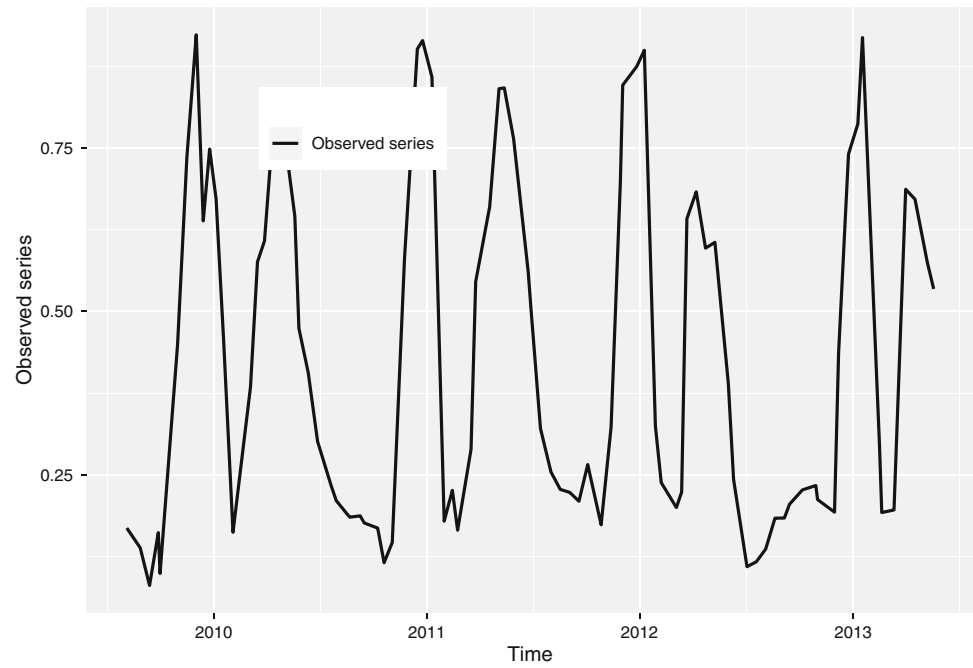
$$\pi_X(h) \triangleq \frac{\rho_X(h) - \alpha_1 \rho_X(h-1) - \dots - \alpha_{h-1} \rho_X(1)}{1 - \alpha_1 \rho_X(1) - \dots - \alpha_{h-1} \rho_X(h-1)},$$

where α_i , for $i = 1, \dots, h-1$, are obtained by determinant relations involving $\rho_X(h)$ (Wei 2006, p. 13).

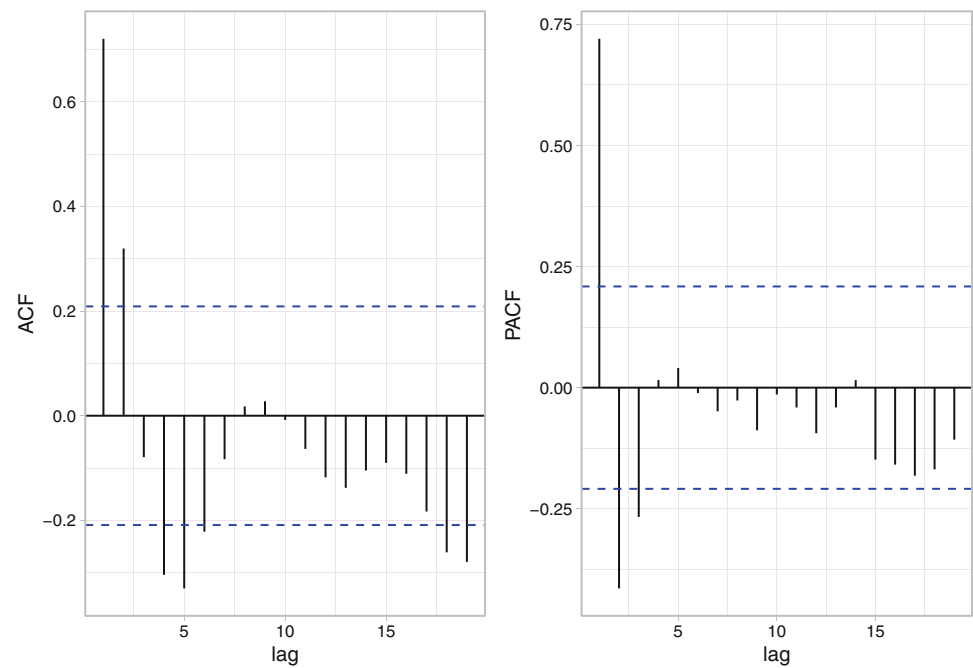
Introducing an Example in Geoscience

To illustrate some of the concepts discussed, consider the introduction of the application context used in this chapter. Remote sensing using imagery data is one of the most commonly used methods for measuring changes in land use and land cover. The time series to be modeled refers to the Enhanced Vegetation Index (EVI) obtained from a series of MODIS MOD13Q1 (Didan 2015) images acquired from 2009-08-05 to 2013-07-31 with ninety-three observations. Figure 1a shows the EVI series fluctuating around the mean value 0.4221. Figure 1b shows the corresponding samples acf and pacf, which indicate an expressive dependence structure possibly indicative of a third-order autoregressive pattern, suggesting the cut after the third lag in the sample pacf. Some descriptive measures of the observed time series are given in Table 1 (for abbreviations Coefficient of variation (CV), Standard Deviate (SD), Minimum (Min.), and Maximum (Max.)).

It follows that EVI values vary in the range (0.0809, 0.9227), the differences between pairs (Min., 1° Quartile),

Time Series Analysis,**Fig. 1** EVI observed series and its sample acf and pacf

(a) EVI observed series



(b) The sample acf and pacf

Table 1

Min.	1st Quartile	Median	Mean	SD	CV	3rd Quartile	Max.
0.0809	0.1924	0.3227	0.4221	0.2619	62.0498%	0.6455	0.9227

(Median, Average), and (3° Quartile, Max.) are expressive, and the CV indicates a meaningful dispersion. In what follows, two modeling strategies are presented.

Smoothing Methods

Hyndman et al. (2008) have presented an important class that includes various smoothing models. Let $\{X_t\}$ be a time series and $\{y_t\}$ a vector of unobserved components (called states) to describe trend (level and inclination) and seasonal effects in X_t . Then the SSM is given by

$$\begin{cases} X_t = \mathbf{w}^\top \mathbf{y}_{t-1} + Z_t \in \mathbb{R} \text{ [Relation between } (X_t) \text{ and states } (\mathbf{y}_t)], \\ \mathbf{y}_t = \mathbf{F} \mathbf{y}_{t-1} + \mathbf{g} Z_t \in \mathbb{R}^p \text{ [transition equations]}, \end{cases} \quad (1)$$

where p represents the number of transition equations, $\{Z_t\}$ represents the Gaussian white noise (GWN) having $\mathbf{E}(Z_t) = 0$ and $\text{Var}(Z_t) = \sigma^2$, and $\mathbf{F}$, $\mathbf{g}$, and $\mathbf{w}$ are the associated coefficients. The nonlinear version for Eq. (1) is given by

$$\begin{cases} X_t = w(\mathbf{y}_{t-1}) + r(\mathbf{y}_{t-1})Z_t \in \mathbb{R}, \\ \mathbf{y}_t = f(\mathbf{y}_{t-1}) + g(\mathbf{y}_{t-1})Z_t \in \mathbb{R}^p, \end{cases} \quad (2)$$

where $w(\cdot)$ and $r(\cdot)$ are real-valued function and $f(\cdot)$ and $g(\cdot)$ are vector functions.

Hyndman et al. (2008) coined the notation (E, T, S) = (Error, Trend, Seasonality) for the SSMs. Some well-known smoothing models are particular cases in the ETS class. For instance, assuming the structures Additive “A,” Multiplicative “M,” and None “N,” (A, N, N) represents the simple exponential smoothing, (A, A, N) indicates the Holts linear model, (A, A, A) provides the Holt-Winters’ additive method, and (A, A, M) yields the Holt-Winters multiplicative model.

In practice, the initial states, say $\mathbf{y}_0$, and vector of parameters, $\boldsymbol{\theta}$, involved in models of Eqs. (1) and (2) are unknown and, therefore, should be estimated. According to Hyndman et al. (2008), the maximum likelihood estimators for $\boldsymbol{\theta}$ and $\mathbf{y}_0$ are obtained by minimizing the expression: Given a time series $X_1, \dots, X_n$ following Eq. (2),

$$S(\boldsymbol{\theta}, \mathbf{y}_0) = \left| \prod_{t=1}^n r(\mathbf{y}_{t-1}) \right|^{2/n} \sum_{t=1}^n Z_t^2$$

which is known as the augmented sum of squared errors criterion.

First Application

Now, consider the ETS model to describe the EVI series, discussed previously. We aim to model the first 88 observations and after to predict the five latter. By applying function `ets(.)` of the package `fpp` of the software R, the ETS (A, N, N) model was chosen:

$$\begin{cases} X_t = \ell_{t-1} + Z_t, \\ \ell_t = \ell_{t-1} + \alpha Z_t, \Leftrightarrow \ell_t = \alpha y_t + (1 - \alpha)\ell_{t-1} \\ \hat{X}_{t+h|t} = \ell_{t-1} \Leftrightarrow \hat{X}_{t+h|t} = \ell_t, \end{cases} \quad (3)$$

having the estimated smooth parameter $\alpha = 0.9999$, initial stage $\ell_0 = 0.1684$, and estimated standard deviate of Z_t $\sigma = 0.1684$, where $\hat{X}_{t+h|t}$ is the forecast equation in a horizon h . The fitted model suggested $\hat{X}_{88+h|t} = 0.5342041$ to predict the elements in (Table 2):

This fitted model presents the best forecast for $h = 1$. Figure 2 presents the ETS fit to EVI observed series. By visual inspection, it is noticeable that the acceptable fit. Applying the Ljung-Box test (in language R, by using `Box.test(., type = “Ljung-Box”)`), there is evidence at the nominal level of 1% that the associated residuals follow the GWN pattern.

The SARIMA Process

Definition and Some Particular Cases

Let X_t be a time series, $B\{X_t\} = X_{t-1}$ the backshift operator, and $B^s\{X_t\} = X_{t-s}$ its extension at the presence of a seasonal pattern with frequency s . Thus, the following polynomials can be defined: For $p, P, Q, Q_s, D, D_s \in \mathbb{Z}_+$,

$$\begin{aligned} \phi_p(B) &= 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p, \\ \theta_q(B) &= 1 + \theta_1 B + \theta_2 B^2 + \dots + \theta_q B^q, \\ \Phi_P(B^s) &= 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \dots - \Phi_P B^{Ps}, \\ \Theta_Q(B^s) &= 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \dots + \Theta_Q B^{Qs}, \\ \nabla^d &= (1 - B)^d, \end{aligned}$$

and

$$\nabla_s^{D_s} = (1 - B^s)^{D_s}.$$

Table 2

2013-05-26	2013-06-13	2013-06-29	2013-07-15	2013-07-31
0.5434	0.4156	0.1220	0.1163	0.1309

Time Series Analysis,

Fig. 2 EVI observed and fitted series by the ETS model (observed series – solid black, predicted series – solid red, forecast lower and upper limits – dashed red, forecast – dashed blue)

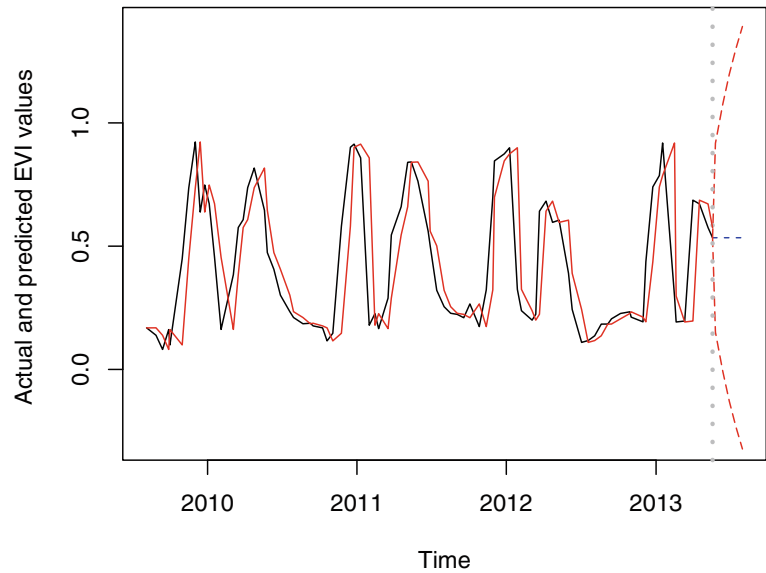
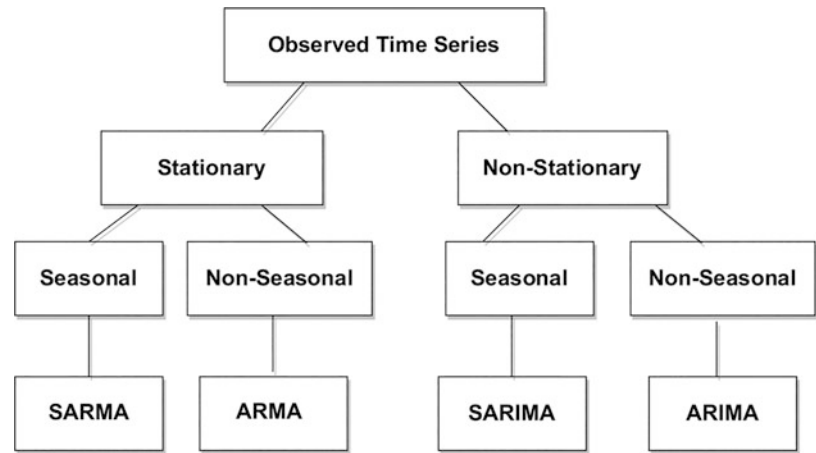
**Time Series Analysis,**

Fig. 3 Class of SARIMA submodels



The SARIMA model is given as follows (Brockwell and Davis 1991): Let X_t be a process of mean zero (without loss of generality),

$$\Phi_P(B^s) \phi_p(B) \{\nabla_s^D \nabla^d X_t\} = \Theta_Q(B^s) \theta_q(B) Z_t$$

or, equivalently,

$$\Phi_P(B^s) \phi_p(B) Y_t = \Theta_Q(B^s) \theta_q(B) Z_t, \quad (4)$$

where $Y_t = \nabla_s^D \nabla^d \{X_t\}$ points out an application of operations to deal with the nonstationarity of X_t from nonseasonal (∇^d) and seasonal (∇_s^D) perspectives, d and D mean the non-seasonal and seasonal orders of differencing, $\phi_p(\cdot)$ and $\Phi_P(\cdot)$ represent the nonseasonal and seasonal autoregressive (AR) operators with orders p and P , $\theta_q(\cdot)$ and $\Theta_Q(\cdot)$ indicate the nonseasonal and seasonal moving average (MA) operators with orders q and Q , and Z_t follows a GWN. This model is denoted as $\text{ARIMA}(p, d, \varrho) \times (P, D, Q)_s$.

Figure 3 exhibits the main submodels in the SARIMA class. These submodels are resulting of the intersection between the stationary and seasonality characteristics:

- $\text{ARMA}(p, q)$ (stationarity and nonseasonality): $\phi_p(B) X_t = \theta_q(B) Z_t$ and $Z_t \sim \text{GWN}(0, \sigma^2)$.
- $\text{SARMA}(P, Q)_s$ (stationarity and seasonality): $\Phi_P(B^s) X_t = \Theta_Q(B^s) Z_t$ and $Z_t \sim \text{GWN}(0, \sigma^2)$.
- $\text{ARIMA}(p, d, q)$ (nonstationarity and nonseasonality): $\phi_p(B) \{\nabla^d X_t\} = \theta_q(B) Z_t$ and $Z_t \sim \text{GWN}(0, \sigma^2)$.
- $\text{ARIMA}(p, d, q) \times (P, D, Q)_s$ (nonstationarity and seasonality): $\Phi_P(B) \phi_p(B) \{\nabla_s^D \nabla^d X_t\} = \theta_q(B) \Theta_Q(B) Z_t$ and $Z_t \sim \text{GWN}(0, \sigma^2)$.

In practice, users of the SARIMA class need to identify from observed time series if there are (not) evidence for rejecting stationarity and seasonality, what is made by using hypotheses tests (Shumway and Stoffer 2011).

Estimation

To determine the estimation for the SARIMA parameters, consider the ARMA process defined from Eq. (4), taking $d = D = 0$ and $\Phi_P(B^s) = \Theta_Q(B^s) = 1$: Let $\{X_t; t \in \mathbb{Z}\}$ be a stationary process following the ARMA(p, q) process and $x_1, x_2, \dots, x_n$ be one of its n -points outcome with mean zero, then the associated joint likelihood is expressed as

$$L(\boldsymbol{\beta}, \sigma^2) \triangleq L(\boldsymbol{\beta}, \sigma^2; x_1, x_2, \dots, x_n) \\ = f(x_1, x_2, \dots, x_n; \boldsymbol{\beta}, \sigma^2), \quad (5)$$

where $f(\cdot)$ is the joint density of the process, $\boldsymbol{\beta}^\top = (\phi^\top, \theta^\top)$, $\phi = (\phi_1, \dots, \phi_p)^\top$, and $\theta = (\theta_1, \dots, \theta_q)^\top$. Now, consider a brief discussion on the likelihood-based estimation for $(\boldsymbol{\beta}^\top, \sigma^2)$. According to Hamilton (1994, ch. 5) and Brockwell and Davis (1991, ch. 8), the Gaussian closed-form likelihood can be derived from the factored density

$$f(x_1, \dots, x_n; \boldsymbol{\beta}, \sigma^2) = f(x_1; \boldsymbol{\beta}, \sigma^2) \times \prod_{t=2}^n f(x_t | x_{t-1}, \dots, x_1; \boldsymbol{\beta}, \sigma^2) \\ = (2\pi\sigma^2)^{-n/2} [r_1^0(\boldsymbol{\beta}) \times r_2^1(\boldsymbol{\beta}) \times \dots \times r_n^{n-1}(\boldsymbol{\beta})]^{-1/2} \exp\left[-\frac{S(\boldsymbol{\beta})}{2\sigma^2}\right],$$

where $S(\boldsymbol{\beta}) \triangleq \sum_{t=1}^n \left[\frac{(x_t - x_t^{t-1}(\boldsymbol{\beta}))^2}{r_t^{t-1}(\boldsymbol{\beta})} \right]$, $x_t^{t-1}(\boldsymbol{\beta}) \triangleq \mathbb{E}(X_t | X_{t-1} = x_{t-1}, \dots, X_1 = x_1)$ is the conditional mean of the Gaussian SP (GSP), and $r_t^{t-1}(\boldsymbol{\beta})$ is such that the conditional GSP variance is

$\text{Var}(X_t | X_{t-1} = x_{t-1}, \dots, X_1 = x_1) = \sigma^2 r_t^{t-1}(\boldsymbol{\beta})$. Thus, the maximum likelihood estimates are given by (Brockwell and Davis 1991)

$$\hat{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta} \in \mathbb{R}^{p+q}} \left[\log f\left(x_1, \dots, x_n; \boldsymbol{\beta}, \frac{1}{n} S(\boldsymbol{\beta})\right) \right]$$

Finally, this idea is easily applied for Eq. (4). Given values p, d, q, P, D, Q , note that $Y_t = \nabla^d \nabla^D X_t$ follows an ARMA($p + sP, q + sQ$) process and, thus, the $(p + P + q + Q)$ -dimensional parameter vector can be estimated by the previous process.

Second Application

Now, consider the SARIMA model to describe the first 88 observations of the EVI series and then predict the five latter. By applying function `auto.arima(.)` of the package forecast of the software R, the SARIMA model was chosen: As indicated by the interpretation of the sample pacf in Fig. 1b, an AR(3) model was fitted such that

$$X_t = 0.4312 + 0.9076(X_{t-1} - 0.4312) - 0.1443(X_{t-2} - 0.4312) \\ - 0.2614(X_{t-3} - 0.4312) + Z_t,$$

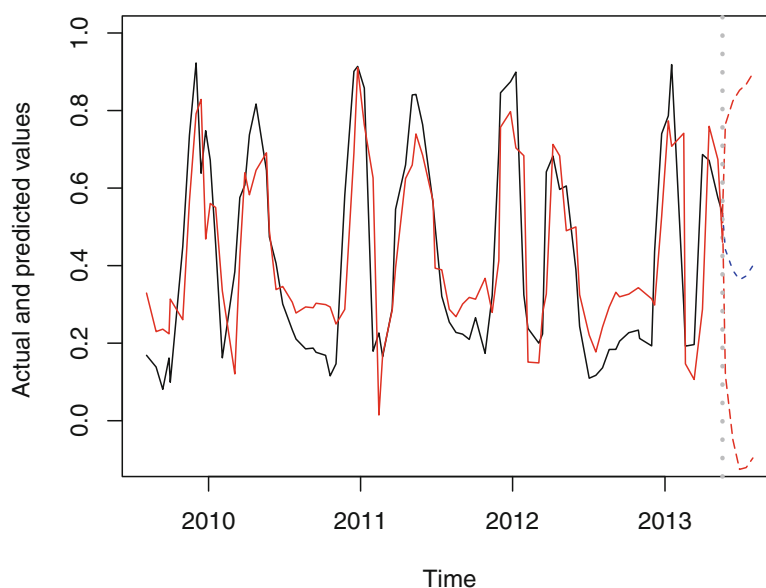
where the estimated variance of Z_t is $\hat{\sigma}^2 = 0.02614$. Using the function `predict(., n.ahead = 5)` of the package

Table 3

Time	2013-05-26	2013-06-13	2013-06-29	2013-07-15	2013-07-31
X_t	0.5434	0.4156	0.1220	0.1163	0.1309
$\hat{X}_{88+h t}$	0.4411681	0.3877664	0.3634413	0.3733872	0.3998829

Time Series Analysis,

Fig. 4 EVI observed and fitted series by the SARIMA model (observed series – solid black, predicted series – solid red, forecast lower and upper limits – dashed red, forecast – dashed blue)



forecast, the fitted model provides the forecast values in (Table 3):

This fitted model presents the best forecast for $h = 2$. Figure 4 presents the AR(3) fit to EVI observed series. By visual inspection, one can observe an acceptable fit. Applying the Ljung-Box test (`Box.test(., type = "Ljung-Box")`), there is evidence at any nominal level smaller than 92.92% that the associated residuals follow the GWN pattern.

Conclusion

In this chapter, we have presented the main concepts for understanding the time series analysis (TSA). Two specific models were discussed and applied to a time series of enhanced vegetation indices extracted from remote sensing radar.

References

- Almeida-Junior PM, Nascimento ADC (2021) ARMA process for speckled data. *J Stat Comput Simul* 91(15):3125–3153
- Brockwell PJ, Davis RA (1991) *Time series: theory and methods*, 2nd edn. Springer, Berlin Heidelberg/New York
- Didan K (2015) MOD13Q1 MODIS/Terra vegetation indices 16-day L3 global 250m SIN grid V006 [data set]. NASA EOSDIS LP DAAC. <https://doi.org/10.5067/MODIS/MOD13Q1.006>
- Gilgen H (2006) *Univariate time series in geosciences: theory and examples*, 1st edn. Springer, Berlin Heidelberg/New York
- Hamilton JD (1994) *Time series analysis*, 1st edn. Princeton University Press/Princeton
- Hyndman RJ, Koehler AB, Ord JK, Snyder RD (2008) *Forecasting with exponential smoothing: the state space approach*, 1st edn. Springer, Berlin Heidelberg/New York
- Maus V, Câmara G, Appel M, Pebesma E (2019) dtwSat: time-weighted dynamic time warping for satellite image time series analysis in R. *J Stat Softw* 88(5):1–31
- Shumway RH, Stoffer DS (2011) *Time series analysis and its applications*, 3rd edn. Springer, Berlin Heidelberg/New York
- Wei WWS (2006) *Time series analysis: univariate and multivariate methods*, 2nd edn. Pearson Addison Wesley/Boston

Time Series Analysis in the Geosciences

Klaus Nordhausen

Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland

Definition

A time series is a collection of data taken sequentially over time and has therefore a natural one-way order where it can be assumed that the current value is more influenced by the

recent past than by observations from a long time ago. Time series analysis comprises methods for the analysis of time series which take this structure into account when extracting meaningful statistics from the data. Time series forecasting concerns the prediction of future values based on an observed time series.

Time Series Data

In geosciences, collecting data that depend on time have been measured for centuries. For example, temperatures have been observed in standardized ways since approximately 1860. But also recordings of river discharge have a long tradition. Records of such phenomena varying irregularly in time and when sequentially observed are called time series. If the measurements are recorded continuously, like from a seismograph, the time series is called continuous time series. On the other hand, if the measurements are taken at certain, usually regular, intervals like hourly, daily, monthly, or yearly, they are denoted discrete time series. Discrete time series with equally spaced intervals are the most common form of time series data and also the focus of this entry.

At a given time point t , it is possible to observe a single observation like temperature; a vector of observations like temperature, precipitation, and wind speed; or a curve like the temperature curve at day t . Then one distinguishes between univariate, multivariate, and functional time series, respectively.

However, independent of the type of time series, the irregularly varying nature of the recordings is addressed by assuming the observed time series is a realization of a stochastic model. And the most convenient assumption is that the stochastic model has some time-invariant structures in which case it is called stationary. If however the stochastic model itself depends on time, it is called nonstationary. The last differentiation between types of time series made here is between linear and nonlinear time series. If the stochastic model can be expressed as the output of a linear model, it is called a linear time series and if not, a nonlinear time series.

Objectives of Time Series Analysis

Time series analysis has three main objectives:

1. **Description:** Methods that describe or summarize characteristics of the time series. Thus this contains tools for visualization or appropriate summary statistics which help to capture the essential features of the time series.
2. **Modeling:** To represent the stochastic nature of the observed time series, many different models have been suggested which help understanding the nature of the

process. An adequate model needs to be selected and its parameters need to be estimated. The choice of a model is often based on the descriptive measures.

3. **Prediction:** Due to the serial nature of the time series, future values can be predicted based on the past and present values. To predict the future, it is important to have a model which represents the data well.

Descriptive Tools for Time Series

The first step of all time series analysis is its visualization. A typical time series plot is given in Fig. 1 which gives the monthly average temperatures in centigrade (C) at the airport of Jyväskylä in central Finland from 1960 to 2005.

Usually in such a figure for a time series $x_t, t = 1, \dots, T$ one evaluates if the mean over time has a tendency, which is usually denoted as trend m_t and if there are regularly appearing patterns known as seasonalities s_t . If both are present, one can consider an additive decomposition of the time series

$$x_t = m_t + s_t + r_t, \quad t = 1, \dots, T,$$

where r_t represents the stochastic feature of data.

To obtain an idea of the trend, usually moving averages of order n are used, which are defined as

$$m_t^n = \frac{1}{n} \sum_{j=t-k}^{t+k} x_j,$$

where $n = 2k + 1$. The larger the n , the smoother the trend estimate will be.

The seasonal component is assumed to reoccur regularly. For example, for our monthly temperature data from Fig. 1, visual inspection indicates no trend, but the values every $l = 12$ months appear to be similar.

To obtain an estimate for a seasonal component of length l , one usually estimates first the trend with $n = l$ if l is odd like for weekly data or $n = 2l$ if n is even as for monthly data. The estimate of the monthly effect is then the average of each

month of the detrended series standardized so that their sum is zero.

The random reminder is then simply $r_t = x_t - m_t - s_t$.

For simplicity, we consider for now the observed time series x_t as a realization of a stochastic process $X_t, t = 0, \pm 1, \dots, \pm \infty$ without a trend or seasonal component, like the reminder r_t from above. The mean value function of X_t is $\mu_t = E(X_t)$ and the autocovariance function at times t and $t - \tau$ is defined as $Cov(X_t, X_{t-\tau}) = E((X_t - \mu_t)(X_{t-\tau} - \mu_{t-\tau}))$. For $\tau = 0$, this reduces to the variance at time t . Thus the serial dependence of time series data is captured especially in the autocovariance function. However, when all these characteristics depend on t , they are not very tractable and therefore a common practical assumption is that

$$E(X_t) = E(X_{t+\tau}) = \mu, \quad Var(X_t) = Var(X_{t+\tau}) = \sigma^2$$

and

$$Cov(X_t, X_{t-\tau}) = Cov(X_{t-l}, X_{t-l-\tau}) = C_\tau,$$

for all τ and all l . Thus the mean and variance do not depend on the time, and the autocovariance depends only on the distance between the observations. Such a time series is called (weakly) stationary. For an observed univariate time series $x_t, t = 1, \dots, T$ then summary statistics to describe these time series features are under the stationarity assumption:

Empirical mean: $\hat{\mu} = \frac{1}{T} \sum_{i=1}^T x_i$,

Empirical variance: $\hat{\sigma}^2 = \frac{1}{T} \sum_{i=1}^T (x_i - \hat{\mu})^2$ and

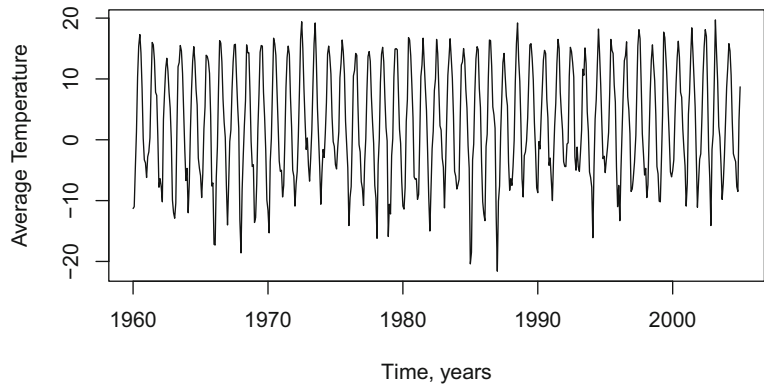
Empirical autocovariance: $\hat{C}_\tau = \frac{1}{T} \sum_{i=\tau+1}^T (x_i - \hat{\mu})(x_{i-\tau} - \hat{\mu})$.

In addition, also the empirical autocorrelation $\hat{\gamma}_\tau = \hat{C}_\tau / \hat{\sigma}^2$ is often considered.

When the analysis of the time series is based mainly on the autocorrelation structure, one says one does the analysis in the time domain. Another approach is to analyze the data in the frequency domain. Then by employing spectral analysis, the time series is decomposed into trigonometric functions at each frequency, and its features are represented with the weights given to the periodic components.

Time Series Analysis in the

Geosciences, Fig. 1 Average monthly temperature at the airport in Jyväskylä in C from 1960 to 2005



If the autocovariance function C_τ rapidly decreases with increasing lag τ and satisfies

$$\sum_{\tau=-\infty}^{\tau=\infty} |C_\tau| < \infty,$$

one can define the Fourier transform of C_τ . For frequencies $-0.5 \leq w \leq 0.5$, the power spectral density function is then

$$p(w) = \sum_{\tau=-\infty}^{\tau=\infty} C_\tau e^{-2\pi i \tau w}.$$

For an observed time series $x_1, \dots, x_T$ one uses then the natural frequencies $w_j = j/T, j = 0, \dots, \lfloor N/2 \rfloor$ to obtain the periodogram

$$p_j = \sum_{\tau=-T+1}^{T-1} \hat{C}_\tau e^{-2\pi i \tau w_j} = \hat{C}_0 + 2 \sum_{\tau=1}^{T-1} \hat{C}_\tau \cos(2\pi \tau w_j),$$

which can be used to obtain the sample spectrum

$$\hat{p}(w) = \sum_{\tau=-T+1}^{T-1} \hat{C}_\tau e^{-2\pi i \tau w_j}$$

by extending the domain to the continuous interval $[0, 0.5]$. Thus, there is a direct relation between the sample spectrum and the sample autocovariance

$$\hat{C}_\tau = \int_{-0.5}^{0.5} \hat{p}(w) e^{2\pi i \tau w} dw, \quad \tau = 0, \dots, T-1.$$

Notice that in practice, the periodogram is usually computed with the fast Fourier transform (FFT), and in the time domain, plots of the autocorrelation function are often inspected while in the frequency domain, the periodogram is considered. Which approach is preferable often depends on the concrete application. For more details on these topics, see, for example, Gilge (2006), Kitagawa (2010), and Shumway and Stoffer (2017).

Time Series Modeling

Exploring the autocorrelation structure and/or the periodogram of a time series provides the starting point for modeling.

A highly popular class of time series models are autoregressive moving average (ARMA) models. For this purpose, we define first a white noise process. A process ε_t is called white noise process if $E(\varepsilon_t) = 0$, $\text{Var}(\varepsilon_t) = \sigma^2$ and $\text{Cov}(\varepsilon_t, \varepsilon_{t-\tau}) = 0$ for all t and all $\tau > 0$. Thus a white noise process is centered with constant variance and serially uncorrelated. In the following, we will also assume that ε_t is Gaussian.

An ARMA model expresses then a time series x_t as a linear combination of its past values x_{t-i} and a white noise process ε_t :

$$x_t = \sum_{i=1}^p \alpha_i x_{t-i} + \varepsilon_t - \sum_{i=1}^q \beta_i \varepsilon_{t-i}.$$

In this model, p is the autoregressive order with autoregressive coefficients $\alpha_1, \dots, \alpha_p$ while q is the moving average order with moving average coefficients $\beta_1, \dots, \beta_q$. A time series x_t which follows an ARMA model is called an ARMA process and is usually denoted ARMA(p, q). The special case ARMA($p, 0$) is called simply an autoregressive process AR(p) and ARMA($0, q$) a moving average process MA(q). It is meanwhile well established that almost all stationary time series can be modeled using autoregressive processes.

Interesting is also that, under certain conditions, all ARMA(p, q) models can be expressed as MA(∞) models, i.e., as linear combinations of present and past realizations of a white noise series

$$x_t = \sum_{i=0}^{\infty} \phi_i \varepsilon_{t-i}.$$

The coefficients ϕ_i in this representation are called the impulse response function and are given by the recursive expression

$$\phi_0 = 1 \quad \text{and} \quad \phi_i = \sum_{j=1}^i \alpha_j \phi_{i-j} - \beta_i, \quad i = 1, 2, \dots,$$

where $\alpha_j = 0$ for $j > p$ and $\beta_j = 0$ for $j > q$.

The impulse response functions can then be used to express the autocovariance function of an ARMA(p, q) model

$$C_0 = \sum_{i=1}^p \alpha_i C_i + \sigma^2 \left(1 - \sum_{i=1}^q \beta_i \phi_i \right)$$

and

$$C_\tau = \sum_{i=1}^p \alpha_i C_{\tau-i} - \sigma^2 \left(1 - \sum_{i=1}^q \beta_i \phi_{i-\tau} \right), \quad \tau = 1, 2, \dots$$

The power spectrum of the same model is analogously

$$p(f) = \sigma^2 \frac{\left| 1 - \sum_{j=1}^q \beta_j e^{-2\pi i f j} \right|^2}{\left| 1 - \sum_{j=1}^p \alpha_j e^{-2\pi i f j} \right|^2}.$$

Useful in this context is also the partial autocorrelation $\gamma_{\tau\tau}$ of a stationary time series x_t which gives the correlation

between x_t and $x_{t-\tau}$ where the linear dependence on $\{x_{t-1}, \dots, x_{t-(\tau-1)}\}$ has been removed, i.e.

$$\hat{\gamma}_{\tau\tau} = \text{cor}(x_t, x_{t-\tau} | x_{t-1}, \dots, x_{t-(\tau-1)}).$$

Then plots of the autocorrelation and partial autocorrelations can give an indication of the AR and MA orders:

- For a stationary AR(p) process, the autocorrelations decay exponentially and the partial autocorrelations are zero for $\tau > p$.
- For a stationary MA(q) process, the autocorrelations are zero for $\tau > q$ and the partial autocorrelations decay exponentially.
- For a stationary ARMA(p,q) process, both, autocorrelations and partial autocorrelations, decay exponentially.

Figure 2 contains three samples of length 500 from an AR(1), MA(1), and ARMA(1,1) process with $\alpha_1 = \beta_1 = 0.6$ together with the corresponding estimated autocorrelation and partial autocorrelation functions. Values within the blue lines in the autocorrelation and partial autocorrelation plots are usually considered not significantly different from zero. Thus, the figure clearly shows how these functions can assist in order selection. Such visualizations of the (partial) autocorrelation function (acf) are often referred to as (partial) autocorrelogram.

As the periodogram contains the same information also its visualization can be used to get an idea of the order of ARMA processes, here the peaks and troughs of $\log(\hat{p}(f))$ are of interest. See Kitagawa (2010) for details. Rather than a graphical selection of the orders p and q , more common is to use model selection criteria such as AIC or BIC to select the order, and, for example, R (R Core Team 2020) has many tools to select ARMA orders and to estimate the coefficients.

Two concepts for AR and MA processes are still relevant. For an AR(p) process

$$1 - \alpha_1 x - \alpha_2 x^2 - \dots - \alpha_p x^p = 0$$

is called the AR characteristic equation. Similarly, for an MA(q) process

$$1 - \beta_1 x - \beta_2 x^2 - \dots - \beta_q x^q = 0$$

denotes the MA characteristic equation. An MA process is called invertible if all roots of the MA characteristic equation have an absolute value larger than 1. An AR process is only stationary when all roots of the AR characteristic equation also have absolute values exceeding 1. Thus in practical ARMA modeling to work in a stationary framework, it is

required that the AR part is stationarity and the MA part invertible. Invertibility for a time series means that we can transform it into an AR process, which means the time series can be expressed as a function of its past values. This is a key requirement when one is interested in predicting future values.

When we started out the modeling part, we assumed that the mean of the time series is constant and does not contain a trend. In many situations, however, where the nonstationary of the time series comes from the non-constant mean, taking differences of the time series makes it more stationary. The first-order difference is defined as $\nabla x_t = x_t - x_{t-1}$ and the d th-order difference repeats this d times and is denoted $\nabla^d x_t$. Using this approach to yield stationary ARMA processes was incorporated in the so-called autoregressive integrated moving average (ARIMA) models of order (p,d,q) which means that d th difference $\nabla^d x_t$ follows a ARMA(p,q) model. In most practical situations, d is either 1 or 2.

Figure 3 shows a sample of size 200 from an ARIMA(1,1,1) process with $\alpha_1 = \beta_1 = 0.6$ together with the first-order difference of the time series. While the original time series seems highly nonstationary, the differences look much more stable.

ARIMA models are nowadays the standard tool when starting time series models and can also be extended to allow for seasonal components. For further details about parameter estimation, model selection, and inference in the context of ARIMA models, see Brockwell and Davis (2020), Shumway and Stoffer (2017), and Hassler (2019).

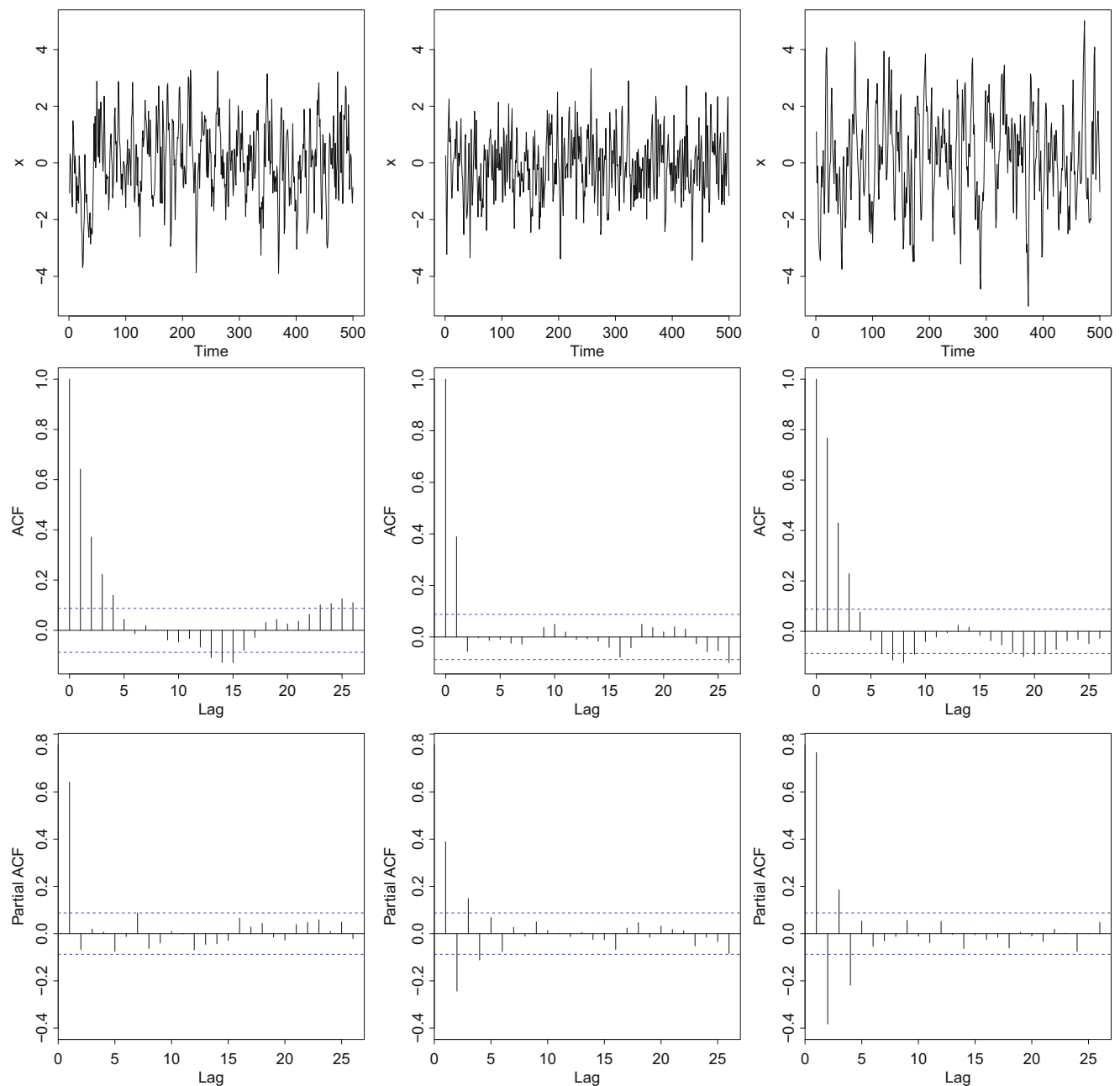
Time Series Forecasting

One of the primary reasons for time series analysis is the prediction of future values, which is usually called forecasting. Based on the available history of the process, $x_1, \dots, x_t$, the goal is thus to forecast the value x_{t+l} , where t denotes the forecast origin and l the lead time. The forecast itself is then denoted $\hat{x}_{t+l}$ and should be preferably the minimum square error forecast given by:

$$\hat{x}_{t+l} = E(x_{t+l} | x_1, \dots, x_t).$$

Clearly, without strong assumptions, such a prediction is difficult. However if the process is Gaussian, it is known that the best predictor is a linear function of the available history. Assuming in addition an ARIMA process yields practical expressions for the predictions and the possibility to evaluate the uncertainties in the predictions by providing confidence intervals.

For stationary ARMA processes with $E(x_t) = \mu$, it holds that $\hat{x}_{t+l} - \mu$ decays to zero as lead l increases. Which means that the further one wants to predict the future, the more naive the forecast, approaching simply the mean value. This is however natural as the dependence in an ARMA process



Time Series Analysis in the Geosciences, Fig. 2 The top row shows from left to right a sample of $T = 500$ observations from an AR(1) process, a MA(1) process, and an ARMA(1,1) process where $\alpha_1 = \beta_1 = 0.6$.

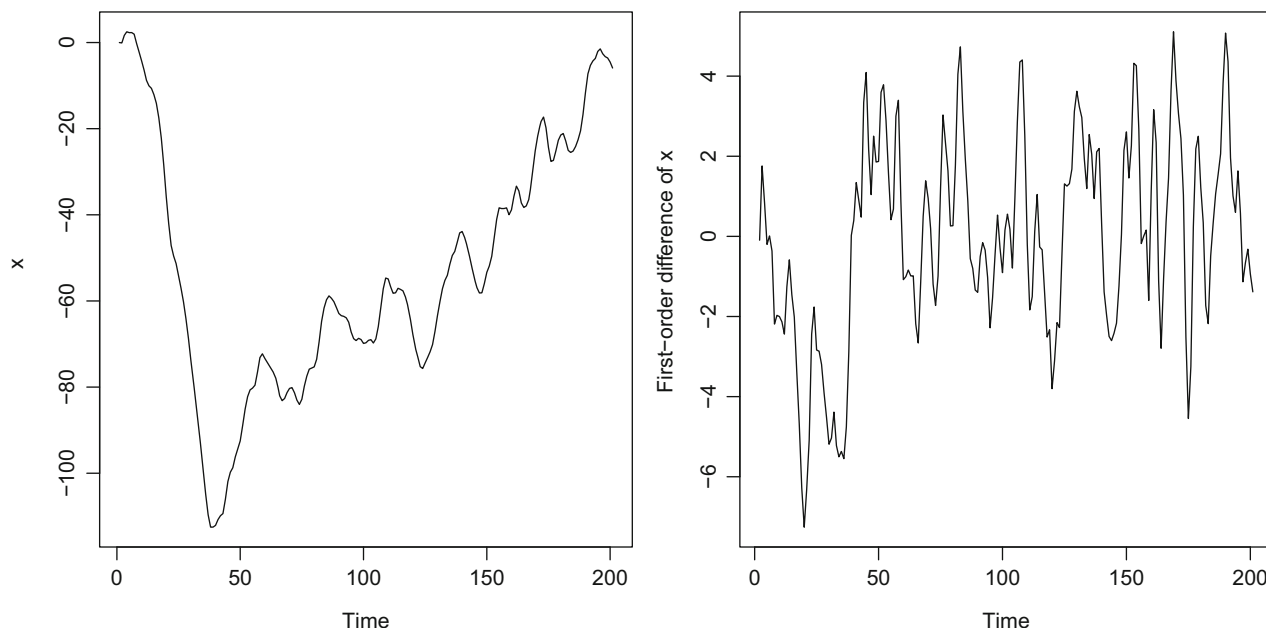
The second row shows the corresponding sample autocorrelation functions and the third the partial sample autocorrelation functions

dies out, and only for shorter leads we can improve the naive forecast and benefit from the available data. Details about ARIMA predictions are found in Brockwell and Davis (2020); Shumway and Stoffer (2017); and Hassler (2019). Using the average monthly temperature data from Fig. 1, a seasonal ARIMA prediction for the years 2006 and 2007 is shown in Fig. 4 as the blue line. The shaded gray area reflects the uncertainty of the predictions. The actual observed average temperatures are given as the dashed black line and shows

that the predictions are quite good, just forecasted less cold winters, but the true values are still within the gray areas.

Various Extensions for Univariate Time Series

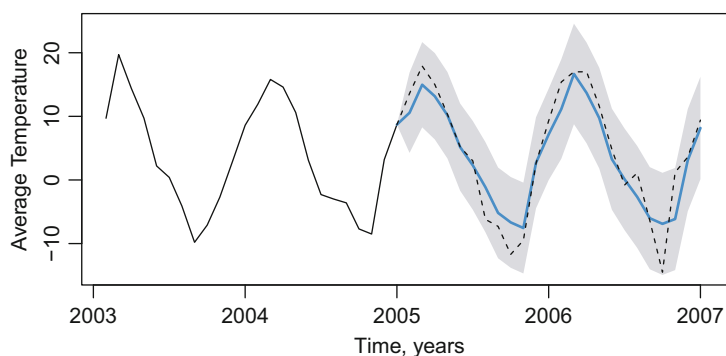
The ARIMA model with $d \in \{0, 1, 2\}$ with a possible seasonal adjustment is one of the standard tools in time series analysis. They are however not always suitable, and in the following, some different time series settings and models are discussed.



Time Series Analysis in the Geosciences, Fig. 3 The left panel shows a sample of size 200 from an ARIMA(1,1,1) process with $\alpha_1 = \beta_1 = 0.6$. The right panel shows the first-order difference of the same process

Time Series Analysis in the Geosciences, Fig. 4

The predicted average monthly temperature for the airport in Jyväskylä in C for 2006 and 2007 (blue line). The shaded area corresponds to the 95% prediction interval and the dashed black line are the observed average temperatures



Time Series with Long Memory. In the ARIMA model as discussed above, it assumed that the time series has a short memory which means that autocorrelation drops to zero after a certain lag or has an exponential rate of decay. For hydrological or meteorological time series like annual average rainfall, temperature, or river flow data, it can be however often observed that the autocorrelations are persistent over many lags. This persistence is then often denoted long memory. The extend of long memory is often quantified with the Hurst coefficient $H \in (0, 1)$ where 0.5 indicates the absence of long memory and a value larger than 0.5 indicates a persistent behavior. Long memory time series modeling can be embedded in an ARIMA framework by allowing the parameter d to take also fractional values, i.e., $d \in (0, 1)$, which is in detail discussed in Hassler (2019) and often denoted as fractional ARIMA modeling.

Nonlinear Time Series. The ARMA model introduced above is linear in its terms which is mathematically convenient and proved to be able to approximate many natural phenomena, however not all. Some processes in nature are just not linear. The literature on nonlinear time series models is quite large and an overview of different applications of nonlinear time series models in geosciences is, for example, given in Donner and Barbosa (2008). In the following, only two popular nonlinear time series models are introduced.

The simplest class of nonlinear models are piecewise linear models which are known as threshold models. In this class of models, it is thought that there is a set of K linear submodels and a mechanism that switches between the different submodels. A popular variant in this context is the threshold autoregressive (TAR) model which has the form

$$x_t = \alpha_0^{(j)} + \sum_{i=1}^p \alpha_i^{(j)} x_{t-i} + \beta^{(j)} \varepsilon_t,$$

where $j \in \{1, \dots, K\}$ is an indicator for the active submodel and the submodels have the parameters $\alpha_0^{(j)}, \dots, \alpha_p^{(j)}$ and $\beta^{(j)}$ and ε_t is a white noise process. Which submodel is active is indicated by the series z_t which takes values in $\{1, \dots, K\}$. If the state of Z_t depends only on the past values of x_t , i.e., $z_t = j$ if $x_{t-\tau} \in R_j$ for some fixed $\tau > 0$ with $\bigcup_{j=1}^K R_j = \mathbb{R}$, then it is a self-exciting TAR model. If however $z_t = j$ if and only if $y_{t-\tau} \in R_j$ where y_t is an observable or unobservable covariate time series, one says the TAR model is excited by y_t .

Another popular nonlinear model is the generalized autoregressive conditional heteroskedastic (GARCH) model. It can be formulated as

$$x_t = \sigma_t \varepsilon_t,$$

where ε_t is a white noise process and

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^p \alpha_i x_{t-i}^2 + \sum_{i=1}^p \beta_i \sigma_{t-i}^2$$

with model parameters $\alpha_i \geq 0, i = 0, \dots, p$, and $\beta_i \geq 0, i = 1, \dots, q$. Thus an ARMA(p,q) process is assumed for the variance of x_t , and for the series x_t , small values and large values should cluster.

Tests if nonlinear models are needed are well-established but have usually special nonlinear models as the alternative in mind. Similarly, parameter estimation and model selection for the above models are well investigated together with appropriate forecasting tools. For details about the above nonlinear models as well as many other methods, see, for example, Fan and Yao (2003).

Univariate time series methods from a geoscience perspective are, for example, also discussed in Gilje (2006).

Multivariate Time Series

So far, we have assumed that only one variable is measured at time t . For multivariate time series at the time t , a m -variate vector $\mathbf{x}_t = (x_1, \dots, x_m)^\top$ is observed. Thus, besides the serial dependence, also the dependence between the different variables need to be considered. The characteristics of the underlying stochastic process are then described using the expected value $E(\mathbf{X}_t)$, the covariance matrix $\text{Cov}(\mathbf{X}_t) = E((\mathbf{X}_t - E(\mathbf{X}_t))(\mathbf{X}_t - E(\mathbf{X}_t))^\top)$, and the autocovariance matrix $\mathbf{C}_\tau(\mathbf{X}_t) = E((\mathbf{X}_t - E(\mathbf{X}_t))(\mathbf{X}_{t-\tau} - E(\mathbf{X}_{t-\tau}))^\top)$. A key assumption is again that these quantities do not depend on the time, and multivariate (weak) stationarity is given when

$$E(\mathbf{X}_t) = \boldsymbol{\mu}, \quad \text{Cov}(\mathbf{X}_t) = \boldsymbol{\Sigma}$$

and

$$\begin{aligned} \mathbf{C}_\tau(\mathbf{X}_t) &= E((\mathbf{X}_t - \boldsymbol{\mu})(\mathbf{X}_{t-\tau} - \boldsymbol{\mu})^\top) \\ &= E((\mathbf{X}_{t-l} - \boldsymbol{\mu})(\mathbf{X}_{t-l-\tau} - \boldsymbol{\mu})^\top) = \mathbf{C}_\tau, \end{aligned}$$

for all τ and l . For an observed time series $\mathbf{x}_t, t = 1, \dots, T$ the corresponding sample statistics are:

$$\text{Empirical mean: } \hat{\boldsymbol{\mu}} = \frac{1}{T} \sum_{i=1}^T \mathbf{x}_i,$$

$$\text{Empirical covariance: } \hat{\boldsymbol{\Sigma}} = \frac{1}{T} \sum_{i=1}^T (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top$$

and

$$\text{Empirical autocovariance matrix: } \hat{\mathbf{C}}_\tau = \frac{1}{T} \sum_{i=\tau+1}^T (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_{i-\tau} - \hat{\boldsymbol{\mu}})^\top.$$

The empirical autocorrelation matrix is obtained as $\hat{\boldsymbol{\Gamma}}_\tau = D^{-1/2} \hat{\mathbf{C}}_\tau D^{-1/2}$ where the diagonal matrix D contains the diagonal elements of $\hat{\boldsymbol{\Sigma}}$.

The univariate ARMA model has been extended to the multivariate case and is usually referred to as vector autoregressive moving average (VARMA) model

$$\mathbf{x}_t = \sum_{i=1}^p \mathbf{A}_i \mathbf{x}_{t-i} + \varepsilon_t - \sum_{i=1}^q \mathbf{B}_i \varepsilon_{t-i},$$

where ε_t is an m -variate white noise process and the $m \times m$ matrices $\mathbf{A}_1, \dots, \mathbf{A}_p$ and $\mathbf{B}_1, \dots, \mathbf{B}_q$ are autoregressive and moving average coefficients, respectively. The special case VARMA(p,0) is correspondingly referred to as vector autoregressive process VAR(p) and VARMA(0,q) as vector moving average process VMA(q). The conditions for stationarity and invertibility translate similarly to the multivariate case as do order determination, and one can choose again to consider the problem from a time or frequency point of view. For details, see, for example, Wei (2019) and references therein. A big challenge in multivariate time series is however that already for small values of m , the number of parameters to estimate is quite considerable.

Therefore often dimension reduction methods are of interest in a time series context. On the one side is the dynamic factor model which is especially popular for financial data, and for details, we refer to Hallin et al. (2020). The approach presented here is based on blind source separation (BSS). In BSS, it is assumed that the observable m -variate time series $\mathbf{x}_t$ is a linear mixture of m latent unobservable processes $\mathbf{s}_t = (s_1, \dots, s_m)^\top$. The model can thus be written as

$$\mathbf{x}_t = \boldsymbol{\Omega} \mathbf{s}_t + \boldsymbol{\mu},$$

where the full rank $m \times m$ matrix $\mathbf{\Omega}$ is the mixing matrix and $\boldsymbol{\mu}$ is the m -variate location vector. The goal in BSS is to estimate $\mathbf{s}_t$ based on $\mathbf{x}_t$ alone. Clearly, for that more assumptions are needed. The two basic assumptions are $E(\mathbf{s}_t) = \mathbf{0}$ and $\text{Cov}(\mathbf{s}_t) = \mathbf{I}_m$. However, a third assumption is required, and this third assumption distinguishes between different BSS approaches. A review of BSS methods in the context of time series is, for example, Pan et al. (2021), and in the following only the main principles are introduced.

In the first approach, it is assumed that all m latent processes are weakly stationary linear time series which have different autocovariance functions, i.e., $\mathbf{C}_\tau(\mathbf{s}_t) = \mathbf{D}_\tau$ where $\mathbf{D}_\tau$ is a diagonal matrix where the diagonal elements depend on τ . This model is called the second-order source separation (SOS) model as the separation can be based on second moments alone. For example, AMUSE finds an unmixing matrix $\mathbf{\Gamma}$ as the matrix which simultaneously diagonalizes the covariance matrix and one autocovariance $\mathbf{C}_\tau$ in the sense that

$$\mathbf{\Gamma} \text{Cov}(\mathbf{x}_t) \mathbf{\Gamma}^\top = \mathbf{I}_p \quad \text{and} \quad \mathbf{\Gamma} \mathbf{C}_\tau(\mathbf{x}_t) \mathbf{\Gamma}^\top = \mathbf{\Lambda}_\tau,$$

where $\mathbf{\Lambda}_\tau$ is a diagonal matrix depending on τ . Such simultaneous diagonalization of two matrices can be solved via a generalized eigenvector-eigenvalue decomposition. The decomposition is however only unique if the autocovariances of $\mathbf{s}_t$ are all distinct at the chosen lag τ and therefore the performance of AMUSE depends a lot on its choice. To be less dependent on this choice, the SOBI method suggests to jointly diagonalize K autocovariance matrices plus the covariance matrix. This is nowadays considered a better approach and the added computational complexity is then negligible.

However, for example, in the case of GARCH processes, there is no information in the second moments and higher-order moments are of interest. The third assumption is then generalized in that way that the m latent processes are uncorrelated at all second and fourth (cross)moments, where cross-moments mean computing these moments between observations measured at different time points. In that case, methods like vSOBI and gSOBI extend SOBI by taking higher-order moments into account in the joint diagonalization process.

The third approach assumes that all m latent components have a constant mean, but their variances change over time and the constraint of identity holds only on average for the observed time span $1, \dots, T$. In that case, the model is called nonstationary source separation (NSS) and the idea for an unmixing matrix is to divide the observed time series into $K \geq 2$ non-overlapping intervals. The method NSS-JD then jointly diagonalizes the K covariance matrices obtained from the intervals under the constraint that the total covariance matrix must be the identity. Similarly, NSS-JD-TD chooses, in addition to the K covariance matrices, from the intervals additional autocovariance matrices for selected lags

computed on the separate intervals to be jointly diagonalized under the same constraint.

Given an unmixing matrix estimate, the components $\hat{\mathbf{s}}_t = \hat{\mathbf{\Gamma}}(\mathbf{x}_t - \hat{\boldsymbol{\mu}})$ are computed. Note that none of the BSS methods can recover actually the order of the components and their signs, but this is not relevant for the purpose of BSS which is threefold:

1. The estimated components might be easier to interpret and might give a better impression of the different underlying dynamics. For interpretations in relation to $\mathbf{x}_t$, the loadings from the unmixing matrix can be interpreted similarly as the loadings, for example, in PCA.
2. As the components are assumed uncorrelated, and often even the stronger assumption of independence is made, it is possible to model the m latent components separately from the others. Thus, no multivariate models need to be fitted but one fits m univariate models. This is often considered easier and more flexible.
3. It is often the case that not all m latent components are of interest and some might be simply noise. Thus dimension reduction can be performed by concentrating in the remaining analysis on the interesting components.

One type of data that is by definition multivariate is compositional data. Naturally, compositional data can also be observed over time which is usually denoted then as a compositional time series (CTS). To address the special nature of compositional observations, usually the compositions are first transformed into a coordinate system that follows a Euclidean representation, and then standard multivariate methods can be applied. The same strategy is usually also used for CTS. Kynclova et al. (2015) discuss vector autoregressive models in the context of CTS, while Nordhausen et al. (2021) investigate the use of blind source separation models in that context.

Functional Time Series

If an almost continuous time series can be divided into natural non-overlapping consecutive intervals, one can consider it as a series of ordered curves. Such a series of curves is known then as functional time series. Consider, for example, data coming from a magnetometer. Such a device measures almost continuously the magnetic field which depends heavily on the rotation of the earth. Hence it seems natural to look at the shape of these values based on the time of the day, i.e., the position of the sun, and therefor one has for each day one curve. Such a curve is then denoted $\chi_t(u)$, where t indicates the day and u the time of the day. Usually, it is assumed that the χ_t 's are elements of Hilbert spaces, and descriptive statistics are the functional mean, the covariance operator, and the autocovariance operator which can often be obtained by pointwise evaluations.

While the standard ARMA model has been extended to the functional setting, research mainly focuses on functional autoregressive (FAR) models for which order estimation and forecasting tools are well established. For further details, we refer to Kokoszka and Reimherr (2017) and references therein.

Software

As time series analysis is a common task in many fields, most statistical software packages provide tools to visualize and summarize time series. Similarly estimation and forecasting tools are usually also available. All the methods described here are, for example, available in R R Core Team (2020), and there is an R TASK VIEW dedicated to time series analysis listing the most important functions and packages; see <https://CRAN.R-project.org/view=TimeSeries>.

Summary

In geosciences many measurements are taken at regular intervals, and of interest is to understand the dynamics in the underlying processes and to have an idea of what will happen in the future. Depending on the nature of the data, there is a large selection of time series methods available. For simplicity, usually the modeling starts with stationary linear models. If these models are then not adequately describing the phenomena under consideration, different possibilities exist to address possible flaws like relaxing the requirements of stationarity and/or linearity.

Cross-References

- [Compositional Data](#)
- [Fast Fourier Transform](#)
- [Power Spectral Density](#)
- [Spectral Analysis](#)
- [Stationarity](#)
- [Stochastic Geometry in the Geosciences](#)
- [Time Series Analysis](#)

Bibliography

- Brockwell PJ, Davis RA (2020) Time series: theory and methods, 2nd edn. Springer, New York
- Donner RV, Barbosa SM (2008) Nonlinear time series analysis in the geosciences. Springer, Berlin
- Fan J, Yao Q (2003) Nonlinear time series. Nonparametric and parametric methods. Springer, New York
- Gilge H (2006) Univariate time series in geosciences. Springer, Berlin
- Hallin M, Lippi M, Barigozzi M, Forni M, Zaffaroni P (2020) Time series in high dimensions. World Scientific, Singapore

- Hassler U (2019) Time series analysis with long memory in view. Wiley, Hoboken
- Kitagawa G (2010) Introduction to time series modeling. CRC Press, Boca Raton
- Kokoszka P, Reimherr M (2017) Introduction to functional data analysis. CRC Press, Boca Raton
- Kynclova P, Filzmoser P, Hron K (2015) Modeling compositional time series with vector autoregressive models. *J Forecast* 34(4):303–314
- Nordhausen K, Fischer G, Filzmoser P (2021) Blind source separation for compositional time series. *Math Geosci* 53:905–924
- Pan Y, Matilainen M, Taskinen S, Nordhausen K (2021) A review of second-order blind identification methods. *WIREs Comput Stat* e1550
- R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
- Shumway RH, Stoffer DS (2017) Time series analysis and its applications. Springer, Cham
- Wei WWS (2019) Multivariate time series analysis and applications. Wiley, Hoboken

Tobler, Waldo

Michael Batty
University College London, London, UK



Fig. 1 Waldo Tobler (1930–2018), From Wikipedia © By-SA 3.0

Biography

Waldo Tobler (1930–2018) was born in Portland, Oregon of Swiss parentage. He grew up in Seattle, but at the beginning of World War II, his father who was a consular official in the Swiss Embassy, moved to Washington DC. At the end of the war, the family returned to Bern in Switzerland where Waldo joined the American Army. He learnt several new languages including Russian and he was in the one place where he could

begin to indulge his lifelong interest in maps. On his discharge from the Army, he returned to North America, attending the University of British Columbia as an undergraduate. In 1957 he went to the University of Washington, Seattle where he joined one of the most select groups of Ph.D. students that has ever existed within geography. Almost single-handedly, they led what came to be called the “Quantitative Revolution.” Under the tutelage of Bill Garrison, Waldo with his Ph.D. student colleagues Brian Berry, Richard Morrill, Duane Marble, John Nystein, Arthur Getis, among others, laid out the ground work for a generation who began to use mathematics and the scientific method in the construction of a grand theory. This tied location theory to social physics providing, along the way, powerful theories that would enable rigorous transformations of the way space could be ordered, articulated, and visualized.

Out of all this came spatial analysis and geographic information systems as well as demographic, economic, transportation and many other models of the urban and regional system. Computers were key to making these developments possible. Waldo’s early work was on map projections and cartograms and the skills that he built up enabled him to develop many different transformations of the way one could construct and project maps, color them, and visualize them. Drawing on concepts from classical physics, in particular models of fluid flow, one of the central concepts he developed was the idea of representing vectors of movement between places defining what he called “interaction winds.” This tied his work to spatial interaction models and in this sense he thought about how such models might be built for a world of flows between places, which, as we know, are largely asymmetric. Waldo of course being a cartographer, was one of the first to produce rapid visualizations, and in one of his path-breaking papers (Tobler 1970), he introduced a dynamic model of the Detroit region, which made extensive use of early computer graphics using storage tubes and I had the privilege of seeing the model in action when I visited him in Ann Arbor in 1974. The paper describing the model not only introduced cellular automata modeling into the geography but also coined what has come to be called Tobler’s Law. In the paper, the model he developed was based on the key spatial principle that “. . . everything is related to everything else, but near things are more related than distant things.” It was almost a throw-away-line but it quickly came to be enshrined as his “Law.”

His first appointment after his time in Seattle was at Michigan where he was a faculty member from 1961 to 1977. He did great work on cellular automata, flows, and on statistical map generalizations and variances during these years. When the new department of geography was resurrected at the University of California at Santa Barbara (UCSB) in 1979, he moved to join that august group where he spent the rest of his academic career. There he developed many interesting

and, often it would appear, idiosyncratic mathematical and statistical applications based on patterns and flows. But for me, his great work during these later years was his focus on migration. He resurrected Ravenstein’s Laws and I am the proud possessor of a copy of his enlargement of Ravenstein’s famous map of migration flows from the 1881 British census. Waldo digitized, enlarged, colored the vector flows by hand, and the map hangs in my office in University College. He knew I was interested in this as after I went to America in 1990 to the National Center for Geographic Information and Analysis, which was headquartered at UCSB, we talked a lot about our interests in visualization, cellular automata, and how we might get computers to make sense of all this. In fact, I first met Waldo in 1974 much earlier when he came to the University of Reading to look me up, on spec so it seemed, while he was in the UK at a conference in Nottingham. In fact, he was really visiting Peter Hall who on the day in question was nowhere to be seen. I could have never imagined him coming to see me. He was one of the greats whose papers I had read but who I simply assumed in those far off days, I would never meet. But I had published a paper called “Spatial Entropy” in the journal *Geographical Analysis* that year and he was intrigued. It so happened that I was just about to depart to spend a year at the University of Waterloo. I then visited him and Gunnar Olsson at Ann Arbor in Michigan that same year, not far away from Waterloo, thence traveling with them to the North American meetings of the Regional Science Association in Chicago. It was a baptism of intellectual fire.

I last saw him at Keith Clarke’s house in Santa Barbara in February 2013 and then in that same year in April at the Annual Meetings of the Association of American Geographers in Los Angeles. He told me he was not going to travel any longer to meetings and our paths did not cross again. He was, in fact, the founder of analytical cartography. He was one of the first geographers to be elected to the National Academy of Sciences but he was recognized most for being what one might call an “academics academic,” a very learned man with a subtle sense of humor, some might almost say “wicked” at times and a conviction in his research that was bolstered by his many other interests. These ranged from hiking in the lowlands of California to the Swiss Alps, from classical music to a wide interest in everyone else’s academic indulgences. He was a colleague who would give his time to talk to anyone who took an interest in his work and he would always reciprocate. A giant in the field.

Bibliography

- Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. *Econ Geogr* 46(Suppl):234–240

Topology in Geosciences

Florian Wellmann

Computational Geoscience and Reservoir Engineering,
RWTH Aachen University, Aachen, Germany

Definition

Topology is a sub-branch of geometry, concerned with relationships between objects that remain unchanged under certain types of deformation (such as stretching, bending, twisting, and shrinking). Objects that can be transformed into each other under such transformations are topologically equivalent. The two objects are then said to be homeomorphic. In a general sense, homeomorphism is a function that provides a mapping between two objects which are topological equivalent.

Introduction

Compared to *geometry*, topology is a relatively recent field of study. It emerged only in the eighteenth century based on the pioneering works of Leonhard Euler and Gottfried Leibniz, but became quickly an important and independent field of study (e.g., Earl 2019). While geometry is the study of objects in space, topology is the study of how objects are connected, in a more abstract point of view. In geology, these relationships are often derived from geological events in the local history – e.g., sedimentary deposition, magmatic intrusions, metamorphic overprinting, or deformation in tectonic events. Note that these relationships between surfaces (2-manifolds) are low-dimensional topologies. A simple general example is the connectivity in letters of the alphabet (Fig. 1a). The letters in each row (with this specific choice of font) can be transformed into each other by stretching and bending of the line segments. In topology terminology, the letters in each row are *homeomorphic* to each other. Note the important difference to the geometric term congruence: the letters cannot be transformed into each other through a combination of translation, rotation, or reflection. The important difference between the letters in each row is the number of *T-junctions*, highlighted with red dots in Fig. 1a: these junctions denote a fundamental difference between the letters that cannot be removed through processes of stretching and bending.

Another popular example is the math pun that a topologist cannot tell the difference between a coffee mug and a doughnut: both objects have one hole and can be transformed into each other through processes of stretching and bending, they are *topologically invariant* (Fig. 1b). This type of invariance

under stretching and bending has also led to the description of topology as “rubber sheet geometry.”

This abstract view of objects and their relationships has important applications in mathematics and physics, and the concepts have been applied to geoscientific investigations. Some basic principles and their applications are described in the following.

Euler’s Formula

The mathematician Euler was one of the first to investigate topological concepts with rigor in his study on the relationship between vertices, edges, and faces in closed polyhedra. In one of the first proofs in the field of topology, he could prove that there exists a fixed relationship in these objects, which is today known as Euler’s formula:

$$V(\text{vertices}) - E(\text{edges}) + F(\text{faces}) = 2 \quad (1)$$

This relationship is directly obvious when we consider a common cube in 3-D (Fig. 2a). It is easy to see that Euler’s formula is valid here, with 8 vertices, 12 edges, and 6 faces ($8 - 12 + 6 = 2$). When we insert a new vertex (Fig. 2b), we obtain triangular faces in addition to the quadrilateral ones in the regular cube, resulting in 9 vertices, 16 edges, and 9 faces ($9 - 16 + 9 = 2$). The principle is also valid for more complex objects, for example, a low-polygon version of the Stanford Bunny in (Fig. 2c) with 157 vertices, 464 edges, and 309 faces ($157 - 464 + 309 = 2$).

Other types of geometric objects can equally be identified by their value in the Euler formula, commonly referred to as Euler characteristic. Closely related to the Euler characteristic is the term genus, which, broadly speaking, describes the number of holes in a surface (a torus has genus 1, whereas a sphere has genus 0). More examples are provided in Table 1.

The fundamental idea to analyze systems based on relationships that remain unchanged under certain types of deformation such as stretching and folding and the related comparison of objects based on topological invariants can be applied to a range of geoscientific questions. Examples are outlined in the following.

Topological Analyses of Connectivity

Geoscientific analyses are often concerned with connections between objects, for example between bodies with different rock properties or faults in fault networks. Egenhofer and Herring (1990) provided a fundamental treatment of the possible relationships between two cells A and B in space based on the comparison of the interiors (A^0, B^0), the boundaries ($\partial A, \partial B$), and the exteriors (A^-, B^-) of these objects. With all possible combinations, this leads to the definition of the possible topological relationships in the *9-intersection-model* (9IM), also referred to as the Egenhofer-Matrix:

Topology in Geosciences,

Fig. 1 Examples of topological concepts: (a) letters with different numbers of junctions; (b) coffee cup and doughnut

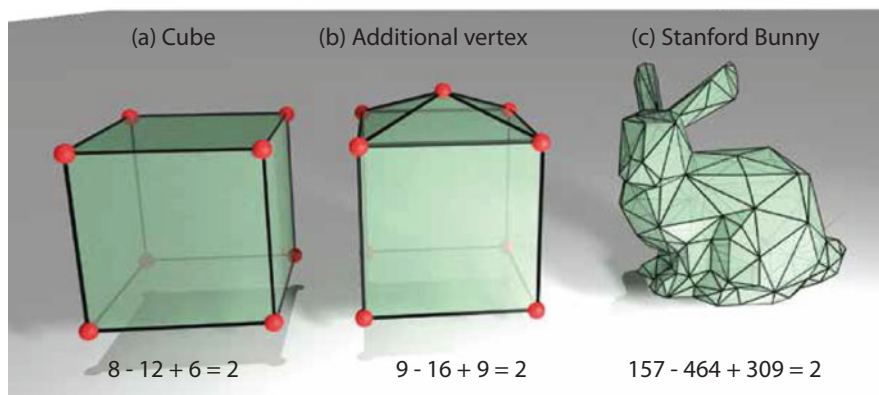
(a) Letters



(b) Coffee cup and doughnut

**Topology in Geosciences,**

Fig. 2 Euler's formula for different objects: (a) common 3-D cube; (b) cube with one inserted vertex; (c) low-polygon model of the Stanford bunny



$$w(A, B) = \begin{pmatrix} A^0 \cap B^0 & A^0 \cap \partial B & A^0 \cap B^- \\ \partial A \cap B^0 & \partial A \cap \partial B & \partial A \cap B^- \\ A^- \cap B^0 & A^- \cap \partial B & A^- \cap B^- \end{pmatrix} \quad (2)$$

The entries in this matrix are either true (T) or false (F). Cell relationships with the same intersection matrix are topologically equivalent. An extension of this concept that includes the dimensionality of the intersections has been proposed by Clementini and Di Felice (1996). This model has been adapted for spatial data query concepts and is implemented in GIS-systems.

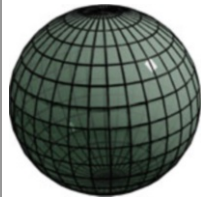
With the 9 Boolean entries in the matrix, a total of $2^9 = 512$ relationships would theoretically be possible. However, only a subset of these matrices is meaningful to describe topological regions. More specifically, if we limit the view to simple regions with a connected boundary separating interior and exterior parts in 2-D (i.e., regions that are homeomorphic to a unit closed disc), only 8 relationships are realizable (Fig. 3).

The same relationships exist between two 3-D regions in 3-D space (Zlatanova et al. 2004). Extensions to more general cases are possible (Li 2006).

Application to Structural Models

The most basic aspect in the context of geoscientific investigations often concerns the question if two cells (e.g., geological bodies) are connected or not: a typical example is a full space-filling tessellation with non-overlapping domains (also referred to as a water-tight discretization). In this case, a first-order relationship between discrete cells can be derived in the form of a connectivity graph (also referred to as topology graph), where nodes represent the cells and edges the neighboring relationship between two cells. Note that other representations are possible, for example on the basis of shared boundaries or nodes (see Thiele et al. 2016, for more examples).

Topology in Geosciences, Table 1 Euler characteristic numbers for some typical shapes

Type	$V - E + F$	Genus	Representation
Sphere (and all closed polyhedra)	2	0	
Torus (and, famously, a coffee cup) tabular	$0 - 2 + 2 = 0$	1	
Double torus	$0 - 4 + 2 = -2$	2	
Line with vertices	$2 - 1 + 0 = 1$	–	
Circle (line closed at same vertex)	$1 - 1 + 0 = 0$	–	
Disc	$1 - 1 + 1 = 1$	–	

The concept to use connectivity graphs (also based on the intersection model) has been applied to a range of investigations in the field of structural geology and geomodeling.

A typical example is provided in Fig. 4 with a topology graph (Fig. 4b) encapsulating the connectivity of geological bodies in a 3-D structural model framework (Fig. 4a).

The concept of connectivity graphs can be extended to consider the specific type of relationship between the objects. In the case of an analysis in the context of geological models, typical relationships would be a simple continuous stratigraphic succession, a contact related to an unconformity, an intrusive contact, or a connection across a fault (Thiele et al. 2016). On this basis, it is then also possible to distinguish a connectivity graph based on differences in lithology (lithology graph) from a connectivity graph for different structural domains, separated by faults.

In addition to the purely static analysis of connectivity in geological models, it is also possible to decipher the temporal relationship in the form of geological events. A fundamental framework for such a temporal geological event analysis has already been proposed by Burns (1975) and can be used to analyze geological event sequences from geological maps (e.g., Jessell et al. 2021). Such a topological analysis considering the time evolution can be considered as a temporal topology, in distinction to the purely spatial topology described above (e.g., Thiele et al. 2016).

Topology of Fault and Fracture Networks

Methods of topology have also been used to describe connectivity in fracture networks. The motivation to use topology here comes from the strong links between connectivity and flow, as an essential aspect in the analysis of fracture networks, which is often more important than the exact geometry of the fracture network itself (e.g., Manzocchi 2002; Sanderson and Nixon 2015; Sanderson et al. 2019). In 2-D space, Manzocchi (2002) defined different types of line intersections that are relevant in the characterization of a fracture network:

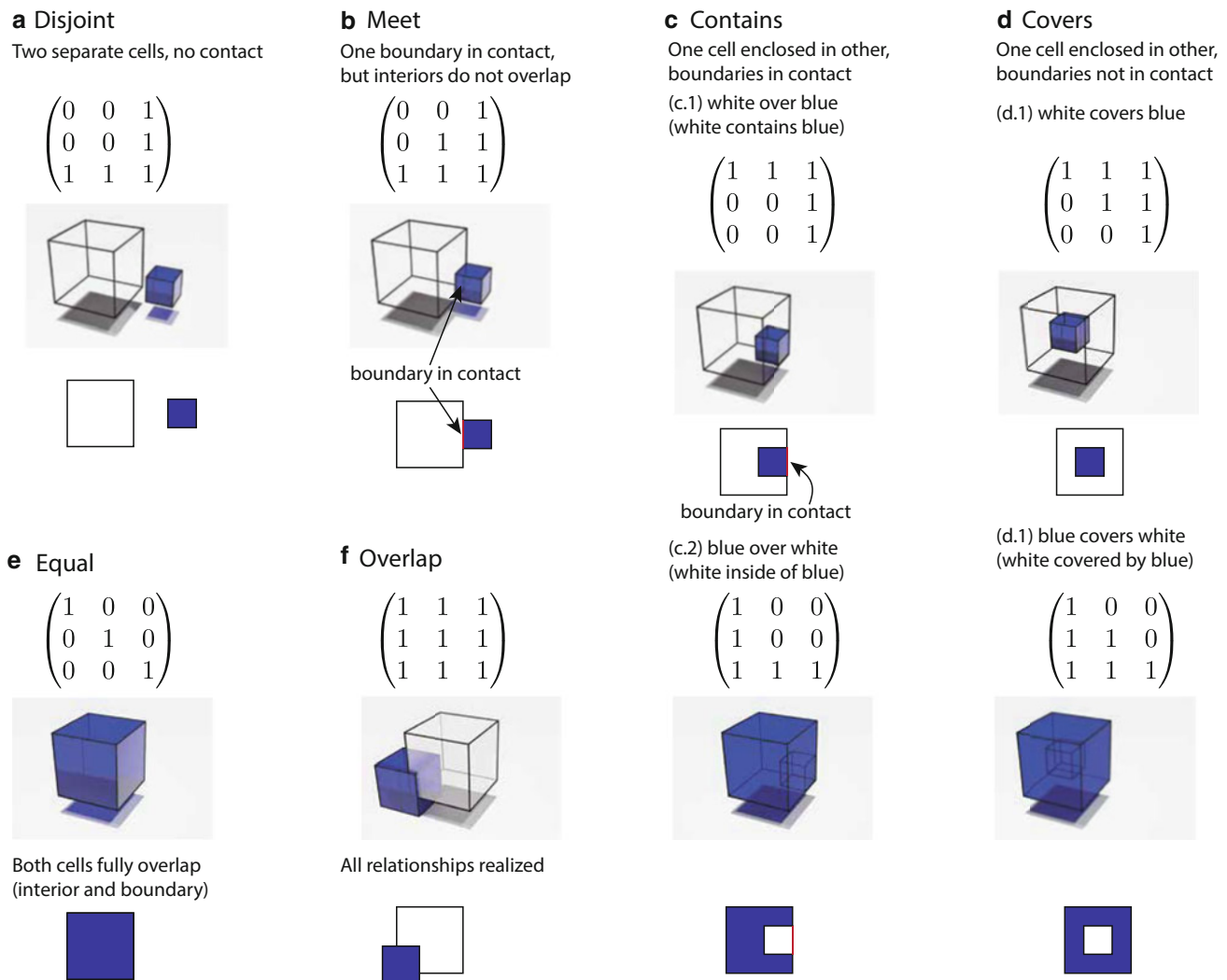
X-nodes points at intersections of two lines, where both lines extend beyond the intersection point itself.

Y-nodes (also: T-nodes): points at a T-shaped intersection between two lines (i.e., one line ending at the intersection point). This type of node is also an indication of an age relationship.

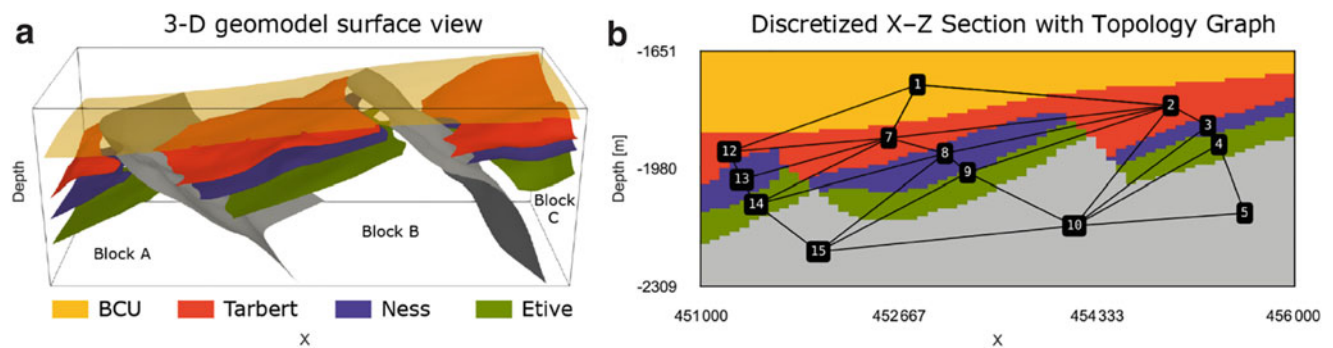
nodes isolated line end points.

These three node types can be related to different types of fracture systems and used as a classification Scheme. A simple example is presented in Fig. 5 with a fracture network analysis on the basis of these three types of nodes. The node ration can be visualized in a ternary diagram for comparison with other types of fault, joint, and fracture

T



Topology in Geosciences, Fig. 3 Topological relationships that can be realized between simple regions with examples in 2-D and 3-D and the corresponding matrix entries (1 = true, 0 = false)



Topology in Geosciences, Fig. 4 Example of a geological model (a) and the corresponding topology graph (b). (From Schaaf et al. (2021))

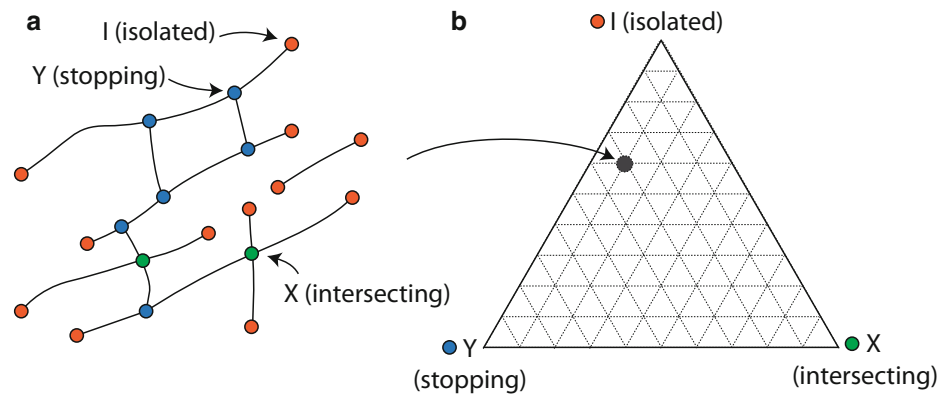
networks and with theoretical fault network models with different types of length distributions (Manzocchi 2002).

Based on these node classifications, a number of summary statistics can be calculated (e.g., Nyberg et al. 2018):

- Number of branches $N_B = \frac{N_I + 3N_Y + 4N_X}{2}$
- Number of lines $N_L = \frac{N_I + 2N_Y}{2}$
- Average connections per branch $C_B = \frac{3N_Y + 4N_X}{N_B}$

Topology in Geosciences,

Fig. 5 Example of a fracture network topology analysis: (a) schematic view of a fracture network with isolated nodes (red), nodes at truncated fractures (blue), and nodes at intersection points (green); (b) ternary diagram showing the node types



An important aspect here is that the important characteristics of these networks with respect to connectivity and, therefore, fluid flow would remain unchanged under purely continuous deformation of the space (i.e., only considering stretching, rotation, shearing, with no change in topology). This aspect enables then an investigation of connectivity and fluid flow through the investigated network in a previous point in time, where the geometry of the fracture network itself may have been different.

The topological description itself can also be extended to an analysis using methods from graph theory (Sanderson et al. 2019), for example, enabling a consideration of fracture width through edge weights, and directed graphs, if flow directions are known (e.g., on the basis of simulations).

Summary

Topology evolved over the last century from a mostly fundamental mathematical topic to an area of active research with a wide range of applications. In the field of geosciences, notable applications are in the area of network analyses for geological maps and models, fault and fracture networks, as well as river and Karst systems, and in geomorphological analyses. In all of these examples, the aim is to find a scientifically relevant abstraction of a geoscientific system using topological descriptions and methods. Examples are the simplification of a river system to inflow linkages or the abstraction of a complex geological model to cells which are in contact with each other. The motivation for the abstraction is here to limit the view to the dominating aspect that is relevant for fluid flow, for example, and to relate these essential characteristics to generic concepts. Similar aspects are behind the motivation to describe fault networks with connectivity graphs. Similar to the connectivity analyses shown above, topological connectivity analyses have also been used to analyze river networks (e.g., Mantilla et al. 2010) and connectivity in Karst

systems (e.g., Collon et al. 2017). Topological concepts are also widely used in a variety of operations in Geographic Information Systems (GIS) analyses (e.g., Phillips et al. 2015).

In a wider context, the field of topology goes far beyond these current applications in geosciences. This is evident in the 2016 Nobel Prize in Physics, which was awarded to a team of scientists who pioneered the use of topological concepts for a better understanding of phase transitions in condensed matter physics. Two important sub-fields of topology are differential topology and algebraic topology. Differential topology addresses the general topological study of smooth maps and manifolds. In contrast to differential geometry, the focus is not on geometric aspects, such as angles and curvature but on broader aspects such as the number of holes in a manifold. Important terms related to this topic are homotopy type and diffeomorphism groups. Algebraic topology is an extension of these concepts using methods from algebra to investigate topological spaces, extending the concepts of homology to homological algebra and linking methods of topology to algebra. The related field of knot theory, also a sub-branch of topology, investigates the embedding of knots from an abstract mathematical viewpoint, with applications to quantum physics or biological structure investigation. The transfer of these concepts to geoscientific research questions is surely an interesting path for future research.

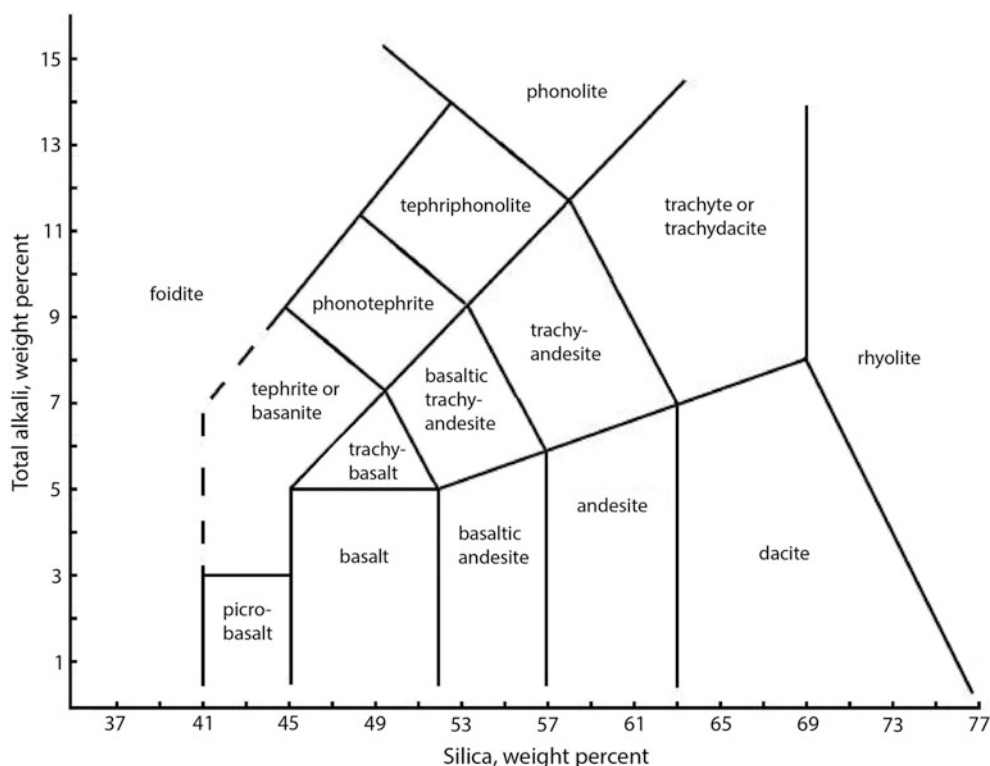
Bibliography

- Burns K (1975) Analysis of geological events. *J Int Assoc Math Geol* 7: 295–321
- Clementini E, Di Felice P (1996) A model for representing topological relationships between complex geometric features in spatial databases. *Inf Sci* 90:121–136
- Collon P, Bernasconi D, Vuilleumier C, Renard P (2017) Statistical metrics for the characterization of karst network geometry and topology. *Geomorphology* 283:122–142

- Earl R (2019) *Topology: a very short introduction*. Oxford University Press, Oxford
- Egenhofer MJ, Herring J (1990) Categorizing binary topological relations between regions, lines, and points in geographic databases. *The* 9:76
- Jessell M, Ogarko V, de Rose Y, Lindsay M, Joshi R, Piechocka A, Grose L, de la Varga M, Ailleres L, Pirot G (2021) Automated geological map deconstruction for 3d model construction using map 2loop 1.0 and map 2model 1.0. *Geosci Model Dev* 14:5063–5092
- Li S (2006) A complete classification of topological relations using the 9-intersection method. *Int J Geogr Inf Sci* 20:589–610
- Mantilla R, Troutman BM, Gupta VK (2010) Testing statistical self-similarity in the topology of river networks. *J Geophys Res Earth* 115 (F03038):1–12
- Manzocchi T (2002) The connectivity of two-dimensional networks of spatially correlated fractures. *Water Resour Res* 38:1–1
- Nyberg B, Nixon CW, Sanderson DJ (2018) Networkgt: a gis tool for geometric and topological analysis of two-dimensional fracture networks. *Geosphere* 14:1618–1634
- Phillips JD, Schwanghart W, Heckmann T (2015) Graph theory in the geosciences. *Earth Sci Rev* 143:147–160
- Sanderson DJ, Nixon CW (2015) The use of topology in fracture network characterization. *J Struct Geol* 72:55–66
- Sanderson DJ, Peacock DC, Nixon CW, Rotevatn A (2019) Graph theory and the analysis of fracture networks. *J Struct Geol* 125:155–165
- Schaaf A, de la Varga M, Wellmann F, Bond CE (2021) Constraining stochastic 3-d structural geological models with topology information using approximate bayesian computation in gempy 2.1. *Geosci Model Dev* 14:3899–3913
- Thiele ST, Jessell MW, Lindsay M, Ogarko V, Wellmann F, Pakyuz-Charrier E (2016) The topology of geology 1: topological analysis. *J Struct Geol* 91:27–38
- Zlatanova S, Rahman AA, Shi W (2004) Topological models and frameworks for 3d spatial objects. *Comput Geosci* 30:419–428

Total Alkali-Silica Diagram,

Fig. 1 Scatterplot defining the names of 15 different types of volcanic rocks. Tephrite or basanite are two alternative names for the same type of rock. The same is valid for trachyte or trachydacite



Total Alkali-Silica Diagram

Ricardo A. Olea

Geology, Energy and Minerals Science Center,
U.S. Geological Survey, Reston, VA, USA

Definition

The total alkali-silica (TAS) diagram is a scatterplot of the chemical concentrations of silica oxide (SiO_2) versus total alkali-sodium oxide (Na_2O) plus potassium oxide (K_2O) – in volcanic rocks (Fig. 1). This graph is used as a template for classifying volcanic rocks based on measured chemical concentrations when fine-grain size makes determining the mineralogic composition difficult to impossible. Under this provision, the TAS diagram is supported by the International Union of Geological Sciences Subcommittee on the Systematics of Igneous Rocks (Le Maitre 2002). As an example, if a volcanic rock has 49.8% SiO_2 and 4.1% alkali oxides, it is classified as basalt.

Properties

The TAS diagram is the results of international efforts and was preceded by several other similar diagrams postulated by various petrologists (Le Bas et al. 1992). By the time the TAS

diagram was released, there was a general consensus on the names of most types of volcanic rock. There were, however, problems with details. Over the years, Le Maitre and Ferguson (1978) had assembled a large database with over 15,000 analyses taken from the literature, each one with its geochemical composition and rock name assigned by the person(s) responsible for each analysis. Plots were prepared for each rock type to observe tendencies and dispersions. The centroid for each tile in Fig. 1 were selected to be as close as possible to the center of gravity for their corresponding point clouds, which were created by plotting information from the database on the same axes. It was further agreed that boundaries between tiles would have to be straight lines, and that the coordinates for their extremes, as much as possible, should be integer numbers.

Only rocks with analyses showing less than 2% water and less than 0.5% CO₂ were used in the preparation of the diagram. If both of those two components were present in concentrations below their respective cutoffs, the concentrations of water and CO₂ were discarded, and the rest of the proportions rescaled to add to 100%.

Summary and Conclusions

The TAS diagram is a graphical tool for the consistent naming of volcanic rocks. The diagram summarizes prior conclusions reached by multiple petrologists over several decades. The class boundaries are far from arbitrary since they represent the results of an international consensus.

Bibliography

- Le Bas MJ, Le Maitre RW, Woolley AR (1992) The construction of the Total Alkali-Silica chemical classification of volcanic rocks. *Mineral Petrol* 46(1):1–22
- Le Maitre RW, ed (2002) *Igneous rocks: a classification and glossary of terms: recommendations of the International Union of Geological Sciences Subcommission on the systematics of igneous rocks*, 2nd ed. Cambridge University Press, 236 pp.
- Le Maitre RW, Ferguson AK (1978) The CLAIR data system. *Comput Geosci* 4:65–76

Trend Surface Analysis

Frits Agterberg
Central Canada Division, Geomathematics Section,
Geological Survey of Canada, Ottawa, ON, Canada

Definition

Trend surface analysis is an early geoscientific application of the general linear model of mathematical statistics. It can be

regarded as the two-dimensional extension of polynomial curve-fitting. Suppose that the geographic coordinates of a point are written as u and v for the east-west and north-south directions, respectively. An observation made at a point k can be written as $Y_k = Y(u_k, v_k)$ to indicate that it represents a specific value assumed by the variable $Y(u, v)$. In trend surface analysis, it is assumed that $Y_k = T_k + R_k$ where $T_k = T(u_k, v_k)$ represents the trend and R_k is the residual at point k . T_k is a specific value of the variable $T(u, v)$ with

$$T(u, v) = b_{00} + b_{10}u + b_{01}v + b_{20}u^2 + b_{11}uv + \dots + b_{pq}u^p v^q$$

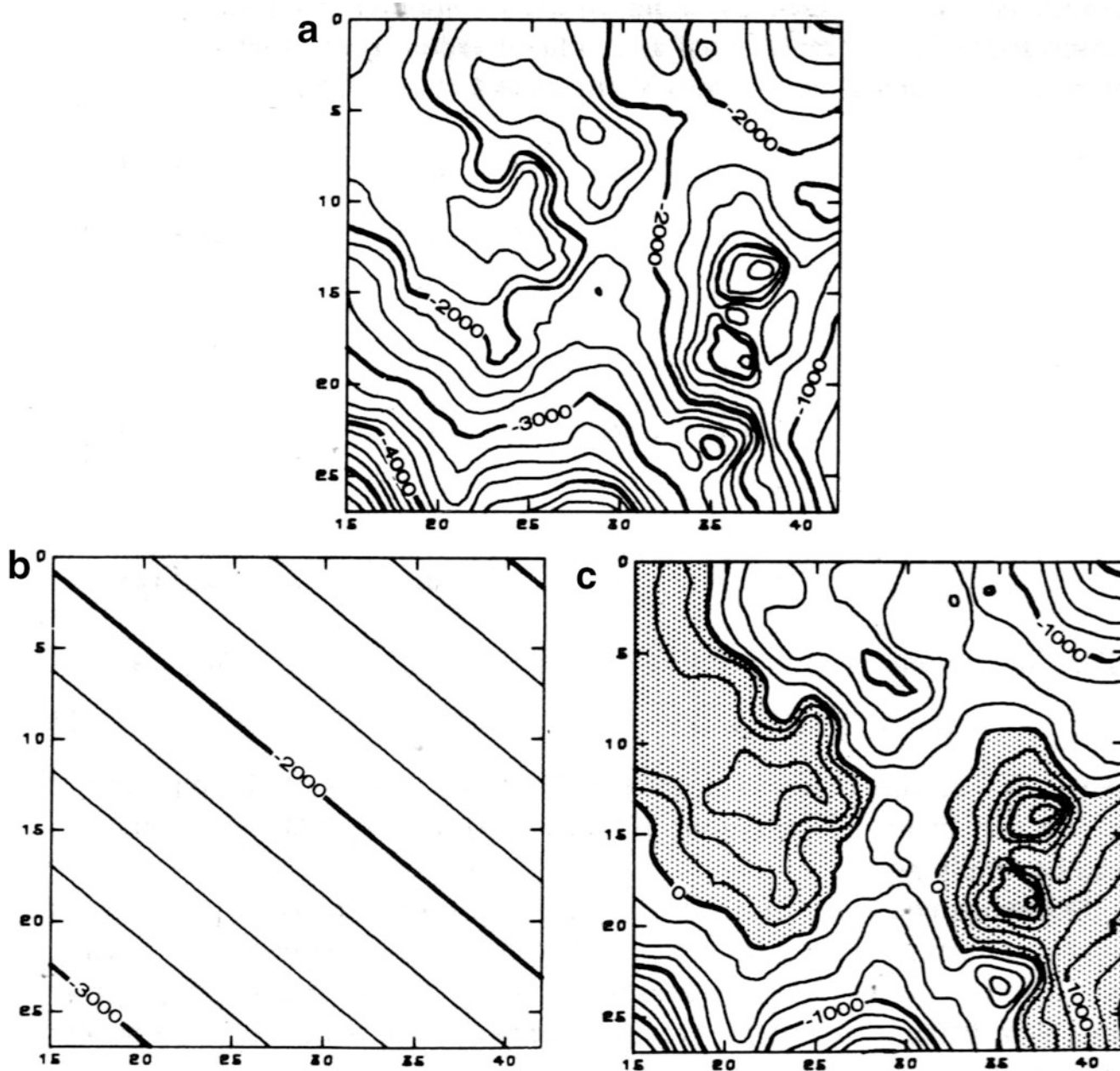
In general, u and v form a rectangular coordinate system. However, latitudes and longitudes can also be used. In specific applications, $p + q \leq r$ where r denotes the degree of the trend surface. Depending on the value of r , a trend surface is called linear ($r = 1$), quadratic ($r = 2$), cubic ($r = 3$), quartic ($r = 4$), or quintic ($r = 5$). Higher-degree surfaces have also been used. A trend surface of degree r has $m = 0.5 \cdot (r + 1) \cdot (r + 2)$ coefficients b_{pq} . It is possible to extend the preceding equations from two to three dimensions.

Introduction

The method of trend surface analysis is one of the early applications of the general linear model in the geosciences (Krumbein and Graybill 1965). In the late 1960s, this technique was competing with universal kriging originally developed by Huijbregts and Matheron (1971). Today, both techniques remain in use for describing spatial trends or “drifts” in variables with a mean that changes systematically in two- or three-dimensional space. Simple moving averaging as practiced by Krige (1966) or inverse distance weighting methods can be equally effective when there are many observations.

Trend surface analysis was one of the first computer-based methods widely applied in geophysics, stratigraphy, and physical geography in the 1960s. Initially, it was assumed that the residuals from a best-fitting trend surface should be independent and normally distributed but Watson (1971) clarified that polynomial trend surfaces are unbiased if the residuals satisfy a stationary random process model. An early example of trend surface analysis is shown in Fig. 1. Davis (1973) fitted the linear trend surface shown in Fig. 1b to the contours of the top of the Arbuckle Formation in Kansas shown in Fig. 1a. The residuals, of which the average value is equal to zero, are shown in Fig. 1c.

Other examples of 2D trend surface analysis will be given in the following sections. They include variations in mineral composition (enstatite in orthopyroxene) in the Mount Albert Peridotite Intrusion, eastern Quebec. A 3D extension of the method applied to Mount Albert specific gravity data shows



Trend Surface Analysis, Fig. 1 Top of Arbuckle Formation in central Kansas. Contours are in feet below sea level (after Davis 1973). Geographic coordinates are in arbitrary units. (Area shown measures 320 km

on a side.) (a) Contour map of original data; (b) Linear trend surface; (c) Residuals from linear trend surface. (Source: Agterberg 1984, Fig. 1)

that, geometrically, serpentinization of this peridotite body occurred along a northward dipping inverted pyramid (also see later). Trend surface analysis can be applied to data that have been subjected to transformation as will be illustrated for copper concentration values in the Whalesback Mine, Newfoundland. In another example, mosaics of trend surfaces fitted to contours of the bottom of the Matinenda Formation, Elliott Lake area, northern Ontario, show residuals that are pathfinders for occurrences of uranium deposits.

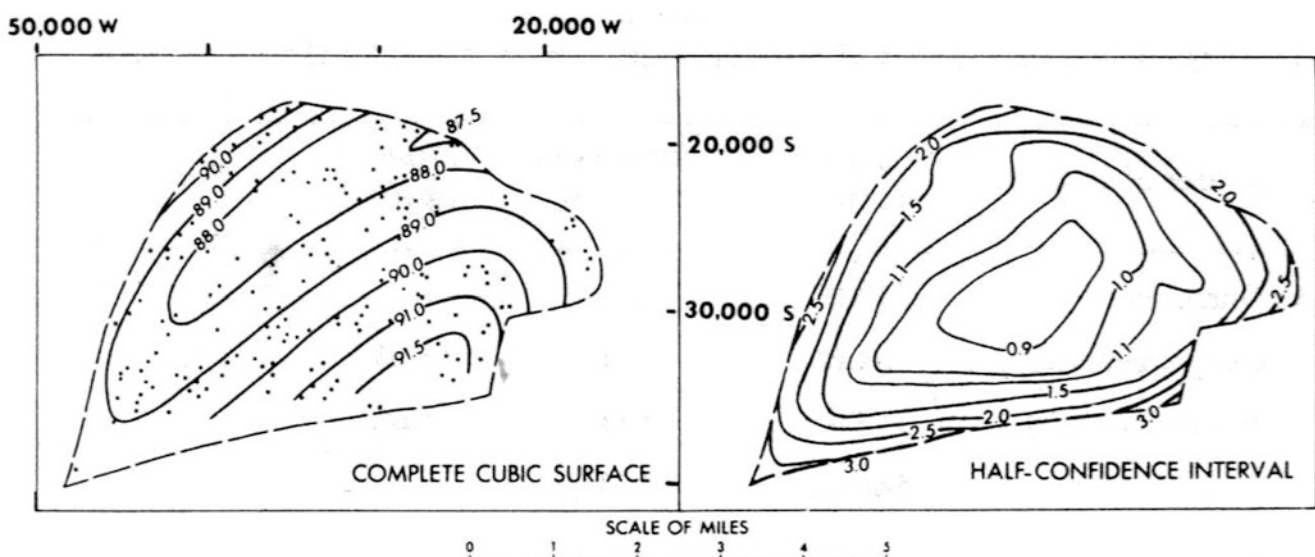
By means of these examples, it will be illustrated that trend surface analysis can provide meaningful results for establishing: (1) regional trends by eliminating local distortions that are either due to localized random distortions of broad regional patterns or to measurement errors and (2) extracting meaningful patterns of residuals retained after eliminating regional trends that obscure these local patterns which are of interest for mineral exploration.

Mount Albert Example

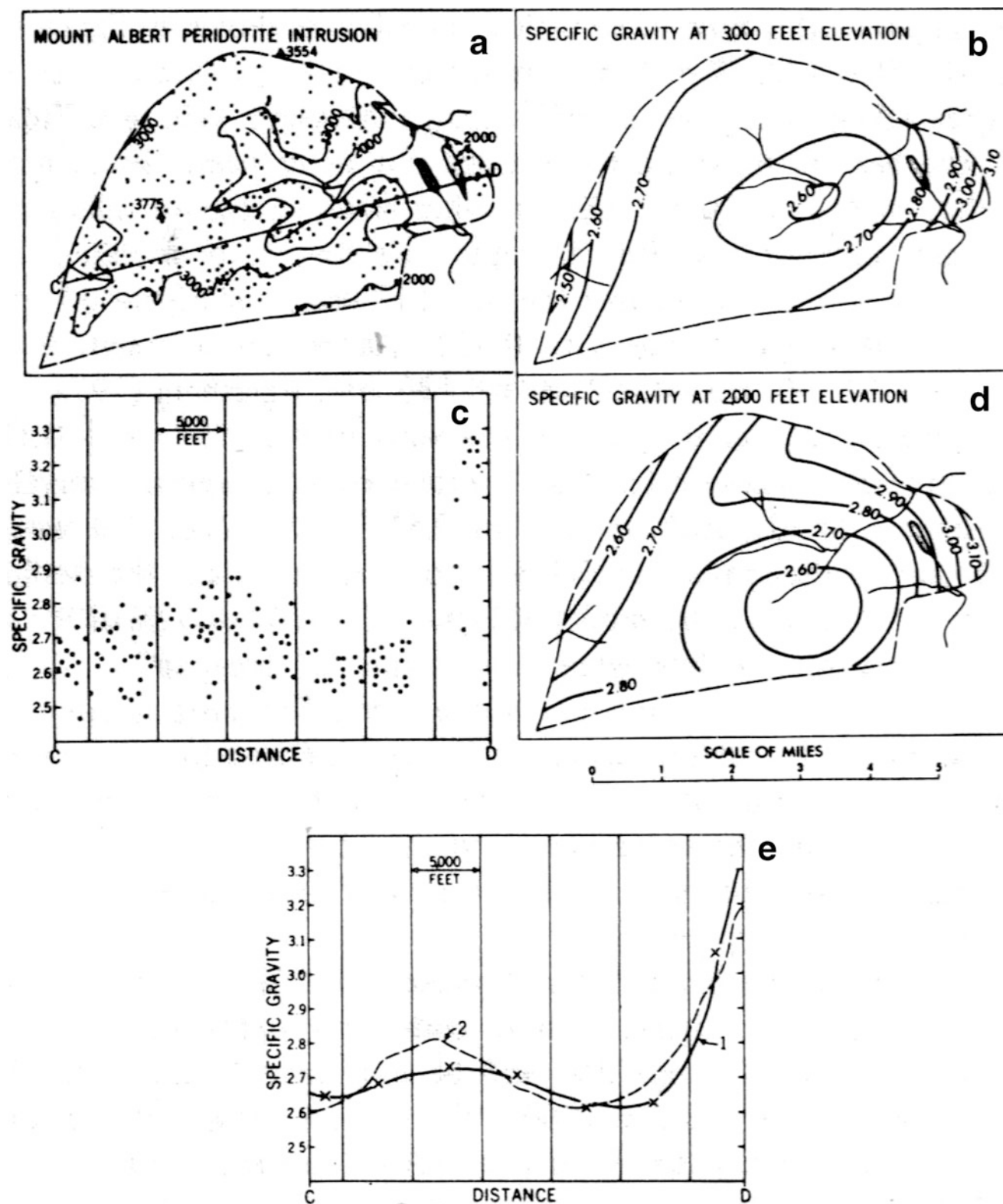
The Mount Albert intrusion is the largest ultramafic mass (approx. 44 km²) in the Gaspé (Quebec) portion of the so-called Appalachian ultramafic belt. It is probably 530 my old. The intrusion was mapped and sampled in 1959 by C.H. Smith and I.D. MacGregor, who made their data available to the author for performing trend surface analysis (Agterberg 1964). The petrography of the body is relatively simple. Prior to serpentinization, it consisted of from 80% to 90% olivine, the remainder being primarily orthopyroxene with some chrome spinel (up to 1%) and diopsidic clinopyroxene. It was attempted to collect rock specimens from the nodes of a rectangular grid with 1000 feet spacing. The actual number of specimens that could be collected was less because of overburden. The number of mineralogical determinations was further reduced by serpentine alteration. The following four variables were determined for as many collected specimens as possible: (1) cell edge d_{174} of olivine; (2) refraction index N_z of orthopyroxene; (3) unit cell dimension of chrome spinel; and (4) specific gravity of the whole rock. The first two variables were converted into percentage magnesium in olivine and orthopyroxene, respectively, and the results reported as mol. percent forsterite (Mg-olivine) and enstatite (Mg-orthopyroxene). The relationship between percent serpentine and rock density in samples from the Mount Albert intrusion is approximately linear. Serpentinization was a process that continued after the ultramafic body was already in place. The specific gravity measurements range from 2.5 (nearly pure serpentine) to 3.3 (unaltered peridotite). In most of the study area, average specific gravity is about 2.7 indicating that, overall, about 75% of the original peridotite was altered into serpentinite.

For example, the cubic trend surface for percentage enstatite in orthopyroxene is shown in Fig. 2, along with its 95% confidence interval. It suggests that there occurs a %Mg maximum in the southeastern corner of the intrusion. The shape of the 95% confidence interval for this trend surface mirrors the boundary of the intrusion because all observation points are contained within this boundary. It confirms the existence of the elongated ENE-WSW trending %Mg minimum but not necessarily the maximum in the southeastern corner of the intrusion. There is a pocket of practically unaltered peridotite in the eastern part of the body where the transition from high-density to lower-density material is relatively rapid. Although the Mount Albert intrusion was originally a peridotite, later it became primarily a serpentinite body. Recently, serpentinites have become important as a probable means of widespread rapid carbon mineralization for permanent disposal of anthropogenic carbon dioxide emissions (Matter and Stute 2016). This warrants more detailed studies of the shapes of such bodies (cf. Deschamps et al. 2013) (Fig. 3).

A schematic contour map for Mount Albert elevation is given in Fig. 3a. A cross-sectional CD was constructed and in Fig. 3b all observations within 2500 ft from the section line were projected onto it. Then, average values for blocks measuring 5000 ft on a side were calculated (Fig. 3c). These block averages almost exactly coincide with the intersection of a separately fitted quintic trend surface with the cross-sectional CD. The percentage explained sum of squares (ESS) of the quintic trend surface amounts to ESS = 56.0%. This shows that (1) the quintic trend surface provides a good fit to the specific gravity trend and (2) a trend surface also can be



Trend Surface Analysis, Fig. 2 Outline of Mount Albert peridotite intrusion. Cubic trend surface for percent enstatite in orthopyroxene is shown together with locations of specimens and 95% half-confidence interval. (Source: Agterberg 1974, Fig. 43)



Trend Surface Analysis, Fig. 3 (a) Schematic contour map of Mount Albert Peridotite; elevations are in feet above sea level. (b) Variations in specific gravity along section CD (see A for location); all values within 2500 ft from section line were perpendicularly projected onto it. (c) and (d). Cubic trend surfaces for 3000-ft and 2000-ft levels obtained by contouring three-dimensional cubic hypersurface in two horizontal

planes. In most of the area, today's topography resembles the 3000-ft pattern more closely than the 2000-ft pattern. (e) Crosses represent averages for values within blocks measuring 5000 ft on a side (see b). Curve 1 is intersection of quintic trend surface (Fig. 7.6) with CD; curve 2 represents variation along topographic surface (see a) according to three-dimensional cubic hypersurface (Source: Agterberg 1974, Fig. 46)

obtained by the relatively simple method of moving averages. The method of moving or running averages consists of calculating arithmetic averages for a large number of overlapping blocks and contouring the results. Obviously, such method can only be applied when there are many observations. If few observations are available, trend surface analysis is to be preferred.

Because as many as 359 data points were available for specific gravity, 3D trend analysis can be attempted by incorporating the elevation of each observation point. The 3D cubic trend equation contains 19 explanatory variables ($u, v, w, u^2, uv, uw, v^2, vw, w^2, u^3, u^2v, u^2w, uv^2, uvw, uw^2, v^3, v^2w, vw^2$, and w^3) and 20 coefficients. Its ESS-value amount to 55.1% which is close to the previously mentioned ESS = 56.0% for the quintic trend surface. Calculated values for the 3D cubic along the topographic surface are shown for the CD profile in Fig. 3e. Because all observation points lie in the topographic surface, the elevation w could be approximated by a 2D polynomial in u and v . Consequently, a 3D trend $f(u, v, w)$ becomes a 2D trend $g(u, v)$ if w , as contained in $f(u, v, w)$, is replaced by its polynomial in u and v . The reliability of the 3D trend $f(u, v, w)$ decreases rapidly below the topographic surface. Figures 3c and d show extrapolations of the 3D cubic trend in the vertical direction obtained by setting w equal to 2000 ft and 3000 ft, respectively. This, of course, results in two ordinary cubic trend surfaces as shown in these two figures. Comparison with the contour map for the topographic surface (Fig. 3a) suggests the following interpretation. As mentioned before, specific gravity on Mount Albert is, by approximation, linearly related to volume percentage serpentine. Thus, a low value indicates a soft rock relatively sensitive to erosion. Two of the three rivers originating on Mount Albert follow zones of weakness at the 3000-ft. level rather than at the 2000-ft level. Further comparison with the contour map for elevation suggests present-day topography was controlled by the distribution of less weathering resistant rocks at higher levels. It can be concluded that the pipe of maximum serpentinization moved northward as well as upward.

Modeling of Residuals

As mentioned before, Watson (1971) pointed out that trend surface analysis produces valid result if the residuals are autocorrelated according to a functional relationship that is independent of location within the study area. In this section two examples of such autocorrelated residuals will be given. Both examples are based on a set of exploratory boreholes. The first is for copper concentration values in adjoining samples taken from along subhorizontal boreholes on the 425-ft mining level in the Whalesback copper deposit, Newfoundland. The second example consists of depth

measurements for the bottom of the Matinenda Formation, Elliott Lake area, northern Ontario, measured in boreholes drilled from the topographic surface downward for the purpose of finding uranium deposits.

The Whalesback copper deposit near Springdale, Newfoundland, provides an example of trend surface analysis in a situation that the values are autocorrelated and have a frequency distribution that is positively skewed with a long tail of large values. The orebody which later was mined out consists of chloritic altered volcanics containing pyrite and chalcopyrite mainly in disseminated form. However, these minerals also tend to cluster and form stringers and occur in well-defined narrow veins with maximum width of 1.2 m. The deposit coincides with a shear zone. It is about 400 m long at the surface. Average width of the central copper zone which averages more than 1% copper is approximately 20 m, and vertical depth is 270 m. The copper zone is enveloped by a rather strongly altered chlorite zone with 0.35% Cu on the average.

Trend surface analysis (Agterberg 1964) was applied to copper concentration values from the 425 ft level that occurs 425 ft below the surface. The core for underground drill holes, which were 50 ft apart, was divided into pieces of 5 ft length, and assay values (percentages copper) were determined from these. About 15 volume percent of the rocks in the deposit consist of dikes that do not contain copper. This yields zero assay values which were omitted from the data to be used for statistical analysis and trends will be projected across the dikes. Therefore, if average grade values are to be determined for larger blocks of ore, these values, after calculation, should be reduced by 15% for dilution caused by barren dikes. Of course, more precise corrections can be made in places where the precise location of dikes is known.

Because trend surfaces are imprecise at the edges of clusters of observation points, overlapping sets of six drill holes, or crosscuts which have been sampled in the same manner, were used, and quadratic and cubic trend surfaces $S(u, v)$ were fitted to logarithmically transformed copper values. The fitted values were converted back to ordinary copper values by using the transformation $T(u, v) = \exp [S(u, v) + s^2/2]$ where s^2 is the residual variance for the $S(u, v)$ surfaces. The central part for each converted surface $T(u, v)$ for a 150 ft wide zone situated between the second and fifth hole in each situation is shown in the mosaic of Fig. 4 for the cubic trend surface. This transformation assumes that at any location the copper values are lognormally distributed with mean $T(u, v)$ and logarithmic variance s^2 . Figure 4 shows histograms of the natural logarithms of the original 516 copper values and also of the residuals from the cubic trend surface. Autocorrelation of residuals affects the results of trend surface analysis. The variance of the mean of n autocorrelated data from a series with the first order Markov property satisfies

$$\begin{aligned}\text{var}(\bar{X}) &= n^{-2} \sum_{i=1}^n \sum_{j=1}^n \text{cov}(X_i, X_j) \\ &= \sigma^2 \left[1 + 2 \frac{\rho}{1-\rho} \left\{ (1-n^{-1}) - \frac{\rho}{1-\rho} (1-\rho^{n-1}/n) \right\} \right] / n\end{aligned}$$

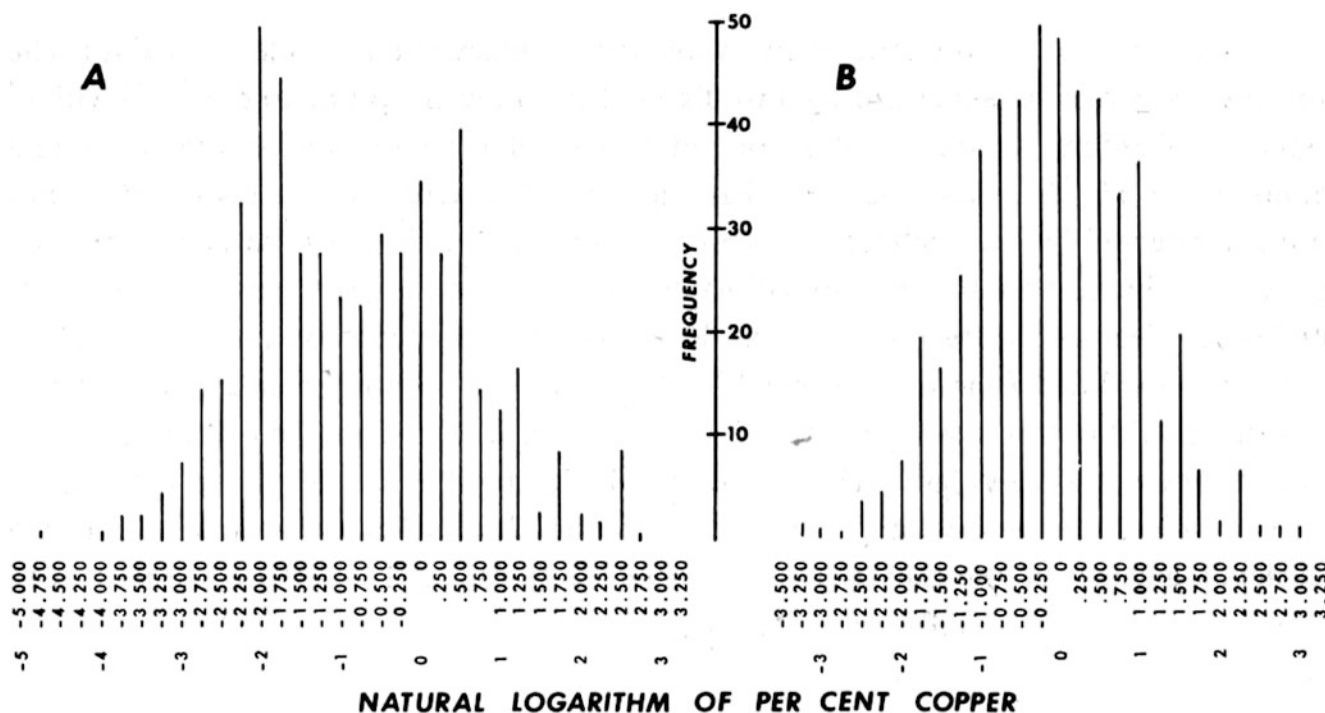
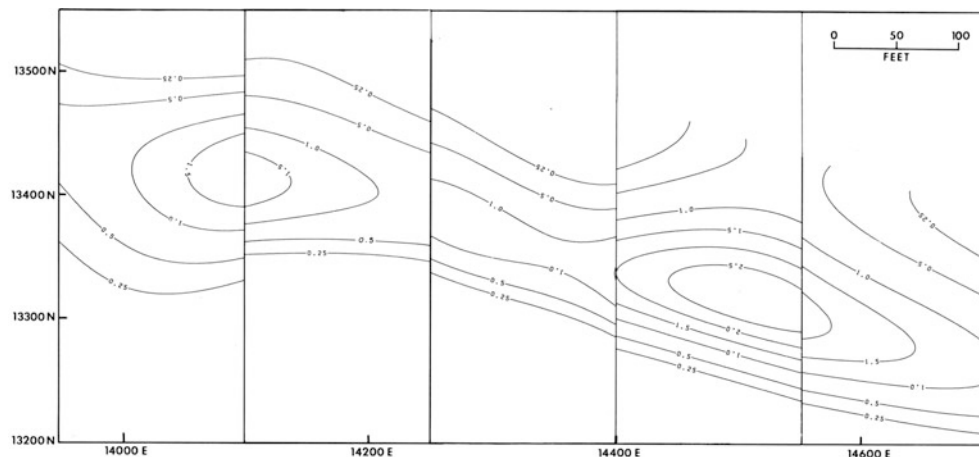
where ρ is the autocorrelation coefficient for adjoining values (cf. Cressie 1991, p.14). Suppose an equivalent number of independent data (n') is defined as $\text{var}(\bar{X}) = \sigma^2/n'$. For $n > 10$, the preceding equation then gives approximately

$$n/n' = 1 + 2r_1 \left[n/(1-r_1) - 1/(1-r_1)^2 \right] / n$$

(cf. Agterberg 1974, p.302). For large n this gives $n' \approx n(1-r_1)/(1+r_1)$ illustrating that autocorrelation strongly affects statistical inference even in large samples. In Agterberg (2014) it is discussed in more detail that (a) the residuals in Fig. 5b are approximately normally distributed and (b) a good estimate of the autocorrelation of the residuals

Trend Surface Analysis,

Fig. 4 Mosaic of exponential cubic trend surfaces for copper concentration value on the 425-ft level of Whalesback Mine; based on underground borehole data; drill holes were north-south oriented, approximately 50 ft apart. (Source: Agterberg 1974, Fig. 49)



Trend Surface Analysis, Fig. 5 (a) Histogram of natural logs of 516 copper values; (b) ditto for residuals from cubic trend surfaces shown in Fig. 4; this frequency distribution is approximately Gaussian. (Source: Agterberg 1974, Fig. 51)

is $r_1 = 0.50$. Consequently, $n' = n/3 = 172$ illustrating the strong effect of autocorrelation of the residuals.

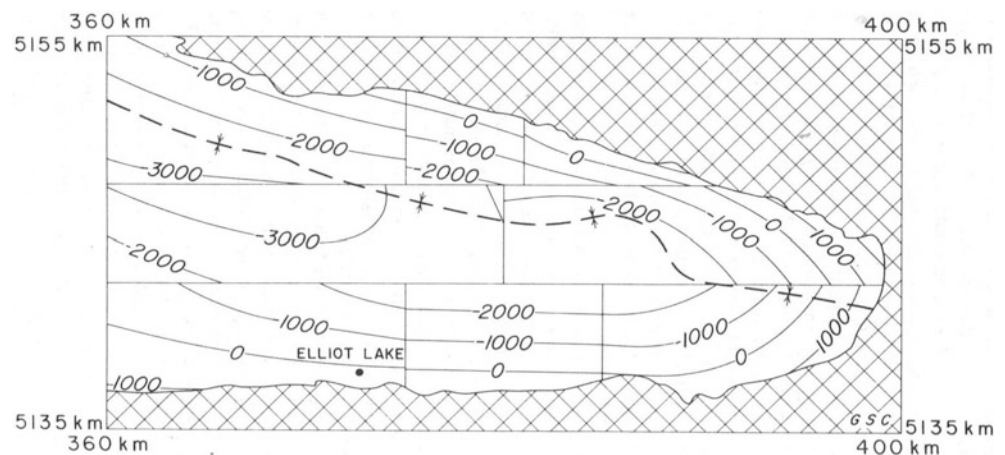
The second example of the preceding trend+signal+noise model is for the Elliot Lake area (Figs. 6, 7, and 8). The following is a brief summary of this statistical analysis previously published in Agterberg (1984). Figure 6 shows contours of parts of eight quadratic trend surfaces fitted to depth measurements of the bottom of the Matinenda Formation in boreholes of which the locations are shown in Fig. 7. In this area the Archean basement rocks of the Canadian Shield are unconformably overlain by Proterozoic rocks which later were subjected to folding. This resulted in the Quirke Syncline of which the axis is shown as a trace in Figs. 6, 7, and 8. Locally, the bottom of the Matinenda Formation contains uranium mineralization. In Fig. 7 the squares indicate boreholes in which significant uranium mineralization was encountered and the triangles represent other boreholes. The quadratic trend surfaces were fitted to these data from overlapping clusters of observation points. Figure 8 shows the

sum of this trend and the signal. The strong similarity of patterns in Figs. 6 and 8 illustrates that changes in magnitude of the signal (Fig. 8) are small in comparison with those of the trend (Fig. 6).

Robertson (1966) had attempted to reconstruct the topography of the top of the Archean basement as it was before the later structural deformation. He assumed that the Matinenda formation was deposited in topographical depressions in the top of the basement. This geological model provides a good reason for application of the trend+signal+noise model introduced in the previous section. Ideally, the trend in this model would represent the structural deformation and the signal would represent the original topographical surface on which the Matinenda Formation was deposited. Thus, it would be meaningful to extract the signal from the data by elimination of (1) the trend and (2) the noise representing irregular local variability which is unpredictable on the scale at which the sampling (by drilling) was carried out.

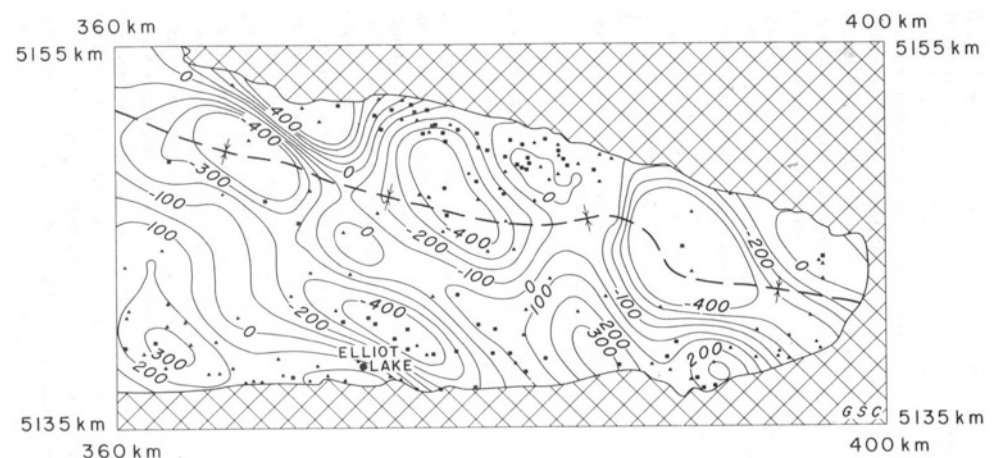
Trend Surface Analysis,

Fig. 6 Contours (in feet above sea level) of parts of eight quadratic trend surfaces fitted to separate parts of bottom of Matinenda Formation, Elliot Lake area. Trace of axial plane of Quirke Syncline is shown also. Archean basement rocks are shown by pattern. Universal Transverse Mercator grid locations of corner points of map area are shown in km. (Source: Agterberg 1984, Fig. 8)



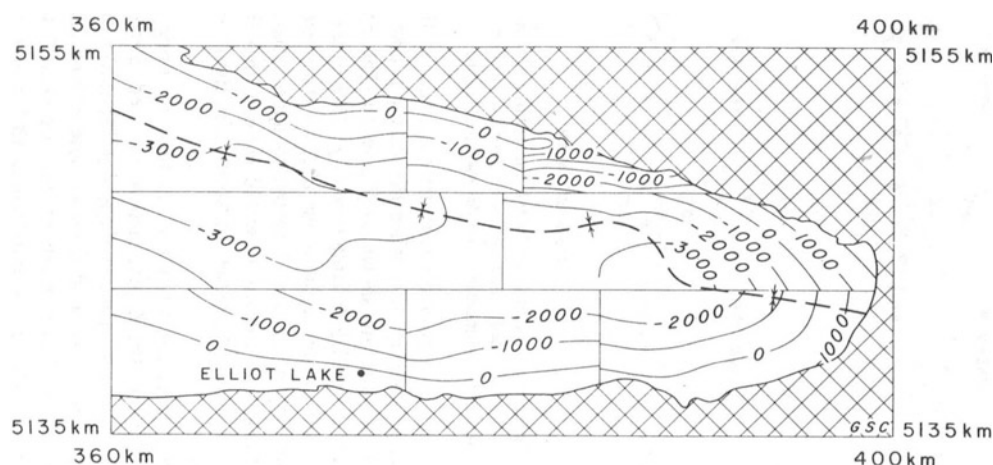
Trend Surface Analysis,

Fig. 7 "Signal" for bottom of Matinenda Formation, Elliott Lake area, Ontario, obtained by contouring kriging values (in feet) computed at points with 1 km-spacing. Squares and triangles represent locations of drill holes with and without significant mineralization, respectively. (Source: Agterberg 1984, Fig. 7)



Trend Surface Analysis,

Fig. 8 Sum of trend (see Fig. 6) and signal (see Fig. 7) for bottom of Matinenda Formation. Note strong resemblance between patterns of Figs. 6 and 8 illustrating that the fluctuations (signal) contoured in Fig. 7 have relatively small magnitudes. (Source: Agterberg 1984, Fig. 9)



The following procedure was followed to obtain a 2D autocorrelation function for residuals from the trend shown in Fig. 6. By using polar coordinates, distances between pairs of boreholes were grouped according to (1) distance (1-km spacing) and (2) direction of connecting line (45° wide segments). This gave domains of different sizes. Autocorrelation coefficients for pairs of values grouped according to this method were assigned to the centers of gravity of their domains. A 2D quadratic exponential (Gaussian) function with superimposed noise component (nugget effect) was fitted by trend surface analysis of logarithmically positive autocorrelation coefficients obtained from the grouped data. Both the input pattern of autocorrelation coefficients and the fitted function are symmetrical with respect to the origin where the observed autocorrelation coefficient is equal to 1 because it includes the noise component. A profile through the origin across the quadratic trend surface fitted to logarithmically transformed autocorrelation function is a parabola with its maximum at the origin. Each parabola becomes a Gaussian curve when the antilog is taken. The maximum of the fitted function fell at 0.325 indicating that only 32½ % of the variance of the residuals is explained by the signal versus 67½ % by the noise. The function has elliptical contours which are elongated in approximately the NW-SE direction. This directional anisotropy would indicate that the actual topography features (channels?) which constitute the signal are elongated in the NW-SE direction.

The signal was extracted from the residuals by using the theoretical autocorrelation function. The kriging method was used to estimate the values of the signal at the intersection points of a grid (UTM grid) with 1-km spacing. Every kriging value was estimated from all observations falling within the ellipse described by the 0.01 contour of the theoretical Gaussian autocorrelation function. This ellipse is approximately 13 km long and about 7 km wide. The estimated kriging values on the 1-km grid defined a relatively smooth pattern which was contoured yielding the pattern of Fig. 7.

Summary and Conclusions

Trend surface analysis is a useful application of the general linear model when the dependent variable of interest shows broad regional changes that can be modeled by using simple polynomial functions of the geographical coordinates. In this chapter a two-dimensional application was given using enstatite content of pyroxene crystals in the Mount Albert Peridotite intrusion in Québec. Most of this peridotite body that primarily consisted of olivine and pyroxene was later serpentinized. Because of significant differences in Mount Alberts topography, the history of this alteration process could be described to some extent by means of three-dimensional trend analysis.

Using other examples, it was shown that the residuals from trend surfaces generally are not stochastically independent but autocorrelated. The methods of time series analysis and geostatistics then can be used to extract meaningful patterns superimposed on the trend surfaces.

Cross-References

- [Autocorrelation](#)
- [Exploratory Data Analysis](#)
- [Geostatistics](#)
- [Kriging](#)
- [Mineral Prospectivity Analysis](#)
- [Pattern Analysis](#)
- [Signal Processing in Geosciences](#)
- [Time Series Analysis](#)

Bibliography

- Agterberg FP (1964) Methods of trend surface analysis. Q Colorado Sch Min 59:111–130

- Agterberg FP (1974) *Geomathematics*. Elsevier, Amsterdam
- Agterberg FP (1984) Trend surface analysis. In: Gaile GL, Willmott CJ (eds) *Spatial statistics and models*. Reidel, Dordrecht, pp 147–171
- Agterberg FP (2014) *Geomathematics: theoretical foundations, applications and future developments*. Springer, Heidelberg
- Cressie NAC (1991) *Statistics for spatial data*. Wiley, New York
- Davis JC (1973) *Statistics and data analysis in geology*. Wiley, New York
- Deschamps F, Godard G, Guillot S, Hattori K (2013) Geochemistry of subduction zone serpentinites: a review. *Lithos* 178:86–127
- Huijbregts C, Matheron G (1971) Universal kriging (an optimal method for estimating and contouring in trend surface analysis). *Can Inst Min Metall Spec vol 12*:159–169
- Krige DG (1966) Two-dimensional weighted moving average trend surfaces for ore valuation. In: *Proceedings of symposium on mathematica, statistics and computers applied to ore valuation*. S Afr Inst Min Metall, Johannesburg, pp 13–38
- Krumbein WC, Graybill FA (1965) *An introduction to statistical models in geology*. McGraw-Hill, New York
- Matter JM, Stute M (2016) Rapid carbon mineralization for permanent disposal of anthropogenic carbon dioxide emissions. <https://doi.org/10.1126/science.a2d8132>
- Robertson JA (1966) The relationship of mineralization to stratigraphy in the Blind River area, Ontario. *Geol Assoc Can Spec Pap* 3:121–136
- Watson GS (1971) Trend surface analysis. *J Int Assoc Math Geol* 3: 215–226

Tukey, John Wilder

Frits Agterberg
Central Canada Division, Geomathematics Section,
Geological Survey of Canada, Ottawa, ON, Canada

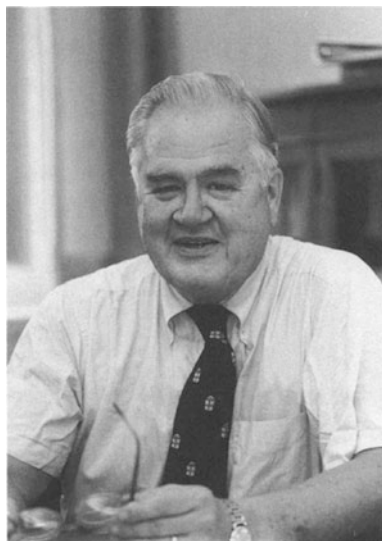


Fig. 1 John Wilder Tukey (1925–2000). Originally published in Anscombe (2003). Reprinted with the permission of the Institute of Mathematical Statistics

Biography

Professor John Wilder Tukey (1915–2000) was born in New Bedford, Massachusetts. As a student at Brown University, he took a course in geology, a subject in which he remained interested during his long and distinguished career. In 1965, he became the founding chairman of the Department of Statistics at Princeton University where he had obtained his PhD in 1939. He also was a director at the AT&T Bell Laboratories. Among his many honors, there was the 1982 Medal of Honor of the Institute of Electrical and Electronics Engineers received for his contributions to spectral analysis of random processes (cf. Blackman and Tukey 1958) and inventing the FFT (Fast Fourier Transform) (cf. Cooley and Tukey 1965). Tukey is probably best known for his book on *“Exploratory Data Analysis”* (Tukey 1977). He added words to the English language: “bit” (in 1947) and “software” (in 1958). At the 53rd Session of the International Statistical Institute (ISI) held in Seoul, Korea (in 2001), John Tukey was declared to be one of the most prominent mathematical statisticians during the second half of the twentieth century.

Tukey has produced over 800 publications. His impact in the field of mathematical geoscience included collaborations with geoscientists. With Bill Krumbein, he worked on multivariate analysis of mineralogical, lithological, and chemical rock composition data (Krumbein and Tukey 1956). Personally, I have had the privilege of working with Tukey on four occasions. In 1966, when I was “petrological statistician” at the Geological of Canada in Ottawa, I participated in an advanced statistical seminar organized by George E.P. Box at the University of Wisconsin. Tukey approached me and subsequently arranged that I received a box with about 2,000 IBM cards to calculate the FFT in one-, two-, or three dimensions on a mainframe computer.

Tukey (1970) was the discussant of a paper I presented at a “Geostatistics” symposium held at the Geological Survey of Kansas. In it, I had divided a map area into three parts. Observations from two parts were used to make predictions in the third, “blind” map area. Tukey pointed out that this form of cross-validation could indeed be used, but a better technique would be to use the jackknife method proposed by Mosteller and Tukey (1968). I also discussed applications of autocorrelation and spectral analysis. According to Tukey, spectra are better. With respect to one example, he pointed out that, although the signal-plus-noise model results for it were within the 95% confidence limits, only the spectrum showed that this approach was not fully satisfactory. It is remarkable that much later, in an application of their universal multifractal method to the same data set, Lovejoy and Schertzer (2007, Fig. 26a) provided a spectrum that perfectly described the lack of fit originally noticed by Tukey.

At the 38th ISI meeting held in Washington, D.C., I presented a paper on mineral potential maps useful to predict the probable occurrences of unknown large copper and zinc deposits. Lithological and geophysical variables quantified for (10 x 10 km) cells were used as explanatory variables in multiple regressions to estimate probabilities of occurrence. Tukey (1972) who was the discussant of this paper pointed out that it might be better to estimate probabilities confined to the [0,1] interval by using logistic analysis. Subsequently, my colleagues and I, in the Geomathematics Section of the Geological Survey of Canada, extensively used logits in similar studies, e.g., in weights of evidence (WofE) modeling. Later, Tukey (1984) commented again on mineral potential maps correctly pointing out that the probabilities of finding new ore deposits were greater on the flanks than on the maxima on these contour maps.

Bibliography

- Anscombe, F. R. (2003): "Quiet Contributor: The Civic Career and Times of John W. Tukey", *Statistical Science*, Aug., 2003, Vol. 18, No. 3 (Aug., 2003), pp. 287-310.
- Blackman RB, Tukey JW (1958) *The measurement of power spectra*. Dover, New York
- Cooley JW, Tukey JW (1965) An algorithm for the machine calculation of complex Fourier series. *Math Comput* 19:297-307
- Krumbein WC, Tukey JW (1956) Multivariate analysis of mineralogic, lithologic, and compositional composition of rock bodies. *J Sediment Res* 26:322-339
- Lovejoy S, Schertzer D (2007) Scaling and multifractal fields in the solid earth and topography. *Nonlinear Process Geophys* 14:465-502
- Mosteller F, Tukey JW (1968) Data analysis including statistics. In: Lindsey LG, Anderson AE (eds) *Handbook of social psychology*, vol 2, 2nd edn. Addison-Wesley, Reading, pp 80-123
- Tukey JW (1970) Some further inputs. In: Merriam DF (ed) *Geostatistics*. Plenum, New York, pp 163-174
- Tukey JW (1972) Discussion of paper by F.P. Agterberg and S.C. Robinson. *Proc 38th Session Intern Statist Inst*, Bull 38:596
- Tukey JW (1977) *Exploratory data analysis*. Addison-Wesley, Reading
- Tukey JW (1984) Comments on "use of spatial analysis in mineral resource evaluation". *J Int Assoc Math Geol* 16:591-594

Turbulence

Frits Agterberg
Central Canada Division, Geomathematics Section,
Geological Survey of Canada, Ottawa, ON, Canada

Definition

Turbulence is the motion of a fluid or gas that is characterized by chaotic changes in pressure as well as in flow or stream velocity. It is created in fluid flow when the kinetic energy

exceeds the damping effect of the fluid's viscosity. Most atmospheric air circulation is affected by turbulence and so are strong oceanic currents and fast-flowing rivers. When river flow is slow, the water moves slowly around obstacles without turbulence but, when flow exceeds a specific limit, turbulence is generated. This critical velocity is characterized by its Reynolds number usually written as R . Sommerfeld (1908) pointed out that the first value of R was established in 1883 by Osborne Reynolds who performed a number of flow visualization experiments in a pipe. A major breakthrough in the understanding of turbulence came with the publication of Kolmogorov (1941)'s similarity principle for R showing that it is the only control parameter for incompressible flow. This principle eventually led to the theory of multifractals anticipated by Mandelbrot (1989) and later firmly established by many authors including Frisch (1995) and Sreenivasan (1991). In respect to the past, the geologic record contains Bouma sequences in sedimentary basins of different ages proving continuous past existence of turbulence in the form of turbidity currents.

Introduction

According to the so-called Orr-Sommerfeld equation:

$$R = \frac{L \cdot V}{\nu}$$

where L is a measure of scale, V represents the fluid's velocity and ν is the kinematic velocity (cf. Frisch 1995). The similarity principle for incompressible flow states that, for a given geometrical shape of the boundaries, the Reynolds number is the only control parameter of the flow.

Applications of the theory of turbulence in geosciences are concerned with movements of air in the atmosphere and with water in oceans, lakes, and rivers. Also, evidence of turbulence resulting in turbidity currents that took place millions of years ago has been found in the bedrock of different geologic ages and has become part of the geologic record. There also exists evidence of turbulence in the past that was connected to the rapid movement of solid matter on land in connection with earthquakes and landslides.

Lee (2013) has characterized turbulence stability through the identification of multifractional Brownian motions (mBm) that constitute a generalization of the fractional Brownian motion concept (fBm). Such models represent mBm as a Gaussian process $W(t)$ with a time-dependent function $H(t)$ replacing the Hurst exponent H . Its variance at time t is assumed to be $\text{Var}[W(t)] = C^2 t^{H(t)}$ with the constant C determining the energy level of the process. It has been shown that atmospheric turbulence is anisotropic requiring different scaling exponents related to the Hurst exponent

(Lovejoy et al. 2010). Lee (2013) separately identified atmospheric turbulence signals for longitudinal, lateral, and vertical velocities as well as for temperature.

Turbidity Currents

Turbulence was involved in turbidity currents that resulted in the rapid deposition of sediments. A simple form of turbidity current occurred during the Quaternary retreat of land-ice that resulted in varves consisting of couplets of thin silt and clay layers. The varves were deposited annually in lakes formed in front of the retreating masses of land-ice. Instantaneous silt formation due to turbidity currents in the summer was followed by slow clay deposition during the remainder of the year as originally shown by Kuenen (1951). Older sedimentary rocks often show so-called Bouma sequences, which are the fossilized remains of turbidity currents in basins that existed at different times during the history of the Earth.

Emplacement of massive turbidites linked to the extinction of turbulence in turbidity currents often resulted in complete or partial manifestations of Bouma sequences (Cantero et al. 2011). A complete Bouma (1962) sequence consists of five distinct layers of sedimentary rocks with specific lithologies. In a fully preserved Bouma sequence, these are, from bottom to top: (1) massive to coarse-grained sandstone, often with pebbles; (2) medium- to fine-grained sandstone, frequently with sole markings such as flute casts indicating the direction of turbidity flow; (3) fine-grained sandstone with ripple marks; (4) parallel-laminated siltstone; and (5) mudstone, possible with trace fossils. Of course, there was no preservation of any water above, within and below any Bouma sequence when it was created.

Various quantitative models have been developed to explain Bouma sequences that can be of any geologic age and which commonly show lateral variations of composition in the flow direction. One such quantitative model is Parker (2008)'s "Turbidity Current with a Roof" (TCR) model of flow driven by suspended sediment within a layer of thickness H , which is confined between two parallel plates with standing water below and above it. Suppose that ρ represents water density, ρ_s is sediment density, s represents channel slope, and ν is the kinematic viscosity of water. According to Cantero et al. (2011), the turbulence for the TCR model then is controlled by the following three dimensionless parameters: the shear Richardson number R_1 , shear Reynolds number R_2 and fall velocity $\hat{V}$. The first two of these parameters satisfy:

$$R_1 = \frac{RgCH}{2u^2}; R_2 = \frac{uH}{2\nu}$$

where $R = \frac{\rho_s}{\rho} - 1$ (≈ 1.65 for quartz), g is the gravitational acceleration, C is the layer-averaged volume suspended sediment concentration, and u is nominal shear velocity. The dimensionless fall velocity is

$$\hat{V} = \frac{V_s}{u}$$

where V_s is the sediment fall velocity. The preceding equations provide only a partial explanation of this version of the TCR model. A complete explanation with graphical illustrations is provided in Cantero et al. (2011).

Summary and Conclusions

Turbulence occurs when fluid or air movement exceeds a critical value characterized by a Reynolds number R . It occurs in the atmosphere, rivers, and oceans. Kolmogorov (1941) established the similarity principle for R stating that it is the only control parameter for incompressible flow. This concept has resulted in the theory of multifractals. Various theoretical models continue to be developed for turbulence with parameters that change in space and time. The geologic record contains Bouma sequences in sedimentary basins of different ages proving the past existence of turbulence in the form of turbidity currents.

Cross-References

- Hurst Exponent
- Multifractals

Bibliography

- Bouma A (1962) Sedimentology of some flysch deposits: a graphic approach to facies interpretation. Elsevier
- Cantero MI, Cantelli A, Pirmez C, Balachandar S, Mohrig D, Hickson TA, Yeh T, Naruse H, Parker G (2011) Emplacement of massive turbidites linked to extinction of turbulence in turbidity currents. *Nat Geosci Lett* 5:42–45
- Frisch U (1995) Turbulence: the legacy of A.N. Kolmogorov. Cambridge University Press
- Kolmogorov AN (1941) Local structure of turbulence in incompressible viscous fluid for very large Reynolds' numbers. *C R Acad Sci URSS (N S)* 30:301–305
- Kuenen PH (1951) Mechanics of varve formation and the action of Naruke turbidity currents. *Geol Förh* 73:69–84
- Lee KC (2013) Characterization of turbulence stability through the identification of multifractional Brownian motions. *Nonlin Process Geophys* 20:97–2013
- Lovejoy S, Tuck A, Schertzer D (2010) Horizontal cascade structure of atmospheric fields determined from aircraft data. *J Geophys Res* 115: D13105. <https://doi.org/10.1029/2009JD013353>

- Mandelbrot B (1989) Multifractal measures, especially for the geophysicist. *Pure Appl Geophys* 131:5–42
- Sommerfeld A (1908) Ein Beitrag zur hydrodynamischen Erklärung der turbulenten Flüssigkeitsbewegung. 4th International Congr. Math Rome 3:116–124
- Sreenivasan (1991) Fractals and multifractals in fluid turbulence. *Annu Rev Fluid Mech* 23:539–600

Turcotte, Donald L.

Lawrence Cathles¹ and John Rundle²

¹Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, NY, USA

²Departments of Physics and Earth and Planetary Sciences, 2131 Earth and Physical Sciences, University of California Davis, Davis, CA, USA



Fig. 1 Donald L. Turcotte, courtesy of Prof. John Rundle

Biography

Donald L. Turcotte is recognized globally for his quantification of nearly all aspects of physical geology.

Professor Turcotte received a Ph.D. in aeronautics and physics from the California Institute of Technology in 1958. After a year at the Naval Postgraduate School in Monterey, he joined the Graduate School of Aeronautical Engineering at Cornell where he addressed seeded combustion, magnetohydrodynamic phenomenon in turbulent boundary layers, and shock waves. He became Full Professor and published his first two books: *Space Propulsion and Statistical Thermodynamics* and *Statistical Thermodynamics*.

Professor Turcotte became an earth scientist in 1965 when he went on sabbatical to Oxford University and met Ron Oxburgh. His quantitative abilities and physical intuition combined with Ron's knowledge of geology to produce

24 papers that quantified the emerging paradigm of plate tectonics, establishing the physical basis of processes in the Earth's thermal boundary layer, ridge melting, subduction zone volcanism, and intraplate tectonics and magmatism. In 1973, he left the School of Aeronautical Engineering and joined the Cornell Department of Geological Sciences where he became an expert on planetary remote sensing, thermal subsidence in sedimentary basins, two-phase hydrothermal porous media convection, lithosphere flexure, cyclic sedimentation, stick-slip earthquakes, and the lithospheres and mantles of the other planets. In 1982, he and Jerry Schubert published *Geodynamics*, which remains the primary reference book quantifying physical earth processes.

In 1985, Bob Smalley introduced Professor Turcotte to fractals. This became his next fascination, perhaps because fractals and chaos apply to almost every Earth process, explaining, at last, why geologists must always place their rock hammers in their photographs to establish scale. In 1992, he published *Fractals and Chaos in Geology and Geophysics*, a primary reference in this new field.

Professor Turcotte is renowned for his productivity, ability to inspire and collaborate, his administrative skills, and pleasant equanimity. He has published over 300 papers in refereed journals and authored or coauthored 4 books, most recently a comprehensive 2001 treatise on mantle convection with Jerry Schubert and Peter Olson. As a member of innumerable academy committees, editorial boards, working groups, past president of the Tectonics Section of the AGU, chair of the Geophysics Section of the National Academy of Sciences, and chair of the Department of Geology at Cornell for 9 years, he has interacted with nearly everyone in the geophysics profession, all of whom appreciate his remarkable ability to routinely turn mountains into molehills.

Professor Turcotte has been recognized by election to the National Academy of Sciences, receipt of the Arthur L. Day medal of the Geological Society of America, the Charles A. Whitten Medal, and the William Bowie Medal of the American Geophysical Union. The Bowie Medal is the AGU's most prestigious award and honors "outstanding contributions to fundamental geophysics and unselfish collaboration in research," which fits Don perfectly.

Bibliography

- Malamud BD, Turcotte DL (1999) Self-affine time series: I. generation and analyses. In: *Advances in geophysics*, vol 40. Elsevier, pp 1–90 <https://www.sciencedirect.com/science/article/abs/pii/S0065268708602939>
- Turcotte DL (1997) *Fractals and Chaos in geology and geophysics*, 2nd edn. Cambridge University Press, Cambridge
- Turcotte D, Schubert G (2014) *Geodynamics*, 3rd edn. Cambridge University Press, Cambridge

Unbalanced Analysis of Variance

R. Webster

Rothamsted Research, Harpenden, UK

Definition

The analysis of variance is a statistical technique, or set of techniques, for separating variance in data into portions that can be attributed to two or more sources of variation. The form of an analysis must fit the experimental or survey design by which the data were obtained. Balanced designs are ones in which all sources at any particular level of classification are observed equally. In the geosciences balance is exceptional; it is rarely possible or even undesirable to measure the same number units in all classes of rock or soil in a region. Data from surveys are almost always unbalanced, either because of circumstances or by deliberate design. The corresponding analyses of such data are therefore also unbalanced. Unbalanced nested designs are especially important in regional geochemistry.

Introduction

In the 1920s RA Fisher set out the principles of experimental design whereby investigators could distinguish effects of their experimental treatments from natural variation in the material being investigated. These comprised replication, to provide information on the variation in responses to the treatments, and randomization to avoid bias that could arise if effects of treatments were confounded with other sources of variation. Fisher devised the analysis of variance (ANOVA) to separate the sources of variation in data from such experiments, to estimate quantitatively the effects of the treatments and to provide inferential tests to judge whether the observed differences could properly be attributed to the imposed treat-

ments rather than their arising by chance. Fisher's innovation was a revolution in agronomy, and it was quickly followed in biological and medical research, and it has been subsequently embraced by geoscientists for surveys in which the same principles apply.

Single-Stage Analysis of Variance

Balanced Designs

Fisher was initially concerned with fertilizer experiments in the field. Treatments were assigned to plots at random, and all treatments would be replicated equally; in other words, the design was *balanced*. For generality we can designate the treatments as "classes." In the simplest cases there is only one level of classification, and the design can be summarized in a single-stage ANOVA table, such as Table 1.

The quantity σ_B^2 , the variance between classes, arises from the differences among the class means. It is a fixed effect determined by the investigator. The quantity σ_W^2 is the within-classes variance, a random quantity estimated by W . It determines the confidence one can have in the class means and any differences between them. The variance of a class mean, say $\bar{z}_{\text{class}}$, would be W/n and its standard error $\sqrt{W/n}$. Agronomists are typically concerned with differences between treatments and compute the standard error of any difference, say $\bar{z}_1 - \bar{z}_2$, as

$$\text{SED} = \sqrt{2W/n}. \quad (1)$$

Unbalanced Design

The same principles apply to surveys. Surveyors, however, rarely have complete control over replication and often do not want it. A geoscientist might sample the soil of a region to estimate the concentration of some element or pollutant in the soil. Different concentrations in soil derived from different classes of parent rock or under different forms of land management would express themselves in the between-classes

Unbalanced Analysis of Variance, Table 1 Analysis of variance for a single-stage balanced completely randomized design with k classes, each replicated n times

Source	Degrees of freedom	Mean square	Parameters estimated	F ratio
Between classes	$k - 1$	B	$\sigma_W^2 + n\sigma_B^2$	B/W
Within classes (residual)	$k(n - 1)$	W	σ_W^2	
Total	$kn - 1$	T		

Unbalanced Analysis of Variance, Table 2 Analysis of variance for a single-stage unbalanced completely randomized design with k classes, replicated $n_1, n_2, \dots, n_k$ times to give N observations in total

Source	Degrees of freedom	Mean square	Parameters estimated	F ratio
Between classes	$k - 1$	B	$\sigma_W^2 + n_0\sigma_B^2$	B/W
Within classes (residual)	$N - k$	W	σ_W^2	
Total	$N - 1$	T		

variance. It is highly unlikely that all classes of rock or land management would be represented equally; the design would be *unbalanced*. Nevertheless, with only one level of classification the ANOVA is still fairly straightforward. Table 1 is replaced by a somewhat more elaborate Table 2.

The contribution of the between-classes variance, σ_B^2 , to the mean square B is somewhat modified. The quantity n of Table 1 is replaced by n_0 given by

$$n_0 = \frac{1}{k-1} \left(N - \frac{\sum_{i=1}^k n_i^2}{N} \right). \quad (2)$$

In an investigation intended to determine the means of each class the differences among the classes comprise a fixed effect, as in the experimental situation. More often in surveys the classes are not predetermined, and the differences among classes arise by chance; they are random effects. In these circumstances σ_B^2 is of interest in its own right and is known as a component of variance.

Further, each class mean $\bar{z}_i$ has its own unique variance estimated by W/n_i and standard error $\sqrt{W/n_i}$ that depend on the number of observations contributing to the mean. The standard error of a difference between two classes, say class 1 and class 2, is

$$\text{SED} = \sqrt{W \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}. \quad (3)$$

Geoscientists are less likely to be interested in the differences between class means than agronomists would be. Nevertheless, Eq. (3) provides the SED.

The simplest of designs represented by Table 1 could be elaborated by blocking, by factorial combinations of treatments, by splitting the plots, and by nesting the treatments one within another; see, for example, Cochran and Cox (1957). Balance would be retained provided that the replication was equal at any level in the design. Balance could be upset if lack of experimental resources or partial failure in an experiment meant that replications were unequal; the analysis would then have to take the lack of balance into account. The most general means for this is by residual maximum likelihood (REML) proposed by Patterson and Thompson (1971), and facilities for it are now available in statistical packages.

Multi-stage Analysis of Variance

The classes of land encountered in surveys are often nested one within another. For example, an agricultural region may comprise administrative districts (stage 1), each divided into farms (stage 2), which in turn are divided into fields (stage 3), and the soil of each of field is sampled at several places (stage 4) to provide the observations. In each case the class at the lower stage is contained completely within the one immediately above it, and each sampling point is contained in one and only one class at each and every stage. The system is a strict hierarchy, a nested structure, and a single observation embodies variation contributed by all of the stages, including an unresolved variance within the classes at the lowest level. The system can be modeled as

$$Z_{ijkl} = \mu + A_i + B_{ij} + C_{ijk} + \varepsilon_{ijkl}. \quad (4)$$

Here Z_{ijkl} is the value of the l th observation of the k th class at stage 3, in the j th class at stage 2, and in the i th class at stage 1. The general mean is μ ; A_i is the difference between μ and the mean of class i in the first stage; B_{ij} is the difference between the mean of the j th sub-class in class i and the mean of class i ; and so on. The final quantity ε_{ijkl} represents the deviation of the observed value from its class mean in stage 3. More stages can be added as desired.

Balanced Designs

A hierarchical design is balanced if all classes at each particular stage are represented equally. The ANOVA table would then appear as Table 3, for which analysis itself is straightforward.

The variance of a mean at any one stage in the design is obtained as the mean square within that stage divided by the number of observations contributing to that mean. If we denote the estimates of the variance components by $s^2 = \hat{\sigma}^2$ then the estimation variances are

Unbalanced Analysis of Variance, Table 3 Analysis of variance for a three-stage balanced completely randomized design with n_1 classes at stage 1, n_2 in each class at stage 2, n_3 in each class at stage 3, and n_4

observations in each class at stage 3 to give $N = n_1 n_2 n_3 n_4$ observations in total and variance parameters

Source	Degrees of freedom	Mean square
Stage 1	$n_1 - 1$	$\frac{1}{n_1 - 1} \sum_{i=1}^{n_1} n_4 n_3 n_2 (\bar{z}_i - \bar{z})^2$
Stage 2	$n_1(n_2 - 1)$	$\frac{1}{n_1(n_2 - 1)} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} n_4 n_3 (\bar{z}_{ij} - \bar{z}_i)^2$
Stage 3	$n_1 n_2(n_3 - 1)$	$\frac{1}{n_1 n_2(n_3 - 1)} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} n_4 (\bar{z}_{ijk} - \bar{z}_{ij})^2$
Residual	$n_1 n_2 n_3(n_4 - 1)$	$\frac{1}{n_1 n_2 n_3(n_4 - 1)} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \sum_{l=1}^{n_4} (\bar{z}_{ijkl} - \bar{z}_{ijk})^2$
Total	$N - 1$	$\frac{1}{N - 1} \sum_{l=1}^N (z_l - \bar{z})^2$
Parameters estimated by mean squares		
Stage 1	$n_1 - 1$	$n_4 n_3 n_2 \sigma_1^2 + n_3 n_2 \sigma_2^2 + n_2 \sigma_3^2 + \sigma_4^2$
Stage 2	$n_1(n_2 - 1)$	$n_3 n_2 \sigma_2^2 + n_2 \sigma_3^2 + \sigma_4^2$
Stage 3	$n_1 n_2(n_3 - 1)$	$n_2 \sigma_3^2 + \sigma_4^2$
Residual (stage 4)	$n_1 n_2 n_3(n_4 - 1)$	σ_4^2
Total	$N - 1$	

The quantities $\bar{z}_i$, $\bar{z}_{ij}$, $\bar{z}_{ijk}$, and $\bar{z}_{ijkl}$ are the means of observations at the four levels of the design

$$\begin{aligned} s^2(\bar{z}_3) &= s_4^2/n_4 \\ s^2(\bar{z}_2) &= s_4^2/(n_4 n_3) + s_3^2/n_3 \\ \text{and } s^2(\bar{z}_1) &= s_4^2/(n_4 n_3 n_2) + s_3^2/(n_3 n_2) + s_2^2/n_1. \end{aligned}$$

If the classes at the several stages can be regarded as random then the corresponding variances σ_1^2 , σ_2^2 , σ_3^2 , and σ_4^2 are components of variance like σ_B^2 above.

Unbalanced Designs: Multi-stage Survey and the Variogram

One particular form of multi-stage design that is finding increasing application in the geosciences and environmental science more widely is one in which the classes are specific distances between sampling points in surveys. Pairs of points, or perhaps sets of three points, are spaced at some fixed distance but in a random orientation from one another; these pairs or triplets are grouped with other pairs or triplets some larger fixed distance apart but again in a random direction; those in turn are grouped with yet other groups a further fixed distance apart, and so on. The structure is spatially nested. The distances typically increase in at least 2.5-fold progression so that the nests at any one stage never overlap in space and are reasonably independent. The design enables one to estimate the contributions of differences at these distances to the total variance from an ANOVA as set out in Table 3. It reveals which spatial scales contribute to the variance and how to plan future surveys efficiently. The procedure was first proposed with that aim by Youden and Mehlich (1937), but it lay dormant until the 1970s.

Miesch (1975) described it in a geochemical setting and linked it to the variogram of geostatistics. He pointed out that if the components of variance were accumulated, starting with the smallest spacing, they formed an estimate of the variogram. For m stages of nesting at distances d_m , d_{m-1} ,

..., d_1 , where d_m is the shortest distance at the m th stage and m_1 is the distance at the first stage, the equivalences are

$$\begin{aligned} \hat{\sigma}_m^2 &= \hat{\gamma}(d_m), \\ \hat{\sigma}_{m-1}^2 + \hat{\sigma}_m^2 &= \hat{\gamma}(d_{m-1}), \\ \hat{\sigma}_{m-2}^2 + \hat{\sigma}_{m-1}^2 + \hat{\sigma}_m^2 &= \hat{\gamma}(d_{m-2}), \end{aligned} \quad (5)$$

and so on.

The scheme is elegant, but to maintain balance the number of sampling points would have to double or triple at each additional stage in the design; the sampling would increase exponentially as the number of stages, that is, distances, increases; it would soon become unaffordable. To make it affordable balance must be sacrificed; only some of the nests in the lower stages contain replicates. The ANOVA is more complex than in the balanced design. The sums of squares are formed in the same way by the method of moments, but the degrees of freedom vary according to the precise structure of the replication in each stage. Further, and more important for a first estimate of a variogram, the coefficients by which the components of variance are multiplied vary from stage to stage. Table 4 sets out the ANOVA for three stages of classification.

The coefficients u can be calculated by the formulae originally provided by Gower (1962):

$$q_{i,j} = \sum_{k=1}^{C_i} \sum_{p=1}^{C_{jk}} \frac{n_{pj}^2}{n_{ik}} \quad \text{and} \quad u_{i,j} = \frac{1}{D_i} (q_{i,j} - q_{i-1,j}). \quad (6)$$

The quantities in the formulae are as follows:

$u_{i,j}$ is the coefficient at stage i for the j th component, C_i is the number of classes at stage i ,

C_{jk}^i is the number of sub-classes in stage j within the k th class in stage i ,
 n_{pj} is the number of sampling points in the p th sub-class at stage j (within class k at stage i) with $i \leq j$, and
 D_i is the number of degrees of freedom at stage i .

Illustrative Example: Lead in the Soil of the Jura

A geochemical survey of potentially toxic metals in the topsoil of the Swiss Jura by Atteia et al. (1994) illustrates the technique. Its aim was to discover the spatial scale(s) of variation and where large concentrations of metals posed a risk to health. A region of 14.5 km² was sampled at 214 nodes of a square grid at 250-m intervals. Thirty eight of the nodes were chosen at random as starting points for nested sampling with four additional stages, only the last of which was replicated two-fold. The design was highly unbalanced. Table 5 sets out the ANOVA for lead (Pb), for which the mean concentration of 57 mg kg⁻¹ exceeded the statutory threshold of 50 mg kg⁻¹ and concentrations were much more in excess locally.

Note that there is a substantial contribution at 15 m and relatively little between 15 and 250 m. The estimated component at 40 m is very small and may be regarded as an estimate of zero. The accumulated components form a rough variogram. Atteia et al. went on to plot the accumulated components on a variogram computed for lag distance increasing in equal increments.

Unbalanced Analysis of Variance, Table 4 Analysis of variance for a three-stage unbalanced completely randomized design

Source	Degrees of freedom	Parameters estimated by mean squares
Stage 1	$D_1 = f_1 - 1$	$u_{1,1}\sigma_1^2 + u_{1,2}\sigma_2^2 + u_{1,3}\sigma_3^2 + \sigma_4^2$
Stage 2	$D_2 = f_2 - f_1$	$u_{2,2}\sigma_2^2 + u_{2,3}\sigma_3^2 + \sigma_4^2$
Stage 3	$D_3 = f_3 - f_2$	$u_{3,3}\sigma_3^2 + \sigma_4^2$
Residual (stage 4)	$D_4 = N - f_3$	σ_4^2
Total	$N - 1$	

Residual Maximum Likelihood Estimation

The ANOVAS as set out in Tables 4 and 5 for unbalanced sampling are sound. The estimates of the variance components are unbiased, though statisticians now prefer residual maximum likelihood (REML) for estimating them. Webster et al. (2006) and Webster and Lark (2013) provide the detail. For balanced designs the results of ANOVA and REML are the same; for unbalanced designs REML estimates of the components of variance differ somewhat from those computed with Gower's coefficients. The right-hand column of Table 5 lists the components of variance as estimated by REML for comparison.

Many other experimental and survey designs may be unbalanced, either by intention for economy, or by chance, or by unforeseen circumstances such as loss of data, failure of instruments, or no access to land. The REML technique, with its great flexibility, will usually enable investigators to cope with the lack of balance whereas ANOVA might not.

Summary

The analysis of variance (ANOVA) devised originally for estimating the effects of imposed treatments in field experiments is readily adapted to surveys provided sampling has been suitably randomized. Whereas experimental treatments can be replicated equally so that designs are balanced, equal replication of classes in survey is exceptional and frequently undesirable. Survey designs are usually unbalanced, and the ANOVAS, which must fit the designs, are unbalanced. Geochemists have a particular interest in nested sampling designs for discovering the spatial scales of ores, metals, and pollutants. The designs are almost necessarily unbalanced for economy. The ANOVAS for such designs are tabulated, and the formulae for computing the components of variance are given. The analysis for a highly unbalanced nested sample survey is illustrated with data on lead (Pb) in the soil from a geochemical survey in the Swiss Jura. The components of variance are estimated by a hierarchical analysis of variance and compared with ones estimated by the now favored residual maximum likelihood.

Unbalanced Analysis of Variance, Table 5 Unbalanced analysis of variance for the concentration of lead in the soil of the Swiss Jura

Source (Distance)	Degrees of freedom ^a	Mean square	Variance component	Cumulative variance	REML estimate
Stage 1 (>250 m)	211	2047	590.2	1532.2	402.7
Stage 2 (100 m)	38	1082	154.0	942.0	282.4
Stage 3 (40 m)	37	856	3.6	788.0	7.0
Stage 4 (15 m)	38	937	458.8	784.4	483.2
Stage 5 (6 m)	38	326	325.6	325.6	277.2
Total	362	1528			

The measurements were in mg kg⁻¹

^aThere were three missing values; the total degrees of freedom are therefore 362 for 363 data from the 366 sampling points

Bibliography

- Atteia O, Dubois J-P, Webster R (1994) Geostatistical analysis of soil contamination in the Swiss Jura. *Environ Pollut* 86:315–327
- Cochran WG, Cox GM (1957) *Experimental designs*. Wiley, New York
- Gower JC (1962) Variance component estimation for unbalanced hierarchical classification. *Biometrics* 18:168–182
- Miesch AT (1975) Variograms and variance components in geochemistry and ore evaluation. *Geol Soc Am Mem* 142:333–340
- Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika* 58:545–554
- Webster R, Lark RM (2013) *Field sampling for environmental science and management*. Routledge, London
- Webster R, Welham SJ, Potts JM, Oliver MA (2006) Estimating the spatial scales of regionalized variables by nested sampling, hierarchical analysis of variance and residual maximum likelihood. *Comput Geosci* 32:1320–1333
- Youden WJ, Mehlich A (1937) Selection of efficient methods of soil sampling. *Contrib Boyce Thompson Inst Plant Res* 9:59–70

Uncertainty Quantification

Behnam Sadeghi^{1,2}, Eric Grunsky³ and Vera Pawlowsky-Glahn⁴

¹EarthByte Group, School of Geosciences, University of Sydney, Sydney, Australia

²Earth and Sustainability Research Centre, University of New South Wales, Sydney, Australia

³Department of Earth and Environmental Sciences, University of Waterloo, Waterloo, ON, Canada

⁴Department of Computer Science, Applied Mathematics and Statistics, Universitat de Girona, Girona, Spain

Definition

An issue in all geochemical anomaly classification methods is uncertainty in the identification of different populations and allocation of samples to those populations, including the critical category of geochemical anomalies or patterns that are associated with the effects of mineralization. This is a major challenge where the effects of mineralization are subtle.

Introduction

There are various possible sources of such uncertainty, including (i) gaps in coverage of geochemical sampling within a study area; (ii) errors in geochemical data analysis, spatial measurement, interpolation; (iii) misunderstanding of geological and geochemical processes; (iv) fuzziness or vagueness of the threshold between geochemical background and geochemical anomalies; and (v) errors associated with instrument measurement and the associated calibration, in

particular errors associated with accuracy and errors associated with precision (Costa and Koppe 1999; Lima et al. 2003; Walker et al. 2003; Bárdossy and Fodor 2004; McCuaig et al. 2007, 2009, 2010; Kreuzer et al. 2008; Kiureghian and Ditlevsen 2009; Singer 2010; Singer and Menzie 2010; Caers 2011; Zuo et al. 2015). However, there are further potential biases induced by data closure effects in compositional data, which may be dealt with by the use of ratios for which a number of methods exist on the basis of Aitchison geometry (Filzmoser et al. 2009; Carranza 2011; Pawlowsky-Glahn and Buccianti 2011; Gallo and Buccianti 2013; Buccianti and Grunsky 2014; Sagar et al. 2018; Zuzolo et al. 2018; Pospiech et al. 2020). The use of appropriate log-ratio transformations, including clr and ilr – see the ► “Compositional Data” entry – is adequate to represent the data in real coordinates, necessary for a proper modeling and outlier detection.

Such issues generate two main types of uncertainty in geochemical anomaly recognition and mapping that are (i) stochastic uncertainty related to the quality and variability of geochemical data, which affects the results of methods and (ii) systematic uncertainty associated with assumptions and procedures for collection, analysis, and modeling of geochemical data (Porwal et al. 2003; Zuo et al. 2015). Due to analytical and interpolation errors in unsampled or under-sampled areas, there will be both stochastic uncertainty and procedural systematic uncertainty in thresholds achieved from geochemical anomaly classification models. Mathematical uncertainties inherent within the different threshold-determining methods defined can easily be quantified using existing statistical measures.

Geospatial interpolation by any method is one of the main sources of uncertainty in continuous field geochemical models (Costa and Koppe 1999). However, errors in the characterization of significant geochemical anomalies are not only due to interpolation errors in sampled/unsampled areas. Error propagation analysis should be considered and can be assessed using methods such as Monte Carlo simulation.

Error propagation from data collection to data analysis results in thresholds to populations spanning a range of values with the relative magnitude of such ranges dependent on the magnitude of sampling, analytical and interpolation errors (Stanley 2003; Stanley and Lawie 2007; Stanley et al. 2010). In effect, every estimate of a geochemical value in a continuous field model represents a range of values due to statistical and interpolation errors. However, existing methods of geochemical anomaly recognition do not consider error propagation and stochastic and systematic uncertainty quantification in the analysis. Additionally, existing methods assume in general errors to be additive, what is clearly not appropriate with compositional data. This, in turn, results in uncertainty whether observed geochemical anomalous samples or populations are statistically significant.

Uncertainty in Geochemical Anomaly Classification Models

A variety of geostatistical and other mathematical methods have been applied to classify or cluster regolith geochemical data. A common objective is to define clusters that can be spatially related to mineralization, based on univariate or multivariate characteristics. Given the range of factors that can affect regolith geochemistry, including variations in parent lithology, regolith processes, and climatic effects, all such methods are subject to uncertainty in sample classification and pattern recognition, and identification of samples which geochemistry has (or has not) been affected by mineralization (i.e., a geochemical anomaly). This affects stochastic geochemical classification methods including fractal-based models.

Uncertainties associated with geochemical data collection and data analysis can result in two types of errors (Koch and Link, 1970) in relation to the null hypothesis connecting detection or lack of detection of a geochemical anomaly with the associated presence or absence of mineralization. As with all such statistical tests, a Type I error or false positive occurs when the null hypothesis is rejected when the proposition is true, and a Type II error or false negative occurs when the null hypothesis is false but is accepted. From geochemical exploration point of view (Table 1), a Type I error occurs when an ore deposit exists in an area, but no mineralization is detected on the classified map; and a Type II error occurs in an area where there is no deposit but has been classified as an anomalous area. Type I errors preclude mineral discovery, whereas Type II errors will entail exploration expenditure that does not result in mineral discovery. Both types of errors have economic consequences (Rose et al. 1979; Table 1), hence there is a need to be able to quantify the overall accuracy (OA) of the geochemical population characterization as represented by the OA matrix contingency table of Carranza (2011) set out in Table 2.

This ideal approach is, however, flawed as neither b or d can be determined due to the inability in many cases to determine whether undetected mineralization is actually

present in an area where the geology would permit the presence of such mineralization (Yilmaz et al. 2017). It is only possible to experimentally gauge the OA in terms of the geochemical response to known mineralization, hence the typical design of most geochemical orientation surveys that focus on patterns associated with known mineralization and commonly provide insufficient assessment of pattern variations away from the direct influence of mineralization. A partial overall accuracy (POA) contingency table is therefore defined as:

$$POA = a/(a + c) \quad (1)$$

In geochemical exploration projects or orientation studies, the number of samples collected is always limited due to budgetary and other constraints. Therefore, in order to define and predict the spatial mineralization patterns using the available samples in sampled and unsampled areas, interpolation of the available data values is required to assign estimated values to unsampled areas (Chilès and Delfiner 2012; Sadeghi et al. 2015).

Interpolation using any single method is one of the main sources of uncertainty in continuous field geochemical models as interpolation estimates are based only on data within the search window (Costa and Koppe, 1999). In order to evaluate the effect of interpolation errors in detecting geochemically anomalous populations, error propagation should be analyzed (Bedford and Cooke 2001; Oberkampff et al. 2002, 2004; Stanley 2003; Helton et al. 2004; Stanley and Lawie 2007; Verly et al. 2008; Stanley et al. 2010; Mert et al. 2016; Sagar et al. 2018).

Monte Carlo Simulation (MCSIM) to Quantify Systematic Uncertainty

One significant method for analyzing the propagation of error in models and evaluating their stability is the Monte Carlo simulation (MCSIM) (Taylor 1982; Heuvelink et al. 1989). The MCSIM is a computational algorithm that generates random numbers or realizations given a specific density function (Pyrcz and Deutsch 2014; Pakyuz-Charrier et al. 2018; Athens and Caers 2019; Madani and Sadeghi 2019), derived from various parameters controlling that probability density function or the cumulative equivalent (the PDF or CDF) as shown in Fig. 1 (Deutsch and Journel 1998; Scheidt et al. 2018). As the MCSIM is based on the PDF and CDF evaluations, it can be applied to high-dimensional and nonlinear spatial modeling approaches to determine the probability of the target mineralization values occurrence based on the probability distributions relating to a given model (Athens and Caers 2019).

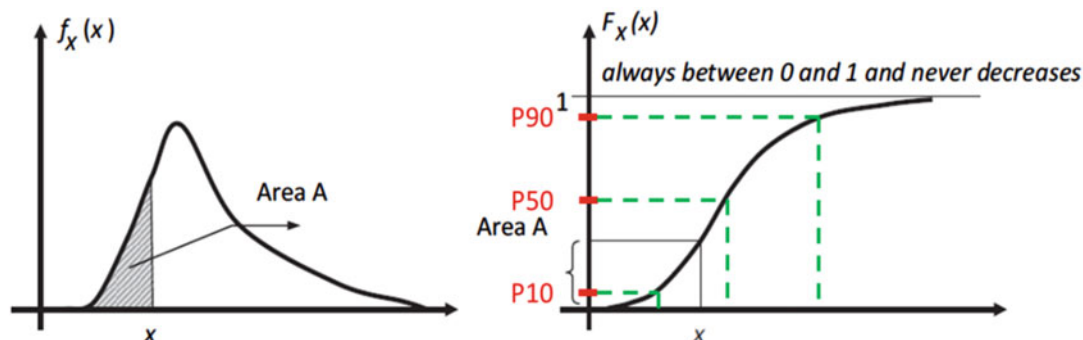
Uncertainty Quantification, Table 1 Classification of samples versus real condition

		Reality	
		Ore present	Ore not present
Assigned class	Anomalous	Correct decision	Type II error – wasted exploration expenditure
	Background	Type I error – mineralization not detected	Correct decision

Adapted from Rose et al. (1979)

Uncertainty Quantification, Table 2 Overall accuracy (OA) matrix to compare performance of a binary model geochemical anomaly or signal recognition with binary ground truth (Sadeghi 2020)

		Geological ground truth	
		Mineralization present and affecting sample geochemistry	No mineralization present
Classification of sample (or pixel)	Anomalous	True anomaly = a	False anomaly = b
	Background	False background = c	True background = d
		Type I error =	$c/(a + b + c + d)$
		Type II error =	$b/(a + b + c + d)$
		Overall accuracy (OA) =	$(a + d)/(a + b + c + d)$

**Uncertainty Quantification, Fig. 1** Schematic PDF and CDF plots and associated quantiles (in this case the P10, P50, and P90 values). (Sadeghi 2020; modified from Scheidt et al. 2018)

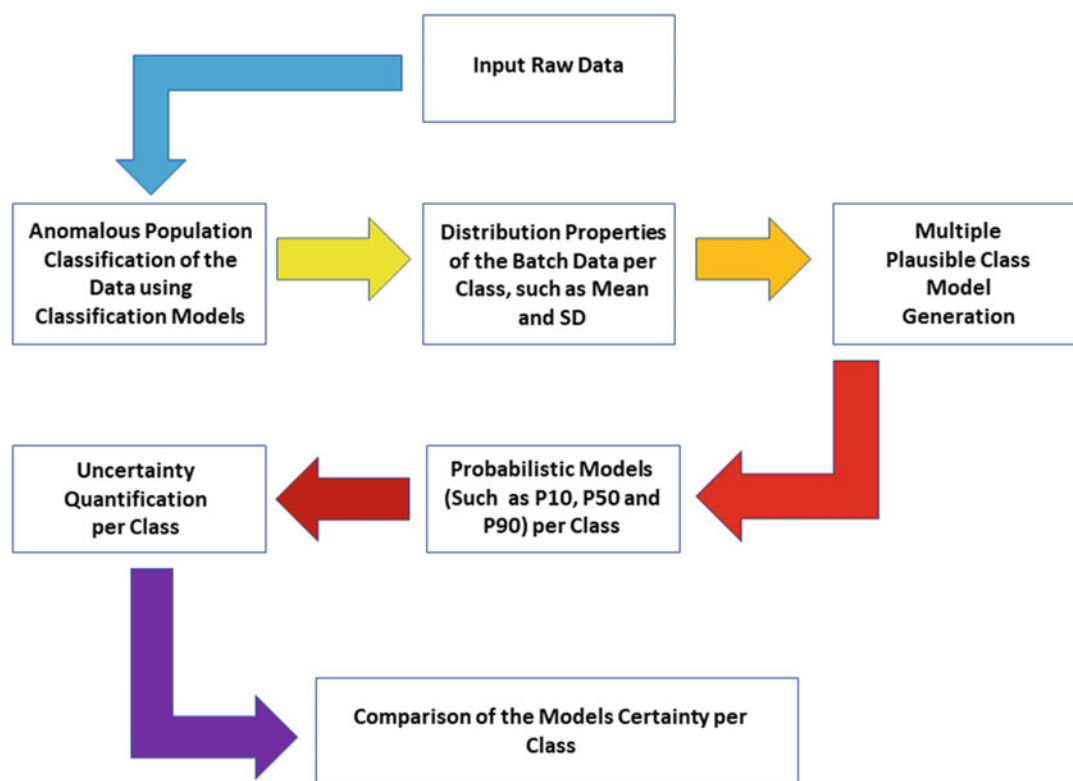
Under the MCSIM approach, the P50 (median) value (the average 50th percentile of the multiple simulated distributions) represents a neutral probability in decision-making (Scheidt et al. 2018), and is defined as the expected “return.” The uncertainty is calculated, in this approach, as $1/(P90-P10)$ for which P10 (lower decile) and P90 (upper decile) are the average tenth and 90th percentiles of the multiple simulated values (Caers 2011; Scheidt et al. 2018), associated with each element.

The implications of the MCSIM modeling mainly relate to sampling density (though may also relate to sample data quality if that can be disentangled from natural geochemical variability). In the recognition of geochemical processes, the sampling density reflects processes that can be adequately represented or processes that are under-represented. Under-represented processes can appear as noise and is apparent when multivariate methods such as principal components are applied to the data (Grunsky and Kjarsgaard 2016). Typically, under-represented processes occur in such methods that are defined by the smaller eigenvalues. Where the quantified return is low or negative, and the quantified uncertainty is high, especially higher than its related return, a decision may be made that additional sampling is required to achieve the minimum required spatial continuity in the data or the stability of the fractal models (Fig. 2). If it goes uncertain, the final models and related classes would not be certain enough, which affects the entire certainty of the models generated in the whole study area.

Spatial Uncertainty

The systematic uncertainty of the thresholds or classes obtained by each classification model was discussed above. However, there is another type of uncertainty that is related to the final maps generated by interpolation of the available values in the sampled areas and assignment of interpolated values to unsampled areas. A single interpolated map has unknown precision (stability) and accuracy. The precision can be estimated by application of simulation methods, such as sequential Gaussian simulation (SGSIM), with a large number of realizations (simulated maps) – see the ► “Simulation” entry. Such simulation models are mostly based on kriging interpolation methods (e.g., ordinary kriging and regression kriging) that minimize or compensate for precision issues.

The difference or dissimilarity between realizations provides the estimate spatial uncertainty of the simulated models. Based on the initial and estimated values in a frequency framework, and the spatial uncertainty obtained using several realizations simulated in a Bayesian framework, areas for targeting are those with highest concentration frequency and lowest spatial uncertainty. The principal target areas for the follow-up exploration are those displaying strongly anomalous or very strongly anomalous populations. For such evaluations and to apply the established simulation algorithms, the rasterized (regular or Cartesian) grids are not required as



Uncertainty Quantification, Fig. 2 Schematic of Monte Carlo uncertainty propagation and decision-making workflow applicable to any classification model (Sadeghi 2020)

well, so they are applied to irregular grids such as point-sets (Remy et al. 2009).

The realizations are often complex and of high dimensionality, hence representing them mathematically in extremely high-dimensional Cartesian space is not feasible. As an alternative to focusing on the realizations themselves, the differences between realizations can be analyzed. To achieve this, the concept of the “distance” has been proposed by Scheidt and Caers (2008, 2009). Conceptually, the distance is a single positive value quantifying the total difference between the outputs from any two realizations (Caers 2011; Scheidt et al. 2018). The average distances or variability between realizations (similar to the standard error of the mean) can also be determined.

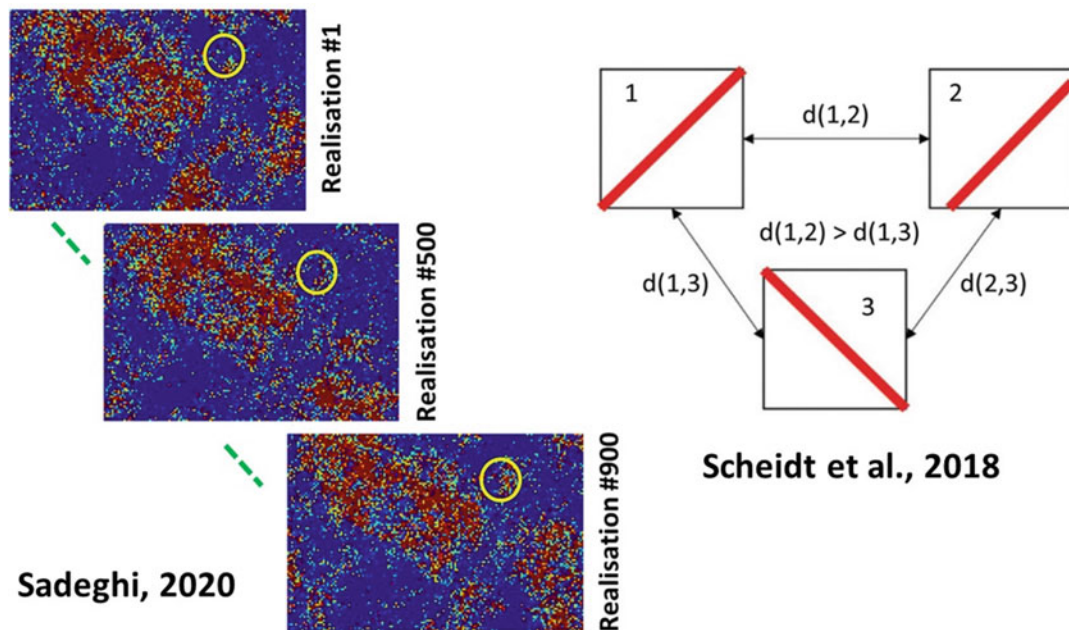
Based on the data types and the associated parameters, various metrics have been developed to quantify the distance between two realizations, such as the Hausdorff distance (Suzuki and Caers 2006), time-of-flight-based distances (Park and Caers 2007), and flow-based distance using fast flow simulators (Scheidt and Caers 2008). The simplest distance calculation in 2D models for non-dynamic data such as regolith geochemical mapping data is the Euclidean distance (Caers 2011; Scheidt et al. 2018; Fig. 3). Of course, for compositional data, all conventional methods can be used, as long as the data are represented in orthonormal log-ratio

coordinates (ilr, clr, balances) (Pawlowsky-Glahn and Buccianti 2011). Here such spatial uncertainty quantifications were introduced in general, and we do not go into mathematical details.

Summary and Conclusions

The uncertainty in the classification models has been approached from two aspects. These are (i) determination of errors in classification and ways in which the errors can be measured and (ii) comparison of different classification models to determine overall stability of the classification and especially samples defined as “anomalous” using the two test datasets where the location of a number of mineral deposits is known. In essence, it is the question of being able to assess the risk that a geochemical “anomaly” is a false indicator of mineralization based on the combination of the nature of the data and the modeling approaches.

In establishing the efficiency of models to accurately classify samples in relation to indications of the actual presence or absence of mineralization, it is again emphasized that the OA model is potentially misleading in quantifying the efficiency of classification models in orientation studies. In the classic contingency tables applied to geochemical sample



Uncertainty Quantification, Fig. 3 Schematic definition of Euclidean distance demonstrating the dissimilarities of different realizations (Sadeghi 2020; modified from Scheidt et al. 2018)

classification and the presence or absence of mineralization, the proportion of “b” samples that display a lack of mineralization signal in an area with no mineralization (or true background) and the “d” samples where there is an anomalous geochemical signal despite no mineralization being present (or the Type II errors) cannot be reliably known. The reason for this is the potential at most scales of geochemical mapping for the existence of unknown or undiscovered mineral deposits and ensuing incorrect allocation of areas or domains on maps as to the “expected background” classification.

The POA approach evaluates the efficiency of the models based only on the known related mineral deposits; hence, it is a partial assessment of efficiency as it relates only to the probability of a geochemical anomaly being present within the dispersion halo of mineralization and not the probability of having no geochemical signal in areas unaffected by mineralization. The POA approach also provides a single scenario in a frequency framework. For cases with different patterns but similar POA, deciding on the most reliable or valid approach to model geochemical data to extract univariate or multivariate patterns or signals related to the effects of mineralization was not valid.

MCSIM modeling, applied to each of the models and their individual characterized population, has partly addressed this problem as a precondition test before either final selection of the models to apply or the decision-making process for selecting the most reliable areas for follow-up exploration. The use of MCSIM on the input data requires initial estimates of (statistical) sampling uncertainty or measurement errors to determine if the potential effects of error propagation on the

final interpolated or simulated models are sufficiently low to allow for modeling to be justified. This is in comparison with a simple visual assessment of the raw data in relation to known mineralization, variations in lithology, structures, and other key supporting data. The MCSIM approach should preferentially be applied to geochemical samples separated into the main domains (typically lithology controlled) but might include landform-regolith associations in terrains different to those in Sweden or Cyprus. The metric to accept the sampling statistical certainty is “return” versus “uncertainty” (Caers 2011; Scheidt et al. 2018). The return must be higher than uncertainty otherwise the sampling (physical or statistical) may need to be improved. This approach has similarity to the question in generating data for purposes of ore reserve estimations of “when to stop drilling” (King et al. 1982).

Cross-References

- [Compositional Data](#)
- [Simulation](#)

Bibliography

- Athens ND, Caers JK (2019) A Monte Carlo-based framework for assessing the value of information and development risk in geothermal exploration. *Appl Energy*. <https://doi.org/10.1016/j.apenergy.2019.113932>
- Bárdossy G, Fodor J (2004) Evaluation of uncertainties and risks in geology. Springer, Berlin

- Bedford T, Cooke R (2001) Probabilistic risk analysis, foundations and methods. Cambridge University Press. ISBN 978-052-1773-20-1
- Buccianti A, Grunsky E (2014) Compositional data analysis in geochemistry: are we sure to see what really occurs during natural processes? *J Geochem Explor* 141:1–5
- Caers JK (2011) Modeling uncertainty in earth sciences. Wiley, Hoboken
- Carranza EJM (2011) Analysis and mapping of geochemical anomalies using logratio-transformed stream sediment data with censored values. *J Geochem Explor* 110:167–185
- Chilès JP, Delfiner P (2012) Geostatistics: modeling spatial uncertainty, 2nd edn. Wiley, Hoboken
- Costa JF, Koppe JC (1999) Assessing uncertainty associated with the delineation of geochemical anomalies. *Nat Resour Res* 8:59–67
- Deutsch CV, Journel AG (1998) GSLIB. Geostatistical software library and User's guide. Oxford University Press, New York
- Filzmoser P, Hron K, Reimann C (2009) Univariate statistical analysis of environmental (compositional data): problems and possibilities. *Sci Total Environ* 407:6100–6108
- Gallo M, Buccianti A (2013) Weighted principal component analysis for compositional data: application example for the water chemistry of the Arno river (Tuscany, Central Italy). *Environmetrics* 24:269–277
- Grunsky EC, Kjarsgaard BA (2016) Recognizing and validating structural processes in geochemical data: examples from a diamondiferous kimberlite and a regional lake sediment geochemical survey. In: Martin-Fernandez JA, Thio-Henestrosa S (eds) Compositional data analysis, Springer proceedings in mathematics and statistics, vol 187. Springer, Cham, pp 85–115. 209 p
- Helton JC, Johnson JD, Oberkampf WL (2004) An exploration of alternative approaches to the representation of uncertainty in model predictions. *Reliab Eng Syst Saf* 85:39–71
- Heuvelink GBM, Burrough PA, Stein A (1989) Propagation of errors in spatial modeling with GIS. *Int J Geog Info Sys* 3:303–322
- King H, McMahon DW, Bujtor GJ (1982) A guide to the understanding of ore reserve estimation. *AusIMM Rpt* 281
- Kiureghian AD, Ditlevsen O (2009) Aleatory or epistemic? Does it matter? *Struct Saf* 31:105–112
- Koch GS, Link RF (1970) Statistical analysis of geological data, vol I. Wiley, New York, 375 p
- Kreuzer OP, Etheridge MA, Guj P, Maureen E, McMahon ME, Holden DJ (2008) Linking mineral deposit models to quantitative risk analysis and decision-making in exploration. *Econ Geol* 103:829–850
- Lima A, De Vivo B, Cicchella D, Cortini M, Albanese S (2003) Multi-fractal IDW interpolation and fractal filtering method in environmental studies: an application on regional stream sediments of Campania region (Italy). *Appl Geochem* 18:1853–1865
- Madani N, Sadeghi B (2019) Capturing hidden geochemical anomalies in scarce data by fractal analysis and stochastic modeling. *Nat Resour Res* 28:833–847
- McCuaig TC, Kreuzer OP, Brown WM (2007) Fooling ourselves – dealing with model uncertainty in a mineral systems approach to exploration. In: Proceedings of the ninth biennial SGA meeting, Dublin
- McCuaig TC, Porwal A, Gessner K (2009) Fooling ourselves: recognizing uncertainty and bias in exploration targeting. *Centre Explor Target* 2:1–6
- McCuaig TC, Beresford S, Hronsky J (2010) Translating the mineral systems approach into an effective targeting system. *Ore Geol Rev* 38:128–138
- Mert MC, Filzmoser P, Hron K (2016) Error propagation in isometric log-ratio coordinates for compositional data: theoretical and practical considerations. *Math Geosci* 48:941–961
- Oberkampf WL, DeLand SM, Rutherford BM, Diegert KV, Alvin KF (2002) Error and uncertainty in modelling and simulation. *Reliab Eng Syst Saf* 75:333–357
- Oberkampf WL, Helton JC, Joslyn CA, Wojtkiewicz SF, Ferson S (2004) Challenge problems, uncertainty in system response given uncertain parameters. *Reliab Eng Syst Saf* 85:11–19
- Pakyuz-Charrier E, Lindsay M, Ogarko V, Giraud J, Jessel M (2018) Monte Carlo simulation for uncertainty estimation on structural data in implicit 3-D geological modelling: a guide for disturbance distribution selection and parameterization. *Solid Earth* 9:385–402
- Park K, Caers JK (2007) History matching in low-dimensional connectivity vector space. *Stanford Univ SCRF Rpt* 20
- Pawlowsky-Glahn V, Buccianti A (2011) Compositional data analysis: theory and applications. Wiley, Hoboken, 378 p
- Porwal A, Carranza EJM, Hale M (2003) Artificial neural networks for mineral-potential mapping: a case study from Aravallia province, Western India. *Nat Resour Res* 12:155–171
- Pospiech S, Tolosana-Delgado R, van den Boogaart KG (2020) Discriminant analysis for compositional data incorporating cell-wise uncertainties. *Math Geosci*. <https://doi.org/10.1007/s11004-020-09878-x>
- Pyrz MJ, Deutsch CV (2014) Geostatistical reservoir modeling. Oxford University Press
- Remy N, Boucher A, Wu J (2009) Applied geostatistics with SGeMS (a user's guide). Cambridge University Press, Cambridge, 264 p
- Rose AW, Hawkes HE, Webb JS (1979) Geochemistry in mineral exploration, 2nd edn. Academic Press, London
- Sadeghi B (2020) Quantification of uncertainty in geochemical anomalies in mineral exploration. PhD thesis, University of New South Wales
- Sadeghi B, Madani N, Carranza EJM (2015) Combination of geostatistical simulation and fractal modeling for mineral resource classification. *J Geochem Explor* 149:59–73
- Sagar BSD, Cheng Q, Agterberg F (2018) Handbook of mathematical geosciences. Springer, Berlin
- Scheidt C, Caers JK (2008) Uncertainty quantification using distances and kernel methods – application to a Deepwater Turbidite reservoir. pangea.stanford.edu, pp 1–29
- Scheidt C, Caers JK (2009) Representing spatial uncertainty using distances and kernels. *Math Geosci* 41:397–419
- Scheidt C, Li L, Caers JK (2018) Quantifying uncertainty in subsurface systems, American Geophysical Union. Wiley, New York
- Singer DA (2010) Progress in integrated quantitative mineral resource assessments. *Ore Geol Rev* 38:242–250
- Singer DA, Menzie WD (2010) Quantitative mineral resource assessments—an integrated approach. Oxford University Press, New York
- Stanley CR (2003) Estimating sampling errors for major and trace elements in geological materials using a propagation of variance approach. *Geochem Explore Environ Anal* 3:169–178
- Stanley C, Lawie D (2007) Average relative error in geochemical determinations: clarification, calculation, and a plea for consistency. *Explor Min Geol* 16(3–4):267–275
- Stanley C, O'Driscoll NJ, Ranjan P (2010) Determining the magnitude of true analytical error in geochemical analysis. *Geochem Explor Environ Anal* 10(4):355–364
- Suzuki S, Caers JK (2006) History matching with and uncertain geological scenario. *SPE Ann. Tech Conf*
- Taylor JR (1982) An introduction to error analysis: the study of uncertainties in physical measurement. Oxford University Press, Sausalito
- Verly G, Brisebois K, Hart W (2008) Simulation of geological uncertainty, resolution porphyry copper deposit. In: Proceedings of the eighth geostatistics congress, Gecamin, vol 1, pp 31–40
- Walker WE, Harremoës P, Rotmans J, van der Sluijs JP, van Asselt MBA, Janssen P (2003) Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *Integr Assess* 4: 5–17
- Yilmaz H, Cohen DR, Sonmez FN (2017) Comparison between the effectiveness of regional BLEG and <80# stream sediment

- geochemistry in detection of precious and base metal mineral deposits in Western Turkey. *J Geochem Explor* 181:69–80
- Zuo R, Zhang Z, Zhang D, Carranza EJM, Wang H (2015) Evaluation of uncertainty in mineral prospectivity mapping due to missing evidence: a case study with skarn-type Fe deposits in Southwestern Fujian Province, China. *Ore Geol Rev* 71:502–515
- Zuzolo D, Cicchella D, Albanese S, Lima A, Zuo R, De Vivo B (2018) Exploring uni-element geochemical data under a compositional perspective. *Appl Geochem* 91:174–184

Unit Regional Production Value

Gunes Ertunc

Department of Mining Engineering, Hacettepe University,
Ankara, Turkey

Definition

The unit regional production value can be described as the accumulated amount of produced resource at a period of time. The proration over a specific region area yields the term unit regional value.

Introduction

The fundamental question underlying the exploration of non-renewable natural resources decision concerns the value of a region when it was developed. The importance of this knowledge emerged when modeling a global strategy for the exploration of nonrenewable natural resources and the feasibility of assessing that particular region. Generally, the value of a region is based on the mineral occurrence which is expressed in terms of prices per area or volume.

The investment decision is based on the potential of the nonrenewable resource in the region. This decision relies on the exploration of this resource, and most of the time, it is risky. One basis, and perhaps the rudimentary way, for estimating the value of resource is the historical production records of similar regions.

In order to compare regions by means of resources and/or reserves, the value of the attribute should be standardized. These standardized values found for the various regions can be used as a comparative tool and can be used as a guide to what to expect in large regions when systematic development is undertaken. Also, this allows to estimate the potential value of a region before the exploration and development phase.

The value of the resources of an area is a variable that varies from one another. Furthermore, their values are independent and do not display conventional random variation. Hence, these values of the mineral resources of a region can be regarded as a good example of a regionalized variable

(actual functions taking a definite value each point of space). Standardizing mineral resource estimations by dividing by its area leads to a regionalized variable called the unit regional value.

Griffiths (1978) suggests mineral resource assessment may be estimated by studying the variation in unit regional value in dollars and/or tons using past production records as a basis. For similar areas or orebodies, this regionalized variable can be used as the basis for resource assessment.

Cybernetic model of mineral resource assessment (Fig. 1) is reproduced from Griffiths (1978). The input of the system is divided into two major characteristics. They are acquired and inherited, respectively. The acquired characteristics are mainly related to financial aspects, such as capital investments, and they can be generalized as socioeconomic factors. On the other hand, inherited characteristics are related to the geological events that control the mineral resources. These inputs pass through the black box processes of exploration to locate the resource, and subsequent exploitation, development, and marketing, and so become transformed into output. The outputs of this system are expressed in dollars, and tonnes are converted to the unit regional value.

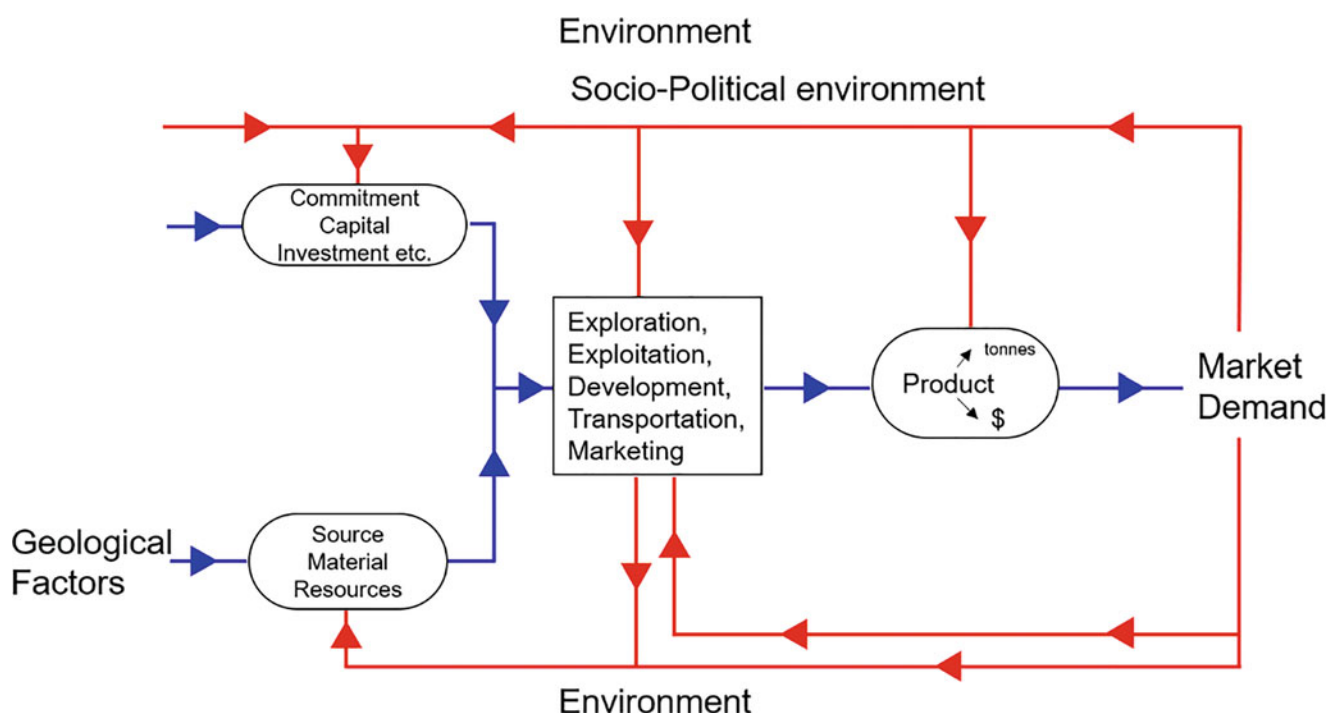
In this model, market demand is the main driving force and feedbacks coming from this demand loops the whole system and has a direct influence on capital investments and processes for technologies. Additionally, new discoveries of a new resource also update the geological knowledge, as an input.

To apply the unit regional production value approach to estimate the resources of an area of interest requires a background of information from areas with substantial production over an extended period of time. All regions are presumed that they are equal in value, and then this hypothesis is tested against the production figures. If the hypothesis is accepted, then the expected value of a large region can be used as a guide to the potential value of its smaller subregions.

For example, in the period of 1880–1964, the USA has produced nonrenewable resources to the value of \$399 B, and roughly it is worth 132,309 \$/mi². Of this total, fuels account for 65%, nonmetallics are 19%, and metals represent 16%. It can be concluded that the fuel resources of coal and petroleum products are thus 3 to 4 times more valuable than either of the other kinds of resources (Griffiths 1969).

The estimated averages obtained for a period of time can be used to reveal the production tendencies as well. For example, the difference between the two estimates of average value for the USA (1880–1964 and 1911–1964) is \$5524 per square mile which indicates that the bulk of the value has been produced after 1911.

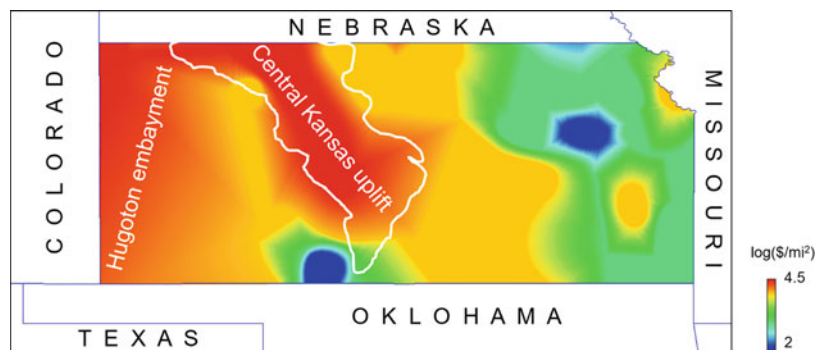
Each state or subregions productions' and the value of that productions can be compared by means of average value. This average value is equivalent to expected value. The situation is not related entirely to size of deposit, position, or geology. In the example above, while the fuel-rich states achieve, states



Unit Regional Production Value, Fig. 1 Simple cybernetic model of mineral resource subsystem. (Modified from Griffiths 1978)

Unit Regional Production

Value, Fig. 2 Log dollar value per square mile for Kansas, 1960. (Reproduced and modified after Griffiths 1969)



with no outstanding fuel resources also exceed their expected values. Their level of achievement depends essentially on their intensity of development.

Griffiths (1969) analyzed the State of Kansas in terms of value per square mile per county. Kansas has produced mineral industry resources to the value of \$16.2B (in 1911 to 1964) against an expected value of \$6.5B. In this study, mineral industry productions were allocated over counties orthogonal state-wide grid. Then the interpolated contour map is made available. In Fig. 2, the thematic representation of dollars per square mile in log scale is represented. From Fig. 2, the contribution of the petroleum resources of the Central Kansas Uplift can be seen clearly.

Clark et al. (1987) utilizes urv technique in order to estimate the mineral resources of the People's Republic of China.

These estimates are based on geologic comparisons with the 50 US states, and revealed that People's Republic of China is well endowed in major metallic and hydrocarbon resources.

Raju and Misra (1994) were evaluated for different coal-bearing states of India by taking West Bengal as standard.

Summary

The exploration and development of nonrenewable natural resources may be profitable, but the decision to explore is based both on its potential value and the risk involved in realizing this potential. Variation in the unit regional production value can be utilized as a standardized tool for the basis for resource assessment.

Cross-References

► Unit Regional Value

Bibliography

- Clark AL, Dorian JP, Fan P (1987) All estimate of the mineral resources of China. *Resour Policy* 13(1):68–84
- Griffiths JC (1969) The unit regional value concept and its application to Kansas, special distribution publication No. 38: State Geol. Surv., The Univ. Kansas, Lawrence, Kansas, 411pp
- Griffiths JC (1978) Mineral resource assessment using the unit regional value concept. *Math Geol* 10:441–472
- Raju MVB, Misra GB (1994) Undiscovered unit regional resources of Gondwana coals in India. *Int J Coal Geol* 25:133–144

Unit Regional Value

Frits Agterberg
Central Canada Division, Geomathematics Section,
Geological Survey of Canada, Ottawa, ON, Canada

Definition

The unit regional value measures the value of all resources of a specific region produced in the past. This total value is prorated over the area of the region considered. The method was invented by J.C. Griffiths at the Pennsylvania State University in 1967 and applied to many regions, especially during the first 20 years after its introduction. The method advocates a systematic approach to the collection of past mineral production data which are integrated using their monetary value. Not only are metal mining and hydrocarbon production data integrated with one another, industrial materials are incorporated as well. Usually, the estimates were conditioned on generalized geological composition of the region.

Introduction

The concept of unit regional value was originally introduced by Griffiths (1967a) in an abstract providing the basic idea and its application to the state of Kansas. This abstract was followed by Griffiths (1968a) explaining the approach in detail. Later it was applied extensively by Griffiths, his PhD and MSc students at the Pennsylvania State University, and by other scientists working on the appraisal of mineral resources. Later detailed descriptions of the method with applications include Griffiths (1978) and Gill and Griffiths (1984).

Basically, the method measures the resources of some specific region in terms of the value of resource produced: this measure is cumulated over the period of production and prorated over the area of the region. For example (*cf.* Griffiths 1978, p. 441), using 1 km² as unit of area, the average unit regional value (*u.r.v.*) of all resources produced over the period 1905 to 1972 for the 48 conterminous states of the USA was \$54,954 (measured in 1967 U.S. dollars). Alaska, on the other hand, possessed a *u.r.v.* of \$2,738 using the same measure. Because the *u.r.v.* of Alaska is about 20 times less than the *u.r.v.* of the conterminous states of the USA, excellent potential for future development on the mineral resources in Alaska is indicated.

A spin-off from the concept of unit regional value was the idea of grid-drilling (Drew 1967). In general, the location of deep drill-holes is based on geological considerations indicating that the probability that a valuable mineral deposit might be located underneath is relatively high at a particular place, for example, as indicated by a geophysical image or its possible occurrence above a geochemical anomaly. However, if mineral resources of different kinds are evenly distributed, a systematic search by grid-drilling might result in unexpected discoveries of hidden resources of many different kinds. Griffiths (1967b) proposed that the USA be systematically explored for nonrenewable mineral resources on a 20-mile square grid. The economic success of this grid-drilling would be based upon this multicommodity approach in which all targets that are economically exploitable are considered desirable and searched for at the same time.

The *u.r.v.* approach and grid-drilling are both based on the assumption that mineral resources of different kinds are evenly distributed across the crust of the Earth when measured systematically and when sufficiently large samples are taken. Systematic inventory is seen as the first step toward discovery of new resources regardless of the nature of the targets that could be of many different kinds.

Incorporation of Geological Information

Originally, the *u.r.v.* estimates were unconditional. However, almost immediately, the method was refined by incorporating quantitative estimates of the geological composition of the regions considered (Griffiths 1968b). Later, a computer system called COMOD was developed (Labovitz et al. 1977). It incorporated quantitative estimates of geological information for any region. Geological maps for the study regions were point-counted yielding proportions by area of each of the lithological units represented on the maps. For purposes of comparing different regions with one another, it is necessary to standardize the rock units distinguished on the maps. The proportions of different rock types were expressed in terms of

65 standardized time-petrographic units (see Griffiths 1978, for details).

Accumulated data for all 50 states of the USA yielded a newly defined diversity index based on 51 rock types. In comparison with other states, Alaska is rich in “detrital high-rank graywacke.” From among the other conterminous states, this made it comparable with the state Nevada according to the COMOD system. Outside the USA, it was found that New Zealand also possesses characteristics which are close to those of Alaska. Since Nevada and New Zealand both possess much higher mineral resource diversity, it is likely that the extra resources produced in these other two regions also should be present in Alaska (Griffiths 1978).

The basic ideas underlying the *u.r.v.* and its generalization incorporating standardized time-petrographic units remain closely associated with John Cedric Griffiths who made other important contributions to mathematical geoscience as well. As discussed in more detail by Donald Singer in Chapter (. . .), Griffiths was one of the founders of the International Association for Mathematical Geosciences which, in 1996, established the Griffiths Award for excellence in the teaching of mathematical geoscience. Griffiths died in 1992 but his geoscientific contributions retain their significance. With respect to *u.r.v.*, incorporation of geological information was later expanded in Griffiths et al. (1997) where, in total, 413 geologic maps at 292 different scales were transformed into a set of 65 three-digit numbers called the time-petrographic index for potential use in other regional *u.r.v.*'s. The approach probably would need significant updating for new applications.

Original Regional Applications

The original *u.r.v.* concept was generalized in different ways. For example, the *u.r.v.*'s of the 48 conterminous states of the USA, when analyzed separately, satisfy a lognormal distribution. Consequently, logarithmically transformed *u.r.v.*'s could be subjected to many standard statistical tests that are based on the normal (Gaussian) distribution. Also, logarithmically transformed *u.r.v.*'s in some applications were subjected to methods of time series analysis (Griffiths 1978). Other generalizations were based on the geology of the regions that were investigated (see previous section). The COMOD software package (Labovitz et al. 1977) standardized mineral resources in terms of 74 commodities and automatically converted the value of resources to deflated 1967 US dollars in all applications.

Gill and Griffiths (1984) listed the regions for which the *u.r.v.* methodology was applied. These include Canada by province (Labovitz 1978), Mexico by state (Chavez-Martinez 1978), South Africa by state (Menzie 1977), Australia by state (Engender 1978), New Zealand (Watson 1977), United Kingdom and Ireland (Walsh 1979), Venezuela by state (Arteaga

1978), and Israel (Gill and Griffiths 1984). In total, areal value estimation studies were carried out for 11 different countries, encompassing a total of 111 politically administratively defined regions such as states or provinces. An important overall conclusion from these many applications was that it could be assumed that all regions above a size of about 5,000 square kms are equally valuable with respect to total endowment in natural resources regardless of their inherent geological characteristics (Griffiths and Singer 1971). Later, the *u.r.v.* methodology was also applied to India by state (Raju and Misra 1991).

A characteristic feature of the *u.r.v.* approach is that it simultaneously considers all commodities produced from the Earth's crust, not only metals and hydrocarbons but also construction materials. In many regions, most of the past production relates to metals, nonmetals, and hydrocarbons. However, in more densely populated regions, most accumulated value can be associated with construction materials. For example, in Israel, the latter accounted for 53.6% of total cumulative output (Gill and Griffiths 1984) followed by non-metals (34.7%), metals (8.0%), and fuels (5.7%).

Summary and Conclusions

The unit regional value (*u.r.v.*) measures the resources of any region in terms of the value of resource produced in the past prorated over the area of the region. In its first example of application (*cf.* Griffiths 1978, p. 441), using 1 km² as unit of area, the average unit regional value (*u.r.v.*) of all resources produced over the period 1905 to 1972 for the 48 conterminous states of the USA was estimated to be \$54,954 (measured in 1967 US dollars). Application to Alaska, on the other hand, resulted in a *u.r.v.* of \$2,738. Because the *u.r.v.* of Alaska is about 20 times less than the *u.r.v.* of the conterminous states of the USA, excellent potential for future development on the mineral resources in Alaska was indicated. In subsequent applications of this approach, the geology of the study regions was incorporated to condition the *u.r.v.* and the computer system COMOD was used. The approach probably would need significant updating for new applications.

Cross-References

- Exploration Geochemistry
- Mineral Prospectivity Analysis

Bibliography

- Arteaga DM (1978) Unit regional value of the mineral resources in Venezuela. MSc thesis, Pennsylvania State University

- Chavez-Martinez L (1978) The unit regional value of the mineral resources of Mexico. MSc thesis, Pennsylvania State University
- Drew LJ (1967) Grid-drilling exploration and its application to the search for petroleum. *Econ Geol* 62:698–710
- Engender PR (1978) An application of unit regional value concept to a study of the mineral resources of Australia. PhD thesis, Pennsylvania State University
- Gill D, Griffiths JC (1984) Areal value assessment of the mineral resources' endowment of Israel. *Math Geol* 16:37
- Griffiths JC (1967a) Unit regional value concept and its application to Kansas. *AAPG Bull* 51:1688
- Griffiths JC (1967b) Mathematical exploration strategy and decision-making. In: *Proceedings, 5th World Petr Cong Mexico*, pp 599–604
- Griffiths JC (1968a) Mineral resource assessment using unit regional value concept. *Math Geol* 10:441–472
- Griffiths JC (1968b) Classification of geological data. *Western Miner* 41:37–42
- Griffiths JC (1978) Mineral resource assessment using the unit regional concept. *J Math Geol* 10:441–472
- Griffiths JC, Singer DA (1971) Unit regional value of non-renewable natural resources as a measure of potential of development of large regions. *Geol Soc Aust Spec Publ* 3:227–238
- Griffiths JC, Pilant AD, Smith CM (1997) Quantitative estimates of the geology of large regions and their application to mineral-resource assessment. *Nonrenewable Resour* 6:157–236
- Labovitz ML (1978) Unit regional value of the Dominion of Canada. PhD thesis, Pennsylvania State University
- Labovitz ML, Menzie WD, Griffiths JC (1977) COMOD: a program for standardizing mineral resource commodity data. *Comput Geosci* 3(3):497–537
- Menzie WD (1977) The unit regional value of the Republic of South Africa. Unpublished PhD thesis, Pennsylvania State University
- Raju MVB, Misra GB (1991) An evaluation of the undiscovered mineral resources of India based on the concept of unit regional value. *Math Geol* 23:841–852
- Walsh DA (1979) An assessment of mineral resources of the U.K. and Republic of Ireland. MSc thesis, Pennsylvania State University
- Watson AJ (1977) An appraisal of the mineral resources of New Zealand. MSc thesis, Pennsylvania State University

Univariate

Alejandro C. Frery
School of Mathematics and Statistics, Victoria University of Wellington, Wellington, New Zealand

Synonyms

Real-valued distributions; Real-valued samples

Definitions

The term “univariate” usually refers to two different entities, namely, distributions and sample statistics. In both cases, it designates a function that depends on a single (typically real) variable.

Univariate Distributions

A real univariate random variable is a function $X: \Omega \rightarrow S \subset \mathbb{R}$. The distribution of X is uniquely characterized by its probability distribution function $F_X(t) = \Pr(X \leq t)$, or by its characteristic function $\phi_X(t) = E(e^{itX})$, where $\mathbf{j} = \sqrt{-1}$ and E denotes the expected value. Examples of univariate distributions are the Poisson and Normal laws.

Univariate Statistics

Consider the sample of real-valued observations $\mathbf{x} = (x_1, x_2, \dots, x_n)$. Sample univariate statistics are transformations of $\mathbf{x}$ that aim at describing interesting quantities.

Overviews

Univariate Distributions

Real random variables can be continuous, discrete, or singular. In practice, only the two first types (and their mixture) are of interest. The distribution of a real continuous random variable is also characterized by its probability density function $f_X(t) = dF_X(t)/dt$; this function is only positive on the support S . The probability function characterizes the distribution of a real discrete random variable. Without loss of generality, assume that the support is a subset of the integer numbers $\mathbb{Z}$. The probability function is the pair $(i, \Pr(X = i))$, for every $i \in S$.

The distribution of both continuous and discrete random variables is the subject of several other Encyclopedia entries. The reader is referred to the compendia by Johnson et al. (1997) for discrete models, and by Johnson et al. (1994, 1995) for continuous distributions.

Univariate Statistics

Some quantities of interest are:

$$\begin{aligned} \text{sample mean } m_1(\mathbf{x}) &= \bar{x} = \frac{1}{n} \sum_{\ell=1}^n x_\ell, \\ \text{sample moment of order } r \ m_r(\mathbf{x}) &= \frac{1}{n} \sum_{\ell=1}^n x_\ell^r, \\ \text{central sample moment of order } r \ \tilde{m}_r(\mathbf{x}) &= \frac{1}{n} \sum_{\ell=1}^n (x_\ell - \bar{x})^r. \end{aligned} \tag{1}$$

Note that $s^2(\mathbf{x}) = m_2(\mathbf{x})$ is the sample variance, and that $s(\mathbf{x})$ is the sample standard deviation. Equation (1) provides an estimate of the location of the distribution that produced the sample $\mathbf{x}$, and $\tilde{m}_2(\mathbf{x}) = m_2(\mathbf{x}) - m_1^2(\mathbf{x}) = s^2(\mathbf{x})$ is an estimate of its variance. All sample moments are well defined, even if their distributional counterparts are not.

Two important higher-order sample moments are the sample skewness and the sample kurtosis. They provide estimates of the asymmetry and the tailedness of the distribution that produced the sample $\mathbf{x}$:

$$\hat{\gamma}_1(\mathbf{x}) = \frac{\tilde{m}_3(\mathbf{x})}{s^3(\mathbf{x})},$$

$$\hat{\gamma}_2(\mathbf{x}) = \frac{\tilde{m}_4(\mathbf{x})}{s^4(\mathbf{x})},$$

and $\hat{\gamma}_2(\mathbf{x}) - 3$ is called “excess kurtosis,” as a reference to the kurtosis of a Gaussian random variable which is 3.

If we may assume that the sample is a collection of independent identically distributed random variables with positive support, it is often interesting to examine the second-kind or log-moments, which are given by

$$k_1(\mathbf{x}) = \frac{1}{n} \sum_{\ell=1}^n \log x_{\ell}, \text{ and}$$

$$\tilde{k}_m(\mathbf{x}) = \frac{1}{n} \sum_{\ell=1}^n (\log x_{\ell} - k_1(\mathbf{x}))^r.$$

Univariate sample statistics can also depend on order statistics. Let $\mathbf{x} = (x_{1:n}, x_{2:n}, \dots, x_{n:n})$ denote the sample $\mathbf{x}$ sorted in nondecreasing order, i.e., $x_{1:n} \leq x_{2:n} \leq \dots \leq x_{n:n}$. The minimum and maximum of $\mathbf{x}$ are, respectively, $x_{1:n}$ and $x_{n:n}$. If the sample size is odd, then the median is $q_{1/2}(\mathbf{x}) = x_{(n+1)/2:n}$. More generally, and regardless of the sample size, the sample quantile of order $\alpha \in (1/n, 1 - 1/n)$ is a weighted average of consecutive order statistics. Hyndman and Fan (1996) provide a comprehensive account of how several computational platforms implement sample quantiles. The interquartile range is the difference $\text{IQR}(\mathbf{x}) = q_{3/4}(\mathbf{x}) - q_{1/4}(\mathbf{x})$.

A summary of sample statistics should contain, at least, the sample size, minimum, first quartile, median, mean, third quartile, and maximum values, as shown in Table 1.

Univariate, Table 1 Summary statistics (rounded to two decimal places)

Name	Notation	Value
Sample size	n	50
Minimum	$x_{1:n}$	0.07
First quartile	$q_{1/4}(\mathbf{x})$	1.09
Median	$q_{1/2}(\mathbf{x})$	2.20
Mean	$\bar{\mathbf{x}}$	2.46
Third quartile	$q_{3/4}(\mathbf{x})$	3.32
Maximum	$x_{n:n}$	6.70

The simplest graphical display of a univariate sample is a “rug plot,” which consists of marks corresponding to each observation along an axis. Rug plots are seldom useful alone; they complement the information provided by the two most commonly seen displays of univariate samples: the boxplot (and variations), and the histogram (and its smoothed versions).

The classical “in the style of Tukey” boxplot shows five summary statistics (the median, two hinges, and two whiskers), and all “outlying” points individually. The hinges are the first and third quartile. The right whisker is set at the smallest value that is larger than or equal to $q_{3/4}(\mathbf{x}) + \frac{1}{2}\text{IQR}(\mathbf{x})$. The left whisker is set at the largest value that is smaller than or equal to $q_{1/4}(\mathbf{x}) - \frac{1}{2}\text{IQR}(\mathbf{x})$. It is useful to plot a notch around the median, usually of size $1.58n^{-1/2}\text{IQR}(\mathbf{x})$ that corresponds to approximately a 95% symmetric confidence interval. Outlying points are above the right whisker and below the left whisker.

Figure 1 shows the elements of the boxplot of the sample whose summary statistics were presented in Table 1. The observations are mapped onto the rug, in maroon (Fig. 1). The only outlying observation is shown as a red dot. Notice that the horizontal grid has been omitted, as it carries no useful information in this plot and may be misleading (Fig. 2).

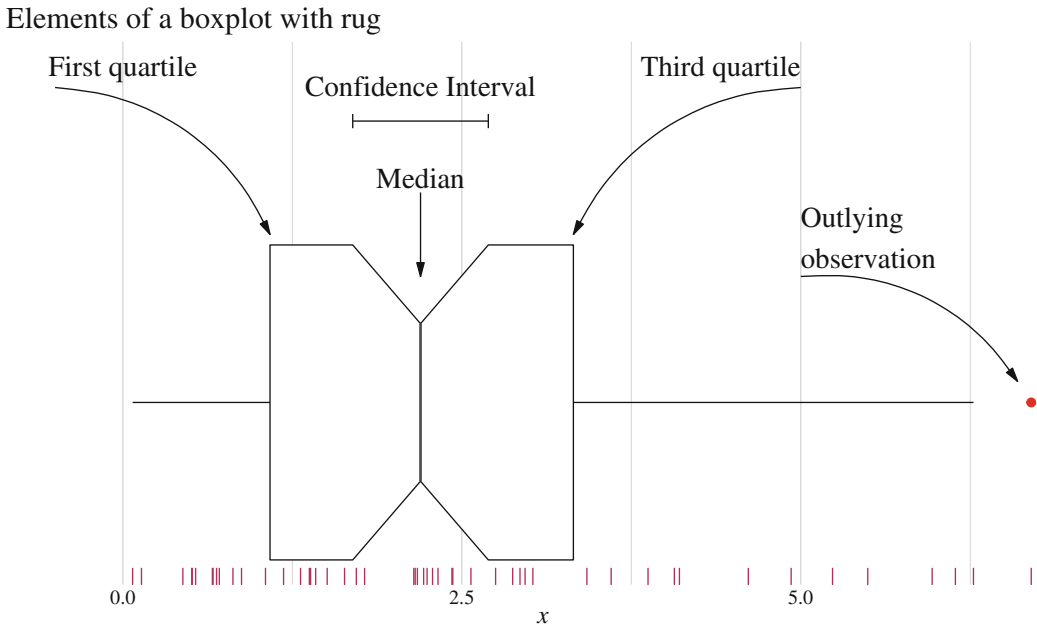
Boxplots are particularly useful at comparing two or more samples. Figure 2 shows the boxplots of three samples. They have approximately the same sample means, but the spread of the data is noticeably different.

While boxplots are built with summaries based on the sorted sample, the histogram and its variations employ all the data. As such, the rug is a histogram. More often than not, histograms show the counts of observations within disjoint intervals (“bins”), instead of marking each observation as in a rug.

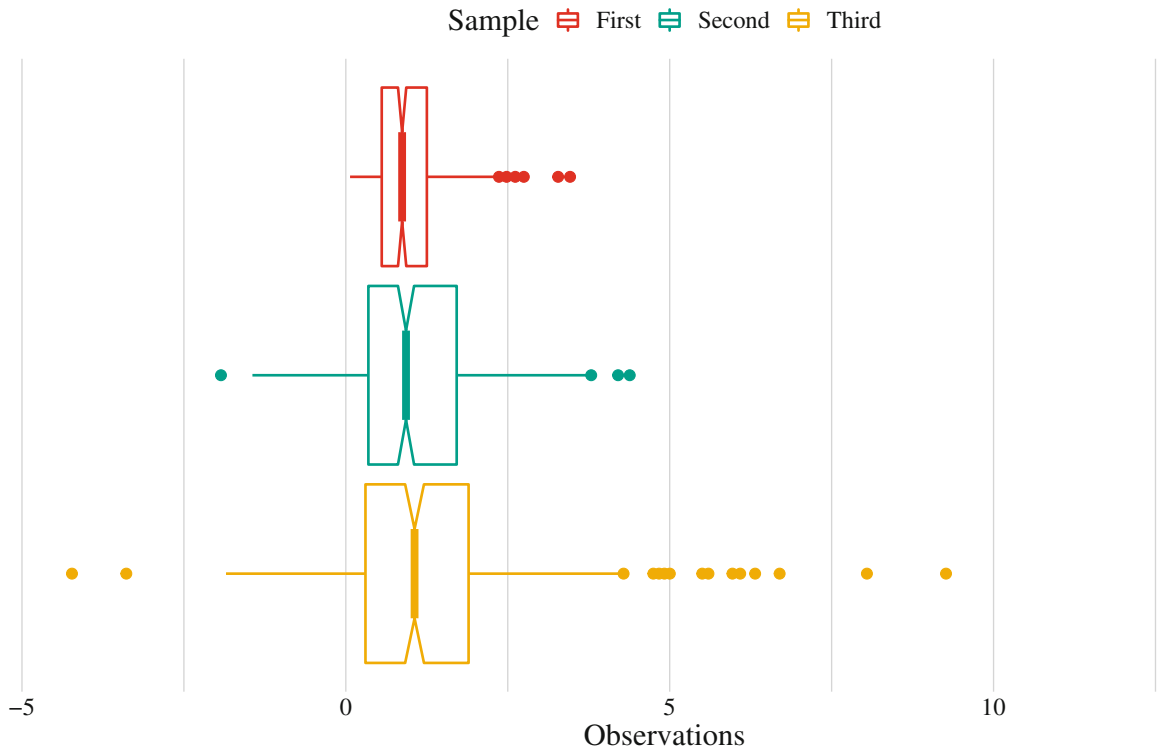
Freedman and Diaconis (1981) studied the properties of histograms as estimators of probability density functions. They proposed the rule known after their names, which states that an optimal bin width is twice the interquartile range divided by the sample size’s cubic root.

The plots in Fig. 3 show the histograms of the same data set: with the default number of bins (30 in the current version of R’s ggplot2 library; cf. Wickham 2016) and with the bin width set by the Freedman-Diaconis rule.

Although histograms as in Fig. 3 provide useful information about the sample, it is hard to show more than one sample at the same time using bars. Kernel smoothing is helpful in such cases, as shown in Fig. 4.



Univariate Fig. 1 Elements of a boxplot with rug

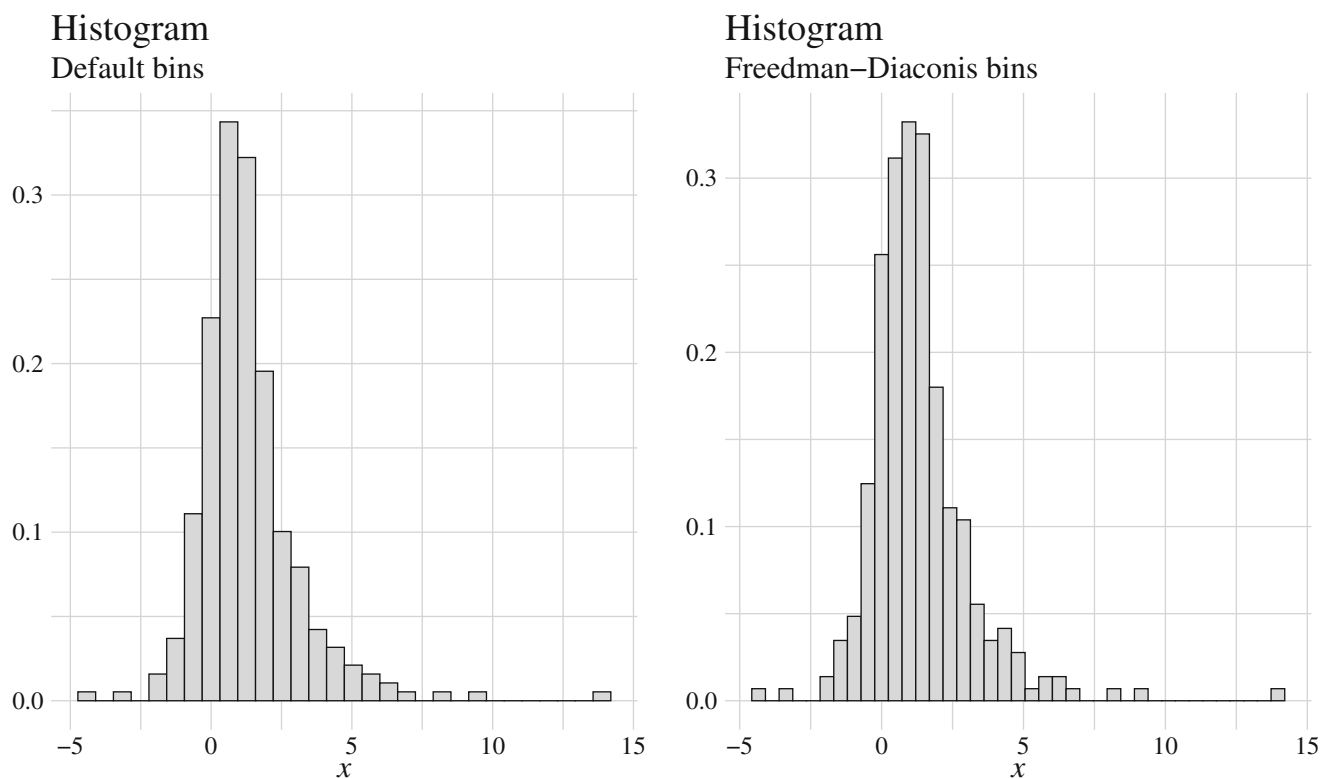


Univariate, Fig. 2 Boxplots of three samples with similar sample mean values

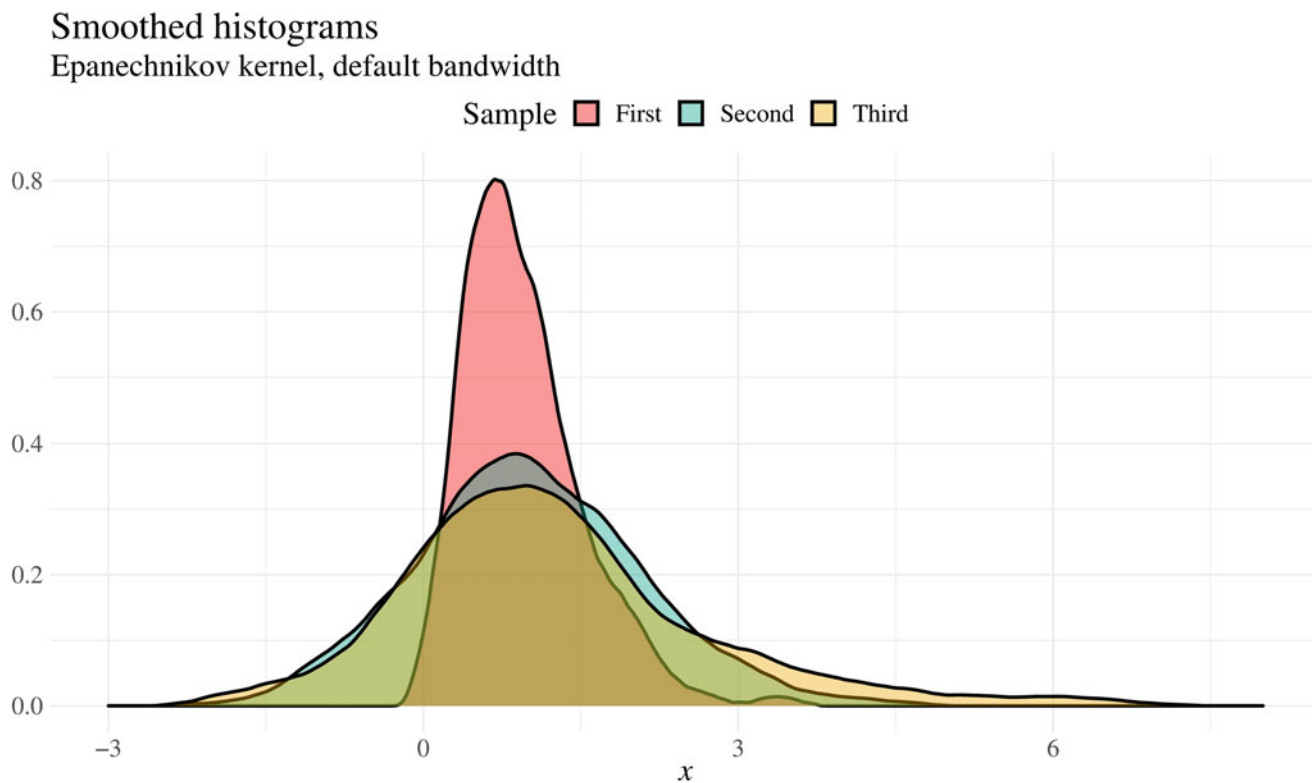
For the sake of reproducibility and replicability, it is a good practice to state the number of bins or the bin width whenever showing histograms, and the kind of kernel and the bandwidth when using smoothed histograms.

Summary

Univariate distributions are always characterized by their cumulative distribution function or by their characteristic



Univariate, Fig. 3 Two histograms and the effect of a judicious choice of the bin width (with the Freedman-Diaconis rule)



Univariate, Fig. 4 Three smoothed histograms with Epanechnikov kernels

function. Univariate statistics provide insights about the underlying model that produced the data.

Cross-References

- [Interquartile Range](#)
- [Standard Deviation](#)

Bibliography

- Freedman D, Diaconis P (1981) On the histogram as a density estimator: L2 theory. *Z Vitaminkd Verw Geb* 57(4):453–476. <https://doi.org/10.1007/bf01025868>
- Hyndman RJ, Fan Y (1996) Sample quantiles in statistical packages. *Am Stat* 50(4):361. <https://doi.org/10.2307/2684934>
- Johnson NL, Kotz S, Balakrishnan N (1994) Continuous univariate distributions, Wiley series in probability and mathematical statistics, vol 1, 2nd edn. Wiley, New York
- Johnson NL, Kotz S, Balakrishnan N (1995) Continuous univariate distributions, vol 2, 2nd edn. Wiley, New York
- Johnson NL, Kotz S, Balakrishnan N (1997) Discrete multivariate distributions. Wiley, New York
- Wickham H (2016) *ggplot2: elegant graphics for data analysis*. Springer, New York. <https://doi.org/10.1007/978-0-387-98141-3>. <https://ggplot2.tidyverse.org>

Universality

Dionissios T. Hristopoulos
School of Electrical and Computer Engineering, Technical
University of Crete, Chania, Crete, Greece

Abbreviations

- ETAS Epidemic type aftershock (model)
RG Renormalization group
SOC Self-organized criticality

Definition

The term universality originates in the theory of phase transitions and self-organized criticality. Phase transitions represent changes between different states of matter which occur when a control parameter (e.g., temperature) attains a critical threshold. The correlations of relevant system properties near a phase transition do not fall off exponentially, and thus they do not exhibit characteristic length scales. In contrast, the decay of the correlations follows power-law functions characterized by respective critical exponents. The theory of self-organized criticality posits that certain systems (e.g., sandpiles, avalanches, and geological faults in seismogenous

zones) are close to a critical threshold without tuning a control parameter such as temperature. The term “universality” indicates that phase transitions which occur in different systems are described by the same set of critical exponents, if they share certain geometric and symmetry features; systems with the same values of critical exponents belong in the same universality class (Kadanoff 2000). In a broader sense, universality refers to empirical laws that have global validity, e.g., Gutenberg-Richter’s law which relates the frequency and intensity of earthquakes. The exponent of Gutenberg-Richter’s law surprisingly takes values close to one regardless of geographical location, geological factors, and seismic history.

Overview

The term universality appears in the theories of phase transitions and self-organized criticality (Bak 2013; Kadanoff 2000). The main feature of systems close to a phase transition is that their response functions lack characteristic length scales. Instead, the correlations in such systems decay as power laws, i.e., scale-free functions determined by respective power-law exponents. The concept of universality posits that systems which are quite different in their microscopic details can exhibit identical, i.e., *universal*, macroscopic behavior. More specifically, universality refers to the fact that the power-law exponents, which describe the macroscopic response of different critical systems belonging in the same universality class, take identical values.

The emergence of universality is often justified in terms of the renormalization group (RG) theory introduced by Wilson (1971). RG provides a systematic mathematical framework for predicting the behavior of physical systems (or idealized mathematical models) under a sequence of coarse-graining transformations. These progressively increase the length scale of observation (coarse-graining), while the short-scale details are effaced; at the same time, the impact of coarse-graining on the system parameters is calculated, and their values are accordingly adjusted. Under repeated RG steps, the system “flows” toward limiting states which are known as fixed points. The defining feature of the latter is that a system remains invariant under further coarse-graining transformations at a fixed point.

Phase transitions often involve a jump between a less ordered and a more ordered state. As the transition threshold is approached from the side of the less ordered state, the system is attracted toward a nontrivial fixed point which determines the properties of the more ordered state and the values of its characteristic critical exponents. A universality class encompasses all the models that converge to the same fixed point under application of the RG procedure. Hence, in a broader sense, the term universality refers to the fact that

identical macroscopic physical laws apply to systems which may vary in their microscopic details. For example, the laws of thermodynamics do not depend on the details of atomic interactions, the statistical properties of turbulence in the inertial range are independent of the dissipation mechanisms, and the effective permeability of statistically isotropic random media is insensitive to the exact form of the local permeability correlations. In this sense, universality is a main reason for the success of the scientific approach, since it enables a general understanding of natural phenomena based on a few fundamental physical principles.

More specifically, in the context of geosciences, the term universality is used to imply the global validity of an empirical law across different geographical areas and times. For example, the Gutenberg-Richter law of seismicity is considered a universal law since its validity has been confirmed by countless observations and analysis of seismic catalogs worldwide. The exponent of Gutenberg-Richter's law takes a value close to one, although the exact value may not be universal. Note that strictly speaking, the term "universality" implies not only the same behavior (i.e., power-law), but more importantly also common critical (power-law) exponents for different systems that belong in the same universality class.

The concept of universality is also linked with fractal geometry. The fractal approach was developed and popularized by Benoit Mandelbrot (e.g., as described in his book *The Fractal Geometry of Nature*). Fractal patterns consist of *self-similar* structures which lack characteristic length scales. For example, a fractal object of radius r has surface area $S(r) \sim r^\alpha$ where $\alpha \neq 2$ is a noninteger (fractal) exponent. In phase transitions, fractal patterns (e.g., clusters of the ordered phase) emerge in systems which are close to the critical state. Fractal patterns in systems that belong in the same universality class share the same fractal exponents.

All notions of universality discussed above have applications in the geosciences. Given the wide range of applications and space limitations, it is not possible to cite important early papers by Leo Kadanoff and Per Bak. The interested readers can find more information about these works in the books by Kadanoff (2000) and Bak (2013).

Universality in Phase Transitions

Phase transitions are mechanisms that enable changes between different states of matter, e.g., water, vapor, and ice. Thus, phase transitions play a prominent role in the water cycle of the Earth. There are many types of phase transitions. Some are controlled by temperature (i.e., thermodynamic transitions), while others by the geometry of the medium in which the process takes place (i.e., geometrical transitions). An example in the first class is

the superconducting transition: Metals and ceramic oxides start to conduct electricity without any resistance if the temperature is lowered below a certain critical threshold. An example in the second class is *percolation*: A random porous medium is permeable to fluids if its porosity exceeds a critical threshold. Percolation is key to the production of an aromatic cup of filtered coffee as well as to the infiltration of rainwater into the ground.

Phase transitions are classified as first order and second order. In first-order transitions, the change of state is accompanied by the absorption or release of latent heat, which changes discontinuously the thermodynamic free energy of the system. This leads to a discontinuity in the first derivative of the free energy with respect to a thermodynamic variable (hence the term "first-order"). Typical examples include the condensation of vapor into water droplets (which involves latent heat release) and the melting of ice (mediated by latent heat absorption). In second-order phase transitions, on the other hand, the free energy changes continuously. The discontinuity appears in the susceptibility which is the second-order derivative of the free energy. A typical example is the transition between paramagnetism and ferromagnetism in magnetic materials that takes place at the Curie point. The magnetization of the material acts as the order parameter: It is zero in the paramagnetic phase and takes a finite value in the ferromagnetic state. Similarly, order parameters are defined for other second-order phase transitions (Kadanoff 2000).

Second-order phase transitions are characterized by a set of *critical exponents* that determine how relevant properties scale near the transition point. For example, the exponent ν describes the divergence of the fluctuations' correlation length, i.e., $\xi \sim |T - T_c|^{-\nu}$, where T is the temperature, T_c is the critical temperature, and ξ is the correlation length. The exponent η describes the power-law decay of the correlations with distance at the transition point, i.e., $C(r) \sim r^{-d+2-\eta}$. The exponents α , β , γ , and δ describe the dependence of: (i) the specific heat $c \sim |T - T_c|^{-\alpha}$, (ii) the order parameter $\phi \sim |T - T_c|^\beta$, (iii) the susceptibility of the order parameter to an external field h , i.e., $d\phi/dh \sim |T - T_c|^{-\gamma}$, and (iv) the dependence of the external field on the order parameter $h \sim \phi^\delta$. The critical exponents depend on the dimensionality of the system and the symmetry of the order parameter, but they may be independent of other details (e.g., lattice structure).

Percolation, Fractals, and Multifractals

Percolation is a geometric phase transition which takes place as a connected (e.g., permeable) structure grows in a random medium until it spans the entire extent of the medium. For example, this happens on a lattice system as the proportion of connected sites (e.g., the porosity) approaches a critical

threshold. There exist different percolation models which include site, bond, and continuum percolation (Isichenko 1992). Percolation can be used as a mathematical model for disordered media including heterogeneous geological porous media. The geometric properties of percolation can be applied to analyze transport in such media. Hence, percolation is an important model for the study of subsurface flow and contaminant transport, especially in highly heterogeneous and fractured soils.

Similarly to other phase transitions, percolation involves a critical threshold p_c . This corresponds to the occupation probability above which a fully connected network (medium) becomes possible. The onset of percolation is marked by the emergence of an incipient infinite cluster (essentially a “giant component”) that extends across the medium. The order parameter is given by the probability that a site belongs to this giant component. Other variables describe geometrical patterns of the percolation network (e.g., size of connected clusters, average radius of clusters of given size, probability that two sites at a given distance belong to the same cluster, etc.). Near the percolation threshold, the behavior of such variables is characterized by critical exponents. Universality in the context of percolation means that the critical exponents depend on the dimension of space in which the percolation model is embedded, but they are independent of the lattice structure and the percolation type (site and bond percolation are in the same universality class).

The critical behavior of percolation implies a fractal geometry for the network patterns (e.g., connected clusters) near the percolation threshold. The fractality is bestowed by the critical exponents which in general take noninteger values. The concepts of universality and scaling have found applications in the geometry of naturally occurring fractal patterns such as river networks, fault systems, coastlines, and mountain topography (Isichenko 1992; Dodds and Rothman 2000).

An extended notion of universality is also used in the theory of multifractals. These mathematical models find applications in natural phenomena such as clouds, rain, fluid turbulence, and weather (Lovejoy and Schertzer 2013). Multifractals have more complex scaling relations than fractals. The fractal exponents (dimensions) of multifractals depend on location, and thus a spectrum of exponents is necessary to fully describe multifractal behavior.

Self-Organized Criticality and Earthquakes

A notable application of the concept of universality is the investigation of avalanches in sandpile models by Leo Kadanoff and coworkers in a paper published in 1989 (c.f. Kadanoff 2000). They found different universality classes based on the microscopic rules used for the generation of

avalanches. Their work paved the way for the development of the theory of self-organized criticality (SOC), which describes systems that are always close to a critical threshold without tuning of an external parameter (e.g., temperature).

Self-organized criticality has been proposed as a theoretical framework that among other things can explain the laws of seismicity. In particular, the concept of self-organized criticality provides a possible explanation for the Gutenberg-Richter law which connects the frequency and magnitude of earthquake events (Bak 2013). The fact that earthquakes all over the globe follow the Gutenberg-Richter law is seen as a signature of universality. An influential paper by Olami et al. (1992) argues that the Burridge-Knopoff spring-block model used to simulate earthquakes exhibits *self-organized criticality* but lacks *detailed universality*. This means that the exponent of the Gutenberg-Richter law depends on the elastic parameters of the model (Olami et al. 1992). Recent numerical experiments and lab-generated quakes also show a dependence of the Gutenberg-Richter exponent on the material properties of the medium in which the events take place.

According to a different theoretical approach, the probability distribution of *earthquake recurrence times* follows a universal law, provided that the time between earthquakes is scaled by the local rate of seismic occurrence (Corral 2004). However, this viewpoint is controversial; other researchers argue in favor of an almost-universal law based on the epidemic type aftershock (ETAS) model of seismicity (e.g., Saichev and Sornette 2007).

Summary and Conclusions

The term universality appears in different theories of natural phenomena (phase transitions, self-organized criticality, fractals, and multifractals), and its meaning may differ in each case. The general idea of universality is that certain “universal properties” depend on a few “macroscopic” parameters but are independent of the specific details. A common signature of universality is the presence of power-law dependence, often when a critical threshold is approached. The connection between power-law functions and universality is reviewed in Marković and Gros (2014). Overall, universality is a very fruitful concept which allows studying seemingly disparate systems in the same theoretical framework and has several applications in the geosciences.

Cross-References

- Allometric Power Laws
- Multifractals
- Scaling and Scale Invariance

Bibliography

- Bak P (2013) How nature works: the science of self-organized criticality. Springer Science & Business Media, Berlin
- Corral A (2004) Long-term clustering, scaling, and universality in the temporal occurrence of earthquakes. *Phys Rev Lett* 92(10):108501. <https://doi.org/10.1103/PhysRevLett.92.108501>
- Dodds PS, Rothman DH (2000) Scaling, universality, and geomorphology. *Annu Rev Earth Planet Sci* 28(1):571–610. <https://doi.org/10.1146/annurev.earth.28.1.571>
- Isichenko MB (1992) Percolation, statistical topography, and transport in random media. *Rev Mod Phys* 64:961–1043. <https://doi.org/10.1103/RevModPhys.64.961>
- Kadanoff LP (2000) Statistical physics: statics, dynamics and renormalization. World Scientific Publishing, River Edge
- Lovejoy S, Schertzer D (2013) The weather and climate: emergent laws and multifractal cascades. Cambridge University Press, Cambridge, UK
- Marković D, Gros C (2014) Power laws and self-organized criticality in theory and nature. *Phys Rep* 536(2):41–74. <https://doi.org/10.1016/j.physrep.2013.11.002>
- Olami Z, Feder HJS, Christensen K (1992) Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes. *Phys Rev Lett* 68(8):1244. <https://doi.org/10.1103/PhysRevLett.68.1244>
- Saichev A, Sornette D (2007) Theory of earthquake recurrence times. *J Geophys Res Solid Earth* 112(B4):B04313. <https://doi.org/10.1029/2006JB004536>
- Wilson KG (1971) Renormalization group and critical phenomena. I. Renormalization group and the Kadanoff scaling picture. *Phys Rev B* 4:3174–3183. <https://doi.org/10.1103/PhysRevB.4.3174>

Upper Approximation

Touhid Mohammad Hossain¹ and Junzo Watada²

¹Department of Computer and Information Sciences, Faculty of Science and Information Technology, Universiti Teknologi PETRONAS, Seri Iskandar, Malaysia

²Graduate School of Information, Production Systems, Waseda University, Fukuoka, Japan

Definition

The principal concepts of rough set theory (RST) are reduct, core, and discovering knowledge as a form of rules. In RST, approximations are defined by using a system of definable sets. Upper and lower approximations of Pawlak Rough Set model can be used to generate certain and possible rules based on three disjoint regions, positive negative regions. The rules can explain the decision state to predict new situation. The results obtained show that in Lithology classification problem, this classification approach is quite beneficial because of high accuracy and its explainability.

Background

Since Pawlak proposed rough set theory (Pawlak 1982), a huge number of research results have been provided by means of Rough set theory (RST). RST is an effective tool especially for data mining in various applications. The benefits of RST enable researchers to experiment that. Pawlak explained some advantages in his book on Rough sets (Pawlak 2002; Hossain et al. 2021). RST is an effective tool to work with quantitative and qualitative features similarly without any preliminary information required (Zhang et al. 2016). Particularly RST was found to be useful for imprecise and inconsistent information systems to generate logical relations and feature selection. RST requires to go through some steps to induce rules from the decision table.

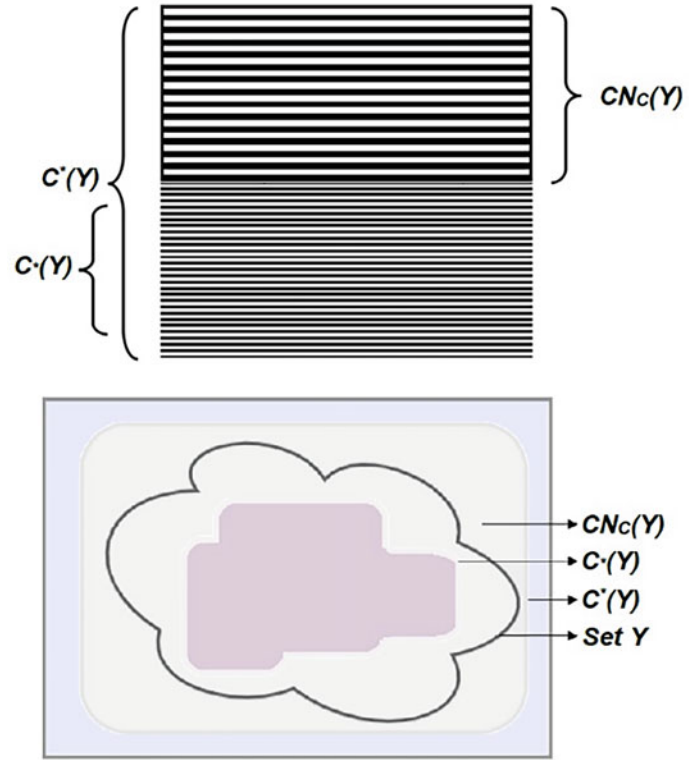
Based on a pair of approximations: upper approximation and lower approximation in RST, it is possible to learn from the examples and generate rules which are known as LERS system (Sudha and Kumaravel 2018). RST handles inconsistent information which allows to calculate lower and upper approximation results as two types of rules: certain rules and possible rules (Grzymala-Busse 1992). By calculating the certain and possible rules from the upper and lower approximations, classification problems can be solved.

This entry includes the application of lower approximation and upper approximation in generating certain and possible rules for solving lithology classification problem which is of major importance in oil and gas exploration because lithology represents the reservoir petrophysical characteristics (Hossain et al. 2020a, b).

Methodology

Decision Table and Informatixon System (IS)

RST requires the dataset to be arranged as a decision table where in the columns there are some dependent features and some independent features. Each column consists the values that indicate the property for the observations. Along the row of the table, the examples or the objects are arranged. Such a table is called an information table. An information system, on the other hand, is a pair $I = (U, K)$, where the universe is denoted as U and the feature set is denoted as K (both U and K are nonempty finite). Mathematically, $k : U \rightarrow V_k$ for $k \in K$, where V_k is denoted as the domain of k . A decision table is a form of IS that represents the retrievable knowledge in the system. Formally, decision table, $I = (U, K \cup D)$, where features in K are called condition feature and D is the decision feature. The conditional features, also known as independent features, consist of the conditional values of an object to get into a decision. Decision features may consist of binary categories I or multiple classes.

Upper Approximation,**Fig. 1** Data partitioning based on RST approximations for a set Y**Approximations in RST**

An important concept of RS is the indiscernibility relation that is formed by information that concerns the data of interest. The idea of the indiscernibility relation is to explain that due to the deficiency of knowledge it is not difficult to distinguish several objects in the concerning information system. Another vital concept of RST is approximations. In RST, there is a pair of approximations: *lowerApproximation* concerns with the object samples which definitely belong to the interest subset. On the contrary, *upperApproximation* concerns about the objects that might possibly belong to the interest subset. *Boundary Region* consists of the samples which cannot be identified as any approximation type.

Mathematically, let a set $Y \subseteq U$, C be an equivalence relation and a knowledge base $N = (U, C)$. So, the approximations can be denoted as:

$$C_*(Y) = \{y \in U : C(y) \subseteq Y\} \quad (1)$$

$$C^*(Y) = \{y \in U : C(y) \cap Y \neq \emptyset\} \quad (2)$$

Above, $C_*(Y)$ and $C^*(Y)$ are called the *C-lower-approximation* and the *C-upper-approximation* of Y , respectively.

The *boundary region* is defined as,

$$CN_C(Y) = C^*(Y) - C_*(Y) \quad (3)$$

If the borderline area of Y is empty which means if $CN_C(Y) = \emptyset$, then dataset Y is considered to be exact in

C . In other words, if $C_*(Y) = C^*(Y)$, then Y is a crisp set. Else if $CN_C(Y) \neq \emptyset$, then the data set Y is a rough set in C . Figure 1 demonstrates the definitions.

RS can be explained by a rough membership function as well (as shown below) (Hossain et al. 2021):

$$\mu_y^C(y) = |Y \cap C(y)| / |C(y)| \quad (4)$$

Apparently,

$$\mu_y^C(y) \in [0, 1] \quad (5)$$

where $\mu_y^C(y)$ is a kind of conditional probability that is represented as a degree of certainty to which y belongs to Y (or $1 - \mu_y^C(y)$).

The rough membership function can also be used for expressing the *approximations* and the *boundary region* of a set,

$$C_*(Y) = \{y \in U : \mu_y^C(y) = 1\}, \quad (6)$$

$$C^*(Y) = \{y \in U : \mu_y^C(y) > 0\}, \quad (7)$$

$$CN_C(Y) = \{y \in U : 0 < \mu_y^C(y) < 1\}. \quad (8)$$

Now we can give two definitions for rough sets shown as follows:

Definition 1 Set Y is rough with respect to C if,

$$C_*(Y) \neq C^*(Y) \quad (9)$$

Definition 2 Set Y is rough with respect to C if for some y ,

$$0 < \mu_y^C(y) < 1 \quad (10)$$

Reduction of Features

Feature reduction is a vital aspect of RST. Reducing the redundant information from the system but at the same time to keep the relation of indiscernibility is the key idea of this step. The target is to search for indispensable features and eliminating the dispensable features.

If k is an indiscernible feature in K , deleting it from the information system I will cause I to be inconsistent. In the other hand, if a feature is discernible, it could be eliminated from the dataset, and in this way the dimensionality of the dataset will be reduced.

Let L be a subset of K and k belongs to L . The following exist:

1. If $I(L) = I(L - \{k\})$, then k is *discernible* in L ; or else k is *indiscernible* in L .
2. If features are necessary, then L is *independent*.
3. If $I(L') = I(L)$ and L' is independent, then L' , the subset of L , is an L 's reduction.

Reduction consists of some functions. First, the core of the features is calculated. *Core* is basically the feature set in where the features are present in all the *reducts*. In another words, a Core consists of the features which are indispensable from the information system devoid of breaking the structure of the similarity class. Let us assume L to be a subset of K . Then the set of the necessary features in L is the *core* of L . *Core* and *reduction* are related in the following way:

$$\text{Core}(L) = \cap \text{Reduct}(L) \quad (11)$$

where $\text{Reduct}(L)$ is the list of all the reductions s of L .

Certain and Possible Rules Generation

Once the reduction is done, the *core* finding is done for the examples. *Core* is derived with the condition that the table still needs to be consistent (Hossain et al. 2021; Hossain et al., 2–5, December 2018). Then certain rules are generated from the consistent portion of the information table. In certain rules generation, lower approximation set is taken into account. Which means, the inconsistent objects are ignored to generate only the definite rules.

Once the certain rules are generated, the possible rules can also be generated by considering upper approximation. Which means the inconsistent objects are also taken into account for generating the rules.

Experimental Steps

Data Collection

A collection of 2747 samples and 13 well log features are considered where ten major lithology classeslithology class found in the well namely Mudstone, Claystone, Sandstone, Siltston, Sandy Siltstone, Sandy Mudstone, Silty Sandstone, Muddy Sandstone, Silty Mudstone, and Granulestone are the categories to be classified by the methodology.

Table 1 shows the information of the lithology classeslithology class and the assigned class number.

Discretization

There are 12 well log conditional features such as Gamma Ray, Neutron Porosity, Density Correlation, Photo Electric Effect, Density Porosity, Conductivity, Caliper, Borehole Volume, Compressional Sonic, Hole Diameter, SQp, and SQs.

For applying the methodology, the above mentioned conditional features need to be discretized. In this experiment, each of these conditional features is discretized into 10 equal length intervals.

Upper and Lower Approximations

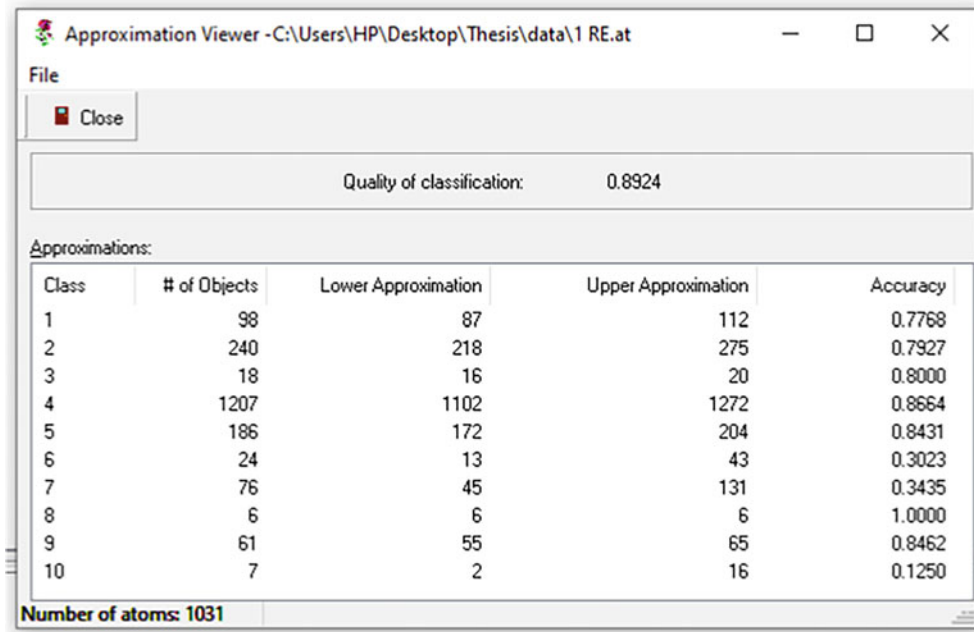
Based on the RST theory, for the training samples belonging to each particular class, the lower and the upper approximations are found.

The quality and accuracy of approximation can be computed with the RST approximations. The accuracy of approximations lies between $[0,1]$.

The accuracy of approximation is defined as

Upper Approximation, Table 1 Lithology classeslithology class and their corresponding information

Lithology	Class no.
Claystone	1
Mudstone	2
Siltston	3
Sandstone	4
Sandy mudstone	5
Sandy siltstone	6
Silty sandstone	7
Silty mudstone	8
Muddy sandstone	9
Granulestone	10



Approximation Viewer - C:\Users\HP\Desktop\Thesis\data\1 RE.at

File

Close

Quality of classification: 0.8924

Approximations:

Class	# of Objects	Lower Approximation	Upper Approximation	Accuracy
1	98	87	112	0.7768
2	240	218	275	0.7927
3	18	16	20	0.8000
4	1207	1102	1272	0.8664
5	186	172	204	0.8431
6	24	13	43	0.3023
7	76	45	131	0.3435
8	6	6	6	1.0000
9	61	55	65	0.8462
10	7	2	16	0.1250

Number of atoms: 1031

Upper Approximation, Fig. 2 RST approximation accuracy. (Note this is based on Rose 2 <http://idss.cs.put.poznan.pl/site/rose.html>)

```
rule 47. (DRHO = 8) & (PE = 6) & (DPHI = 2) => (Lithology = 8); [2, 2, 33.33%, 100.00%][0, 0, 0, 0, 0, 0, 2, 0, 0]
[0, 0, 0, 0, 0, 0, 0, 0]
rule 48. (NPFI = 2) & (DRHO = 7) & (PE = 7) => (Lithology = 8); [1, 1, 16.67%, 100.00%][0, 0, 0, 0, 0, 0, 1, 0, 0]
[0, 0, 0, 0, 0, 0, 0, 0]
rule 49. (DRHO = 9) & (CALI = 4) & (SQP = 3) => (Lithology = 8); [1, 1, 16.67%, 100.00%][0, 0, 0, 0, 0, 0, 1, 0, 0]
[0, 0, 0, 0, 0, 0, 0, 0]
rule 50. (NPFI = 3) & (DPHI = 7) => (Lithology = 10); [1, 1, 14.29%, 100.00%][0, 0, 0, 0, 0, 0, 0, 0, 1]
[0, 0, 0, 0, 0, 0, 0, 0]
rule 51. (DPHI = 5) & (CALI = 6) => (Lithology = 10); [1, 1, 14.29%, 100.00%][0, 0, 0, 0, 0, 0, 0, 0, 1]
[0, 0, 0, 0, 0, 0, 0, 0]
# Approximate rules
rule 52. (GR = 3) & (NPFI = 3) & (CT = 3) & (BHVT = 7) & (SQP = 4) => (Lithology = 4) OR (Lithology = 7); [14, 14, 22.22%, 100.00%][0, 0, 0, 9, 0, 0, 5, 0, 0]
[0, 0, 0, 55, 477, 764, 1186, 1271, 1340, 1445, 1732, 1847], [0, 0], [1, 314, 443, 1015, 1488], [0, 0, 0]
rule 53. (GR = 4) & (FE = 4) & (CT = 4) & (DTC = 8) => (Lithology = 4) OR (Lithology = 7); [12, 12, 19.05%, 100.00%][0, 0, 0, 8, 0, 0, 4, 0, 0]
[0, 0, 0, 558, 701, 1180, 1356, 1403, 1426, 1504, 1618], [0, 0], [35, 426, 861, 1239], [0, 0, 0]
rule 54. (GR = 4) & (DRHO = 4) & (SQP = 4) => (Lithology = 4) OR (Lithology = 5) OR (Lithology = 7); [6, 6, 100.00%, 100.00%][0, 0, 0, 2, 1, 0, 3, 0, 0]
[0, 0, 0, 814, 1045], [827], [0], [36, 1033, 1476], [0, 0, 0]
```

Upper Approximation, Fig. 3 Approximation rules. (Note this is based on Rose 2 <http://idss.cs.put.poznan.pl/site/rose.html>)

$$\text{Approximation accuracy} = \frac{\text{No. of objects in lower approximation}}{\text{No. of objects in upper approximation}} \quad (12)$$

The figure below shows the sample sizes for upper and lower approximations for each class.

Certain and Uncertain Rules Generation

1. In Fig. 2, we can see that, other than class no. 8, the approximation accuracy is less than 1.

2. Since, for class no. 8, lower approximation = upper approximation, only certain rules can be generated from this class and no possible rules are needed.
3. For all the classes except class no. 8, both certain and possible rules can be generated.
4. Sample Certain and Possible rules are shown in Fig. 3. Rules no. 47–51 are certain rules. And Rules No. 52–54 are possible rules.

Classification Using the RST Rules

The rules that are found in section “[Certain and Uncertain Rules Generation](#)” are implemented on the test dataset for classifying the objects and computing the accuracy.

1. First, the certain rules are applied on the testing samples.
2. Then the possible rules are applied to the samples that could not be classified by the certain rules.
3. The classification accuracy is 84.83% and the misclassification rate is 15.17%.
4. The computation of minimizing the decision rules is a combinatorial computation. It is time-consuming. Kim et al. (2013) and Kim et al. (2011) enable us to solve the fast solution of RST by using DNA computation.

Summary

In this entry, a classification method based on upper and lower approximations for generating possible and certain rules is shown and a real time application is also provided to solve the lithology identification problem. The conclusions are as follows:

1. By using the RST approximations, certain and possible rules can be generated and data inconsistency is handled.
2. The generated rules are usable for classifying the decision classes for the future objects.
3. The classification accuracy is 84.83%. Therefore, the model can be adapted as a real-world solution to lithology classification.
4. The rules provide explainability to the model which helps the researchers to extract the reasoning for the classifications of the objects.

Bibliography

- Grzymala-Busse JW (1992) LERS a system for learning from examples based on rough sets. In: Slowinski R (ed) *Intelligent decision support handbook of applications and advances of the rough sets theory*. Kluwer Academic, p 318
- Hossain TM, Watada J, Hermana M, Sakai H (2018) A rough set based rule induction approach to geoscience data. In: *International conference of unconventional modelling, simulation optimization on soft computing and meta heuristics, UMSO 2018, Kitakyushu, Japan, 2–5, December 2018*. IEEE. <https://doi.org/10.1109/UMSO.2018.8637237>
- Hossain TM, Watada J, Aziz IA, Hermana M (2020a) Machine learning in electrofacies classification and subsurface lithology interpretation: a rough set theory approach. *Appl Sci* 10:5940
- Hossain TM, Watada J, Hermana M, Aziz IA (2020b) Supervised machine learning in electrofacies classification: A rough set theory approach. *J Phys Conf Ser* 1529:052048
- Hossain TM, Watada J, Hermana M, Meraj ST, Sakai H (2021) Lithology prediction using well logs: a granular computing approach. *Int J Innov Comput Inf Control* 17:225. Accepted on August, 22, 2020
- Kim I, Chu Y-Y, Watada J, Wu J-Y, Pedrycz W (2011) A DNA-based algorithm for minimizing decision rules: A rough sets approach. *IEEE Trans NanoBioscience* 10(3):139–151. <https://doi.org/10.1109/TNB.2011.2168535>
- Kim I, Watada J, Pedrycz W (2013) DNA rough-set computing in the development of decision rule reducts. In: Skowron A, Suraj Z (eds) *Rough sets and intelligent systems: Professor Zdzisław Pawlak in Memoriam, volume 1 of Intelligent systems reference library*. Springer, pp 409–438. https://doi.org/10.1007/978-3-642-30344-9_15
- Pawlak Z (1982) Rough sets. *Int J Parallel Prog* 11(5):341–356. <https://doi.org/10.1007/BF01001956>
- Pawlak ZI (2002) Rough sets and intelligent data analysis. *Inf Sci* 147: 1–12
- Sudha M, Kumaravel A (2018) Quality of classification with lers system in the data size context. *Appl Comput Inf* 16(1/2):29–38
- Zhang Q, Xie Q, Wang G (2016) A survey on rough set theory and its applications. *CAAI Trans Intell Technol*. <https://doi.org/10.1016/j.trit.2016.11.001>

Variance

Claudia Cappello¹ and Monica Palma²

¹Dipartimento di Scienze dell'Economia, Università del Salento, Lecce, Italy

²Università del Salento-Dip. Scienze dell'Economia, Complesso Ecotekne, Lecce, Italy

Synonyms

Second-order moment; Standard deviation

Definition

The variance of a regionalized variable measures the dispersion around the expected value and it is defined as the expected value of the squared difference between the random variable and its expected value.

Overview

In spatial statistics, the observed value $z(\mathbf{x})$ of a given aspect of interest at the location $\mathbf{x}$ is considered as a particular realization of the corresponding random variable $Z(\mathbf{x})$, and the set of the random variables $\{Z(\mathbf{x}), \mathbf{x} \in V\}$, where $V \subseteq \mathbb{R}^d$, $d \in \mathbb{N}_+$ ($d \leq 3$), is called spatial random field. It is important to point out that in the applications, the sample data could be measured on points or blocks. If the spatial data refer to a block both its location and its volume (in terms of shape and size, which define the support of the sample) have to be considered. Moreover, if the volume of the blocks is very small and equal for all the sample data, it can be neglected and the sample data can be treated as points (Isaaks and Srivastava, 1989).

In the following, the first- and second-order moments of $Z(\cdot)$ are provided and the central role of the variance is highlighted in the geostatistical context.

Given a random variable $Z(\mathbf{x})$ at the point $\mathbf{x}$, its *first-order moment* depends on the spatial location $\mathbf{x}$ and it is defined as follows:

$$E[Z(\mathbf{x})] = m(\mathbf{x}), \quad \mathbf{x} \in V,$$

while the *second-order moments* are given below:

- *Variance*, also known as a priori *variance*, which measures the dispersion of $Z(\mathbf{x})$ around its expected value $m(\mathbf{x})$

$$\text{Var}[Z(\mathbf{x})] = E[Z(\mathbf{x}) - m(\mathbf{x})]^2, \quad \mathbf{x} \in V. \quad (1)$$

The variance can also be written as $\text{Var}[Z(\mathbf{x})] = E[Z^2(\mathbf{x})] - [m(\mathbf{x})]^2$; thus, when the expectation of $Z(\mathbf{x})$ is zero, the variance corresponds to $E[Z^2(\mathbf{x})]$.

- *Covariance*, which is defined as follows:

$$C[Z(\mathbf{x}), Z(\mathbf{x}')] = E\{[Z(\mathbf{x}) - m(\mathbf{x})][Z(\mathbf{x}') - m(\mathbf{x}')]\}, \quad \mathbf{x}, \mathbf{x}' \in V; \quad (2)$$

it exists if the variance of the two random variables $Z(\mathbf{x})$ and $Z(\mathbf{x}')$ is finite.

- *Variogram*, which measures the variance of the increments between $Z(\mathbf{x})$ and $Z(\mathbf{x}')$

$$2\gamma[Z(\mathbf{x}), Z(\mathbf{x}')] = \text{Var}[Z(\mathbf{x}) - Z(\mathbf{x}')] \quad \mathbf{x}, \mathbf{x}' \in V. \quad (3)$$

Note that the function γ is called *semivariogram*.

Looking at the second-order moments of a regionalized variable, the variance measures the spread of the phenomenon around the mean, and is a measure of the dissimilarity between $Z(\mathbf{x})$ and $Z(\mathbf{x}')$.

The covariance (2) and the variogram (3) depend on the support points $\mathbf{x}$ and $\mathbf{x}'$, hence many realizations of the random variables at the points $\mathbf{x}$ and $\mathbf{x}'$ are necessary to make statistical inference on these moments. However, the spatial data are non-repetitive, since only one observation is available in every single location. To overcome this inferential problem, the stationarity hypotheses have to be introduced, which assume that the observations at the spatial points characterized by the same separation vector $\mathbf{h}$ (in length and direction) are considered as replicates of the pair $Z(\mathbf{x})$ and $Z(\mathbf{x}')$. In particular, a random field is second-order stationary if

- The expected value does not depend on the spatial point $\mathbf{x}$: $E[Z(\mathbf{x})] = m, \mathbf{x} \in V$.
- The covariance only relies on the separation vector $\mathbf{h}$ and not on the spatial locations $\mathbf{x}$ and $\mathbf{x}'$:

$$C(\mathbf{h}) = E\{[Z(\mathbf{x}) - m][Z(\mathbf{x} + \mathbf{h}) - m]\}, \quad \mathbf{x}, \mathbf{x} + \mathbf{h} \in V;$$

where $\mathbf{h} = (|\mathbf{h}|, \alpha)$ is the separation vector, whose elements are $|\mathbf{h}|$, which is the length of the vector and α , which is the corresponding direction.

The stationarity of the covariance implies the stationarity of the variance and of the variogram (Journel and Huijbregts 1978), as specified hereinafter.

- For $\mathbf{h} = \mathbf{0}$ the covariance is equal to the variance

$$\begin{aligned} C(\mathbf{0}) &= E\{[Z(\mathbf{x}) - m][Z(\mathbf{x}) - m]\} = E[Z(\mathbf{x}) - m]^2 \\ &= \text{Var}[Z(\mathbf{x})], \quad \mathbf{x} \in V; \end{aligned} \quad (4)$$

hence, all the random variables over the same spatial domain have the same expectation and the same finite variance.

- For any couple of random variables $Z(\mathbf{x})$ and $Z(\mathbf{x} + \mathbf{h})$ the variogram is equal to

$$\begin{aligned} 2\gamma(\mathbf{h}) &= \text{Var}[Z(\mathbf{x} + \mathbf{h}) - Z(\mathbf{x})] \\ &= E[Z(\mathbf{x} + \mathbf{h}) - Z(\mathbf{x})]^2, \quad \mathbf{x}, \mathbf{x} + \mathbf{h} \in V. \end{aligned} \quad (5)$$

From (5) it can be proved that if two random variables are separated by a vector $\mathbf{h}$, then the semivariogram γ is strictly related to the variance and the covariance through the following relation

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}), \quad (6)$$

where $C(\mathbf{0}) = \text{Var}[Z(\mathbf{x})]$. In addition, under the second-order stationarity hypothesis, it is assumed the existence of the covariance and, consequently, of a finite a priori

variance. Moreover, the covariance is a bounded function, namely, $|C(\mathbf{h})| \leq C(\mathbf{0})$, then, according to (6), also the variogram is bounded by $C(\mathbf{0})$. However, for some phenomena the variance and the covariance do not exist in finite form; in this case it is convenient to resort to the intrinsic hypotheses, which are based on the second-order stationarity of the increments between $Z(\mathbf{x} + \mathbf{h})$ and $Z(\mathbf{x})$. In particular, given a separation vector $\mathbf{h}$:

- The expected value of the increments is supposed to be equal to zero:

$$E[Z(\mathbf{x} + \mathbf{h}) - Z(\mathbf{x})] = 0, \quad \mathbf{x}, \mathbf{x} + \mathbf{h} \in V;$$

- The variance of the increments $[Z(\mathbf{x} + \mathbf{h}) - Z(\mathbf{x})]$ is finite and does not depend on $\mathbf{x}$: $2\gamma(\mathbf{h}) = \text{Var}[Z(\mathbf{x} + \mathbf{h}) - Z(\mathbf{x})] = E[Z(\mathbf{x} + \mathbf{h}) - Z(\mathbf{x})]^2, \mathbf{x}, \mathbf{x} + \mathbf{h} \in V$.

Variance for an Indicator Random Field

In some applications it is interesting to evaluate the probability whether or not a random variable at unsampled locations exceeds a fixed threshold. For this aim it is relevant introducing the indicator random field and the corresponding variance.

Given a second-order stationary random field $\{Z(\mathbf{x}), \mathbf{x} \in V\}$ where $V \subseteq \mathbb{R}^d$ and a fixed threshold $z \in \mathbb{R}$, the indicator random field $\{I(\mathbf{x}; z), \mathbf{x} \in V\}$ is such that

$$I(\mathbf{x}; z) = \begin{cases} 1 & \text{if } Z(\mathbf{x}) \leq z, \\ 0 & \text{otherwise.} \end{cases}$$

The variance of an indicator random field is finite and equal to

$$\text{Var}[I(\mathbf{x}; z)] = F(z) - F^2(z), \quad (7)$$

where $F(\cdot)$ denotes the marginal distribution of $Z(\mathbf{x})$.

The remarkable characteristic of the variance of an indicator random field is that its maximum value is 0.25 for $z = z_M$, where z_M is the median of the distribution of $F(z)$, hence $F(z_M) = 0.5$.

Properties of the Variance

Given a random variable $Z(\mathbf{x})$ at the spatial location $\mathbf{x}$, the following properties of the variance are provided:

1. $\text{Var}[Z(\mathbf{x})] \geq 0$,
2. $\forall a \in \mathbb{R}, \quad \text{Var}(a) = 0$,
3. $\forall a \in \mathbb{R}, \quad \text{Var}[Z(\mathbf{x}) + a] = \text{Var}[Z(\mathbf{x})]$,
4. $\forall a \in \mathbb{R}, \quad \text{Var}[aZ(\mathbf{x})] = a^2 \text{Var}[Z(\mathbf{x})]$.

Moreover, given the regionalized variables $Z(\mathbf{x}_i)$, $i = 1, \dots, N$, the variance of a sum of random variables results as follows:

$$\text{Var} \left[\sum_{i=1}^N Z(\mathbf{x}_i) \right] = \sum_{i=1}^N \text{Var}[Z(\mathbf{x}_i)] + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N C[Z(\mathbf{x}_i), Z(\mathbf{x}_j)]$$

where $C(\cdot)$ is the covariance function. Starting from the first property of the variance, the admissibility condition for a function to be a covariogram (i.e. the condition of positive definiteness) can be derived.

The variance of the vector $\mathbf{Z} = [Z(\mathbf{x}_1), Z(\mathbf{x}_2), \dots, Z(\mathbf{x}_N)]$ can also be written in matrix form, which is known in the literature as variance-covariance matrix, usually indicated with the Greek letter Σ . This matrix, of order $(N \times N)$, contains the variances of the regionalized variables $Z(\mathbf{x}_i)$, $i = 1, \dots, N$ on the diagonal, while the off-diagonal entries are the covariances between the variables, i.e.,

$$\Sigma = \begin{bmatrix} \text{Var}[Z(\mathbf{x}_1)] & C[Z(\mathbf{x}_1), Z(\mathbf{x}_2)] & \cdots & C[Z(\mathbf{x}_1), Z(\mathbf{x}_N)] \\ C[Z(\mathbf{x}_2), Z(\mathbf{x}_1)] & \text{Var}[Z(\mathbf{x}_2)] & \cdots & C[Z(\mathbf{x}_2), Z(\mathbf{x}_N)] \\ \vdots & \vdots & \ddots & \vdots \\ C[Z(\mathbf{x}_N), Z(\mathbf{x}_1)] & C[Z(\mathbf{x}_N), Z(\mathbf{x}_2)] & \cdots & \text{Var}[Z(\mathbf{x}_N)] \end{bmatrix}.$$

Note that the Σ matrix is symmetric and positive semi-definite (Schott 2018).

Sample Variance

As previously pointed out, the a priori variance of a second-order stationary random field is finite and corresponds to $C(\mathbf{0})$, on the other hand if Z satisfies the intrinsic hypotheses but not the second-order stationarity ones, then $\text{Var}[Z(\mathbf{x})]$ might not be finite, however, and the variance of the increments between $Z(\mathbf{x} + \mathbf{h})$ and $Z(\mathbf{x})$ is finite.

Given the random variables $Z(\mathbf{x}_i)$, $i = 1, \dots, N$, of a random function Z , the estimator of the variance is the *sample variance*, whose analytic expression is given below:

$$S^2 = \frac{1}{N} \sum_{i=1}^N [Z(\mathbf{x}_i) - \bar{Z}]^2, \quad \mathbf{x}_i \in V, \quad (8)$$

where $\bar{Z} = \frac{1}{N} \sum_{i=1}^N Z(\mathbf{x}_i)$ is the sample mean.

Alternatively, the sample variance can also be expressed in terms of the sum of the squared differences between all the random variables $Z(\mathbf{x}_i)$

$$S^2 = \frac{1}{2N^2} \sum_{i=1}^N \sum_{j=1}^N [Z(\mathbf{x}_i) - Z(\mathbf{x}_j)]^2, \quad \mathbf{x}_i, \mathbf{x}_j \in V. \quad (9)$$

It is important to point out that the sample variance is a random variable; hence it is characterized by first- and second-order moments. In particular, the expected value of the sample variance, denoted by σ^2 , is obtained as follows (Chilès and Delfiner 2012):

$$E(S^2) = \sigma^2 = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \gamma(\mathbf{x}_i - \mathbf{x}_j)^2. \quad (10)$$

Moreover, for second-order stationary random functions, the expected value of the sample variance (10) is also equivalent to

$$E(S^2) = \sigma^2 = C(\mathbf{0}) - \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N C(\mathbf{x}_i - \mathbf{x}_j). \quad (11)$$

Hence, starting from the equivalence in (11) the following relation can be derived: $\sigma^2 < C(\mathbf{0})$. In addition, if the sample locations are so far apart, which means that the regionalized variables are uncorrelated, the expectation of the sample variance is equal to

$$E(S^2) = \sigma^2 = \left(1 - \frac{1}{N}\right) C(\mathbf{0}), \quad (12)$$

where the difference between σ^2 and $C(\mathbf{0})$ vanishes as the sample size increases.

Estimation Variance

It is well known that in Geostatistics the estimation of a regionalized variable at unsampled spatial locations, obtained through different techniques, is a relevant aspect. The estimation error occurs in every estimation technique due to the fact that the quantity to be estimated generally differs from the estimate.

Given a random function Z over a spatial domain V , Z takes its values $Z(\mathbf{x})$ on small regular-shaped areas v centered on $\mathbf{x}$, and if the support v is a point its area is negligible with respect to V . Assuming that the true value $z(\mathbf{x}_i)$ is estimated through $\hat{z}(\mathbf{x}_i)$, the estimation error is equal to $r(\mathbf{x}_i) = \hat{z}(\mathbf{x}_i) - z(\mathbf{x}_i)$. Then, the error $r(\mathbf{x}_i)$ is a particular realization of the random variable $R(\mathbf{x}_i) = \hat{Z}(\mathbf{x}_i) - Z(\mathbf{x}_i)$, $i = 1, \dots, N$. Note that if the random field Z is stationary, also $R(\mathbf{x})$ is stationary; moreover

since $R(\mathbf{x})$ is a random variable it is relevant to evaluate its first- and second-order moments (Journel and Huijbregts 1978) that are:

- the expected value $E[R(\mathbf{x})] = m_E, \forall \mathbf{x}$;
- the variance (or *estimation variance*) $Var[R(\mathbf{x})] = E[R(\mathbf{x})]^2 - m_E^2 = \sigma_E^2, \forall \mathbf{x}$

The estimation of the above-mentioned moments is required since a good estimation technique ensures a mean error equal to zero and the smallest estimation variance. In particular, the estimator $\hat{Z}$ must be a function of $Z(\mathbf{x}_i)$

$$\hat{Z}(\mathbf{x}) = f\{Z(\mathbf{x}_1), Z(\mathbf{x}_2), \dots, Z(\mathbf{x}_N)\}$$

and it is such that:

- The unbiasedness condition is satisfied, i.e., $E[R(\mathbf{x})] = E[\hat{Z}(\mathbf{x}) - Z(\mathbf{x})] = 0$.
- It is possible to compute easily the estimation variance, i.e., $Var[R(\mathbf{x})] = \sigma_E^2 = Var[\hat{Z}(\mathbf{x}) - Z(\mathbf{x})]$

For any function $f(\cdot)$, the calculation of the expected value and of the variance of the random variable $R(\cdot)$ requires that the distribution of $Z(\mathbf{x}_i)$, $i = 1, \dots, N$, is known. Nevertheless, since it is generally not possible to infer this distribution from a unique realization of $Z(\mathbf{x})$, usually the class of the linear $\hat{Z}$ is chosen in estimators (Journel and Huijbregts 1978):

$$\hat{Z} = \sum_{i=1}^N \lambda_i Z(\mathbf{x}_i). \quad (13)$$

Hence, given a stationary random field Z , if $\hat{Z}$ is chosen in the class of the linear estimators the unbiasedness condition is ensured if $\sum_{i=1}^N \lambda_i = 1$ and the estimation variance (Matheron 1963, 1965) corresponds to

$$\sigma_E^2 = 2\bar{\gamma}(v, V) - \bar{\gamma}(V, V) - \bar{\gamma}(v, v), \quad (14)$$

or, more specifically can be calculated as follows:

$$\sigma_E^2 = 2 \sum_{i=1}^N \lambda_i \bar{\gamma}(\mathbf{x}_i, V) - \bar{\gamma}(V, V) - \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j \gamma(\mathbf{x}_i, \mathbf{x}_j), \quad (15)$$

where

- $\bar{\gamma}(\mathbf{x}_i, V) = \frac{1}{|V|} \int_V \gamma(\mathbf{x}_i, \mathbf{x}) d\mathbf{x}$, is the average semivariogram between the sampling points $\mathbf{x}_i$ and all the other points in the domain V ;
- $\bar{\gamma}(V, V) = \frac{1}{|V|^2} \int_V \int_V \gamma(\mathbf{x}, \mathbf{x}') d\mathbf{x} d\mathbf{x}'$, is the average semi-variogram over V

From (14) or (15) it is possible to conclude that σ_E^2 does not depend on the shape and size of the area under study, the relative locations of the sample data within V as well as the spatial correlation structure γ .

Hence, σ_E^2 is strictly related to the model chosen for modeling γ , keeping constant the area V and the location of the sample data.

Note that, by minimizing the estimation variance σ_E^2 , the Ordinary Kriging (OK) estimator is obtained (Stein, 1999). Hence, in the class of the linear weighted average estimators, the kriging provides the best linear unbiased estimator (B.L. U.E.), since the optimal weights λ_N minimize the estimation variance σ_E^2 , under the constraint $\sum_{i=1}^N \lambda_i = 1$. It is well known that different forms of kriging are available in the literature and each type is characterized by a specific estimation variance. In the contributions of Journel and Huijbregts (1978) and Journel (1977) the hierarchy among the estimation variances is detailed, that is $\sigma_{E_{DK}}^2 \leq \sigma_{E_{SK}}^2 \leq \sigma_{E_{OK}}^2 \leq \sigma_{E_{UK}}^2$, where $\sigma_{E_{DK}}^2$ is the estimation variance associated to the disjunctive kriging, $\sigma_{E_{SK}}^2$ is the one associated with the simple kriging, while $\sigma_{E_{OK}}^2$ and $\sigma_{E_{UK}}^2$ refer to the estimation variance in case of ordinary and universal kriging, respectively.

Dispersion Variance

The term *dispersion variance*, as clarified in Journel and Huijbregts (1978), is used in Geostatistics to denote:

- The dispersion around the mean value of a set of data collected within a domain V . This dispersion increases as the dimension of the spatial domain increases.
- The dispersion within a fixed domain V with respect to the support v . This dispersion decreases as the support v on which each datum is defined increases.

The dispersion variance of the support v within the spatial domain V also depends on the variogram:

$$D^2(v|V) = \bar{\gamma}(V, V) - \bar{\gamma}(v, v). \quad (16)$$

From the relationship in (16) it is evident that as the support v increases, then $\bar{\gamma}(v, v)$ increases, hence the dispersion variance $D^2(v|V)$ decreases, keeping other factors constant. For the extreme case in which the spatial unit is a point ($v = 0$), the dispersion variance in terms of variogram, corresponds to

$$D^2(0|V) = \bar{\gamma}(V, V). \quad (17)$$

It is important to note that the dispersion variances in (16) and (17) are linked by the additive relationship that is:

$$D^2(0|V) = D^2(0|v) + D^2(v|V).$$

This property explains the drop of the variance when the support changes from a quasi point sample support to a larger support of interest:

$$D^2(0|V) - D^2(v|V) = D^2(0|v) = \bar{\gamma}(v, v).$$

Summary

In spatial statistics the second-order moments of a random variable $Z(\mathbf{x})$ at the location $\mathbf{x}$ are always evaluated. In particular, the variance measures the dispersion of the random variable around its expected value.

In this chapter, the first- and second-order moments of $Z(\cdot)$ are provided and the central role of the variance is highlighted in the geostatistical context.

Cross-References

- [Kriging](#)
- [Stationarity](#)
- [Variogram](#)

References

- Chilès JP, Delfiner P (2012) *Geostatistics: modeling spatial uncertainty*, 2nd edn. Wiley, Hoboken. 734 pp
- Isaaks EH, Srivastava RM (1989) *An introduction to applied geostatistics*. Oxford University Press, New York. 561 pp
- Journel AG (1977) Kriging in terms of projections. *J Int Assoc Math Geol* 9:563–586
- Journel AG, Huijbregts CJ (1978) *Mining geostatistics*. Academic, London/New York. 600 pp
- Matheron G (1963) Principles of geostatistics. *Econ Geol* 58(8): 1246–1266
- Matheron G (1965) Les variables rgionalises et leur estimation: Une application de la thorie des fonctions alatoires aux sciences de la nature. Masson, Paris
- Schott JR (2018) *Matrix analysis for statistics*, 3rd edn. Wiley, Hoboken. 552 pp
- Stein ML (1999) *Interpolation of spatial data: some theory for kriging*. Series in statistics. Springer, New York. 247 pp

Variogram

Behnam Sadeghi

EarthByte Group, School of Geosciences, University of Sydney, Sydney, NSW, Australia

Earth and Sustainability Research Centre, University of New South Wales, Sydney, NSW, Australia

Definition

A variogram is a tool to describe the data spatial continuity. When a number of data samples are available, the analyst can use a variability measure between each pair of the samples at different distances to calculate the experimental variogram function considering the best type of the variogram fitted (Deutsch and Journel 1998).

Introduction

Geostatistics is the combination of different disciplines of geology and statistics. Its main focus is on spatial and spatiotemporal types of dataset in various fields of geosciences and engineering. One of the significant and basic concepts in geostatistics is variogram, its different types, and the variography. A variogram is a function to describe the data spatial continuity using the variability between sample points located at different distances from each other and affecting each other considering the distances (lags). In actual, it helps with finding the highest distance between the samples in which the samples are spatially in relation and affecting each other. In other words, closer samples have lower variabilities, i.e., higher spatial effects, yet the distant samples provide higher variabilities, i.e., lower spatial effects. The final interpolation would be implemented on those samples which are situated in that effective range.

Samples and Datasets

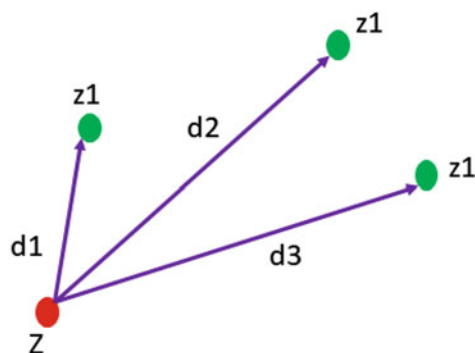
In the geosciences, we may have various types of samples in 2D (on the ground) or 3D (under the ground and subsurface), such as litho, soil, till, rock chip, and stream sediment samples. Based on the study for which the samples are collected, one looks for something specific as the target value, i.e., data. For instance, in geochemical studies, that value could be the concentration of the target elements in various samples.

Interpolation

In general, mostly in geological studies, due to lack of access to some parts of the study areas like mountainous areas, or even lack of budget, which is always critical, the number of the samples collected is limited. Based on the experts' opinion and the exploration level and scale (e.g., reconnaissance, regional, local, and detailed exploration), the sampling network could be regular or irregular and dense or scarce, and even the distance between samples could be short in meters or higher than kilometers. To generate a final continuous map based on the sample values to provide a better imagination about the mineral deposits, their concentrations, and trend, we need to estimate the values in the unsampled areas as well, however based on those in sampled areas. To do so, we need to apply interpolation methods.

A group of multivariate interpolation methods are applicable in geostatistical analysis considering the type of samples and studies. Some of the more conventional methods are bilinear interpolation and nearest-neighbor interpolation, which are mainly based on the distance between samples.

Generally, in geochemical surface map generation procedures and exploration projects, *weighted moving average* (WMA) interpolation techniques have been significantly taken into consideration. Inverse distance weighting (IDW) and kriging are among the most common WMA techniques (Carranza 2009). The IDW interpolation technique considers both distances between samples and the spatial effect of samples on each other considering that distance (assigned as weight). Simply, this method says the closer samples have higher effects (assigned weights) on each other than the distant samples (Fig. 1). This interpolation technique is mostly applicable to 2D modeling/mapping; however, in 3D modeling, it has some limitations, which are out of this entry's scope. To apply this technique, we need to have a knowledge of the point uni-element concentration data. This helps with better adjustments of the "moving average" window (or kernel) size and distance (assigned as weight) parameters (Carranza 2009).



$$Z = \frac{\sum_{i=1}^n \left(\frac{z_i}{d_i^p} \right)}{\sum_{i=1}^n \left(\frac{1}{d_i^p} \right)}$$

Z: target element concentration

z_i : concentration in i th sample

d: distance

n: number of samples

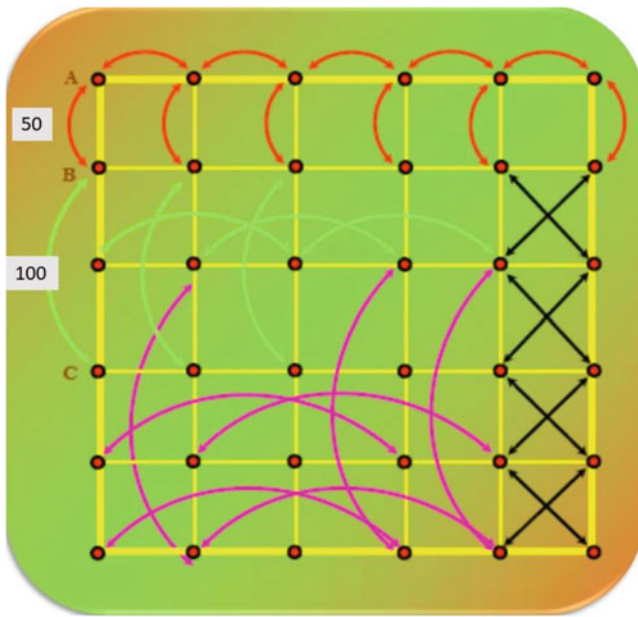
* The optimum value of p is often 2.

Variogram, Fig. 1 The simplified concept of IDW

The main point raised here is we may have a local group of samples together, in addition to several other samples in various distances from them. Can we use all those samples for the interpolation? Does it make the final interpolation biased? Do those samples really have any spatial effect on the main group of samples? If so, how much? If not, considering the samples' values and their variances, what samples could effectively get involved with the estimation and what samples will not have any spatial effect? This means we need to have a distance range in which all samples are in spatial relationship, and they are proper for interpolation, and the rest of the samples beyond that range have no spatial effect and association and will not be used for the interpolation (Fig. 2). To do so, we need a new concept or interpolation method. Kriging interpolation/estimation was developed based on such a search radius (Sagar et al. 2018; see the entry "Kriging"). Kriging, as a Gaussian process regression, is the most important interpolation method in geostatistics that provides the best linear unbiased estimation (BLUE) based on covariances (Deutsch and Journel 1998). It applies Gauss-Markov theorem to demonstrate the estimation/error independence. The main difference between kriging and regression is that kriging is applied to an individual realization of a single random field, but the regression is multiple scenarios of multivariate dataset. To apply kriging interpolation (see the entry "Kriging"), or Gaussian simulation to generate more than a single scenario (i.e., realizations) based on the same data – see the entries "Simulation," "Sequential Gaussian Simulation," and "Uncertainty Quantification" – variogram is required.

Variogram Main Principles

In Fig. 2, imagine the distances between two samples of A and B, and B and C are 50 m and 100, respectively. The differences between the target element concentration values per samples are available, too. In this respect, the distances between the other samples, with defined concentration values,



Variogram, Fig. 2 Variogram calculation process

are also calculated. Having all these details, variogram is generated using the equation below (David 1982; Isaaks and Srivastava 1990; Journel and Huijbregts 2004; Chilès and Delfiner 2012; Pyrcz and Deutsch 2014):

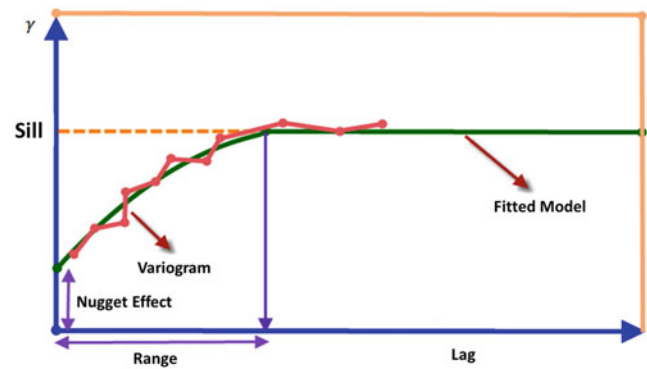
$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [(Z(x_i) - Z(x_{i+h}))^2]$$

where $N(h)$ is the number of sample pairs with h distance (lag distance) from each other. $Z(x_i)$ and $Z(x_{i+h})$ are also the variables with the same distance of h . $\gamma(h)$ represents semi-variograms, generally called variograms in literature although it is half of a variogram, in actuality.

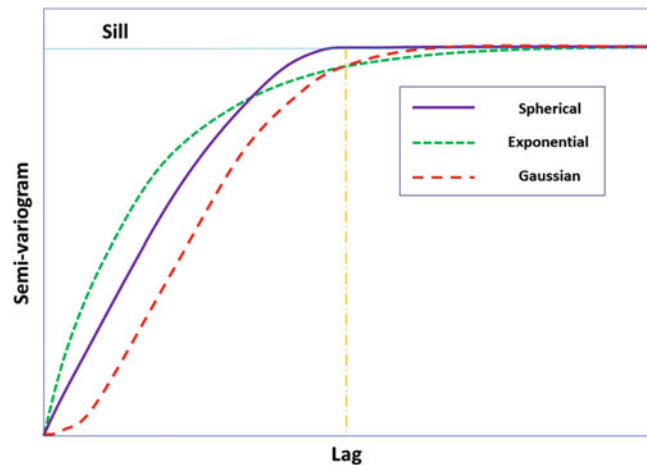
Variogram is generally a tool to evaluate the dissimilarity of a quantitative value, i.e., a local variable, between two samples with h distance. In most of the mineral deposits, such local variables are represented as element concentrations.

Figure 3 simply demonstrates the variogram's structure. In most of the mineralization and mineral deposits, variogram does not start from 0, i.e., the variance is not zero. Such a difference is called the “nugget effect” (Bohling 2005). In general, the nugget effect represents any discontinuities at the origin, even with different causes. It happens due to some reasons as the sources of variance such as (Chilès and Delfiner 2012):

- A structure, microstructure, or “geological noise”; it happens when the range is less than the distance between the samples in sampling network.
- Measurement or positioning errors.



Variogram, Fig. 3 Schematic variogram and its representative parts



Variogram, Fig. 4 Various main types of variograms

The variogram fitted to the points often inclines gently until it reaches the sill, i.e., variance. Such a distance is called “range,” which represents the maximum distance between the samples in which they spatially affect each other, i.e., their covariance is zero.

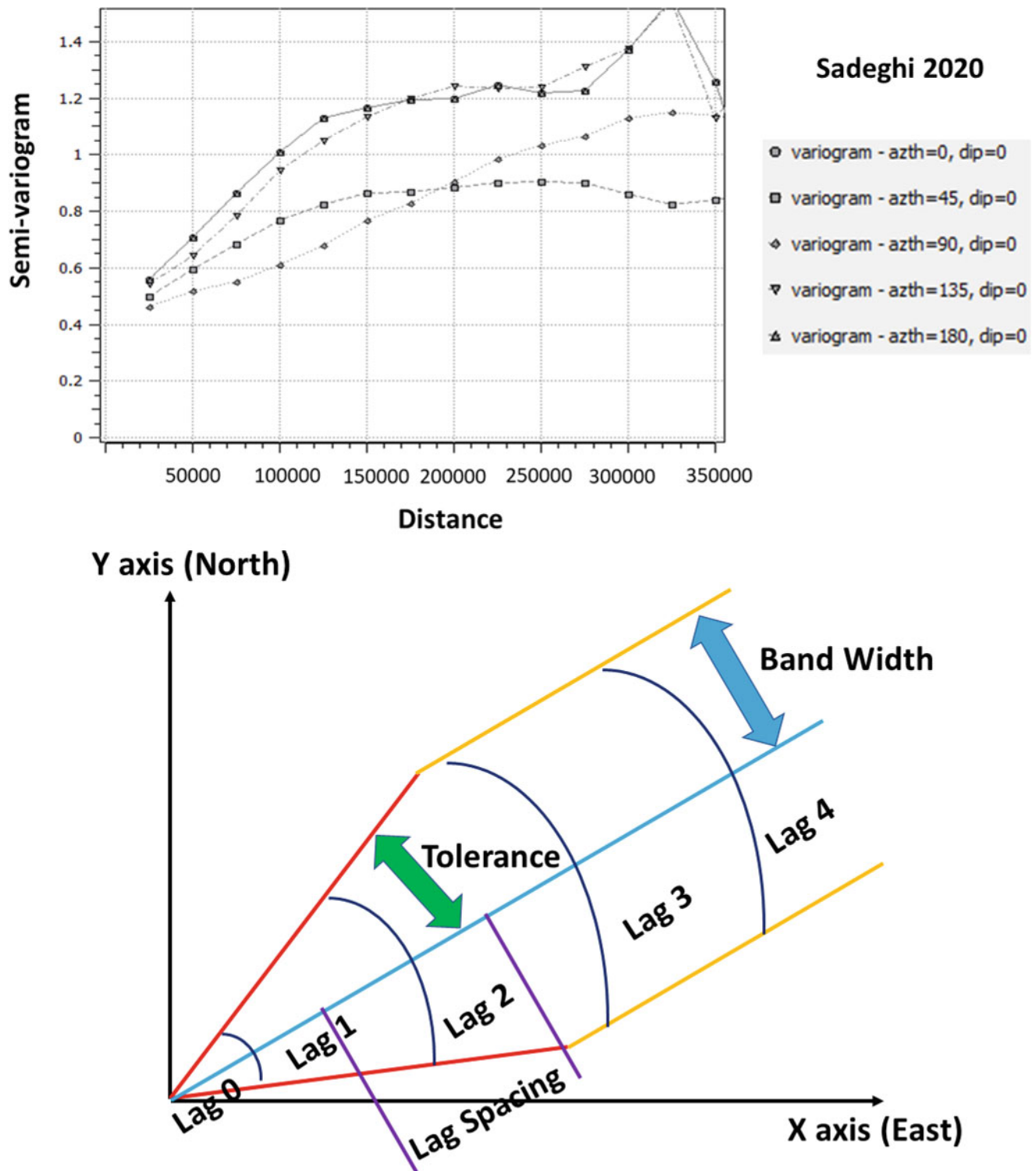
Main Types of Variogram

In order to evaluate variograms, we need to study the standard variograms rather than the experimental ones. The main standard variogram types are (Fig. 4) (Remy et al. 2009):

1. Spherical variogram (Matheron 1963):

$$\gamma(h) = \begin{cases} 0 & ; h = 0 \\ c \cdot \left(1.5 \left(\frac{\|h\|}{a} \right) - 0.5 \left(\frac{\|h\|}{a} \right)^3 \right) & ; \|h\| \leq a \\ c & ; h > a \end{cases}$$

where c represents the difference between the nugget effect and the sill, a is the range, and h is the lag. This equation is



Variogram, Fig. 5 Schematic explanation of the experimental variogram tolerance

applicable when the nugget effect is zero; however, in case it is not, we need to add its value to the equation. This variogram is the most useful variogram in geostatistics.

2. Exponential variogram:

$$\gamma(h) = c. \left(1 - \exp\left(\frac{-3\|h\|}{a}\right) \right)$$

3. Gaussian variogram:

$$\gamma(h) = c. \left(1 - \exp\left(\frac{-3\|h\|^2}{a^2}\right) \right)$$

The covariance equivalent of the above models is defined as (Remy et al. 2009):

$$C(h) = C(0) - \gamma(h), \text{ with } C(0) = 1$$

It is always advised to generate several variograms with different dip and azimuth values to define the optimum values of sill and range using different available scenarios (Deutsch and Journel 1998; Remy et al. 2009; Sadeghi 2020). It helps with the evaluation of the experimental variogram tolerance (Fig. 5). In other words, variogram is calculated per principal direction after any coordinate and data transformation step. Such directions are represented by azimuth and dip angles. The final variogram tolerance is demonstrated in Fig. 5.

Generally, in fitting any types of the abovementioned variograms, one condition should be taken into consideration (Goovaerts 1997; Remy et al. 2009):

$$\text{Nugget effect} + \text{sill} = 1$$

In case we face zonal mineralization, the variogram could have more than a single structure. In other words, it can have one nugget effect but several sills and ranges. Each structure could have its own individual type of variogram. However, again the sum of the nugget effect and all the sill values together should be 1 while fitting the variogram.

Summary and Conclusions

In projects such as mineral or petroleum exploration, the limited number of relevant geological samples on or under the ground has been always critical. This is because of some reasons such as lack of access to some areas, limited budget for further sampling and analysis, and of course limited time for further sampling and analysis. To study of the sample values as the representatives for sampled and unsampled areas, interpolation techniques such as IDW and various types of kriging have been developed and applied to generate estimated models and provide a better view of the

mineralization and spatial continuity and connectivity of the sample data values. In order to apply the kriging interpolation method and, through that, quantify the errors related and the probabilities required for spatial and stochastic uncertainty quantification, (semi-)variograms have been introduced and applied as the main tools. Using variograms and variography, the range that samples can affect each other spatially is calculated. The samples located in the same range will only be used for the interpolations. Such variograms can be studied in various azimuths and dips (e.g., 0, 45, 90, 135, and 180) to provide optimum values of the sill, the nugget effect, range, and other required statistics. Then, the interpolated models will be generated based on such calculated values. Such details are also applied to simulate more than one scenario (i.e., realization) based on the same datasets, to provide both Bayesian and frequency frameworks and calculate the stochastic and spatial uncertainties – see the entries “Uncertainty Quantification” and “Simulation.”

Cross-References

- [Kriging](#)
- [Sequential Gaussian Simulation](#)
- [Simulation](#)
- [Uncertainty Quantification](#)

Bibliography

- Bohling G (2005) Introduction to geostatistics and Variogram analysis. Kansas Geological Survey, 20 p
- Carranza EJM (2009) Geochemical anomaly and mineral prospectivity mapping in GIS. In: Handbook of exploration and environmental geochemistry, vol 11. Elsevier, Amsterdam, 368 p
- Chilès JP, Delfiner P (2012) Geostatistics: modeling spatial uncertainty, 2nd edn. Wiley, Hoboken, New Jersey, USA, 726 p
- David M (1982) Geostatistical ore reserve estimation, 1st edn. Elsevier Science, Amsterdam, 384 p
- Deutsch CV, Journel AG (1998) GSLIB. Geostatistical software library and user's guide. Oxford University Press, New York, USA
- Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York
- Isaaks EH, Srivastava RM (1990) An introduction to applied geostatistics. Oxford University Press, New York, USA, 592 p
- Journel AG, Huijbregts CJ (2004) Mining geostatistics. The Blackburn Press, Caldwell, New Jersey, USA, 600 p
- Matheron G (1963) Trait'e de g'eostatistique appliqu'ee, tome ii, vol 2. Technip, Paris
- Pyrz MJ, Deutsch CV (2014) Geostatistical reservoir modeling. Oxford University Press, New York, NY
- Remy N, Boucher A, Wu J (2009) Applied geostatistics with SGeMS (A user's guide). Cambridge University Press, Cambridge, UK, 264 p
- Sadeghi B (2020) Quantification of uncertainty in geochemical anomalies in mineral exploration. PhD thesis, University of New South Wales
- Sagar BSD, Cheng Q, Agterberg F (2018) Handbook of mathematical geosciences. Springer, Switzerland

Very Fast Simulated Reannealing

Dionissios T. Hristopoulos

School of Electrical and Computer Engineering, Technical University of Crete, Chania, Crete, Greece

Abbreviations

ASA	Adaptive simulated annealing
SA	Simulated annealing
SQ	Simulated quenching
VFSR	Very fast simulated reannealing

Definition

Very fast simulated reannealing (VFSR) is an improved version of simulated annealing (SA). The latter is a global optimization method suitable for complex, non-convex problems. It aims to optimize an objective function $\mathcal{H}(\mathbf{x})$, with respect to the D -dimensional parameter vector $\mathbf{x} = (x_1, \dots, x_D)^T$. Simulated annealing relies on random Metropolis sampling of the parameter space which mimics the physical annealing process of materials. The annealing algorithm treats $\mathcal{H}(\mathbf{x})$ as a fictitious energy function. The algorithm proposes moves which change the current state $\mathbf{x}$, seeking for the optimum state. The proposed states are controlled by an internal parameter which plays the role of temperature and controls the acceptance rate of the proposed states.

Very fast simulated reannealing employs a fast temperature reduction schedule in combination with periodic resetting of the annealing temperature to higher values. In addition, it allows different temperatures for different parameters, and an adaptive mechanism which tunes temperatures to the sensitivities of the objective function with respect to each parameter. These improvements allow fast convergence of the algorithm to the global optimum. For reasons of conciseness and without loss of generality, it is assumed in the following that the optimization problem refers to the *minimization* of the objective function $\mathcal{H}(\mathbf{x})$.

Overview

Annealing is a physical process which involves heat treatment and is used in metallurgy to produce materials (e.g., steel) with improved mechanical properties. It was presumably crucial in the manufacturing of Damascus steel swords which were famous in antiquity for their toughness, sharpness, and strength. The well-kept secrets of Damascus sword-making were lost in time as the interest in swords as weapons waned. However, their legend survived in popular culture,

inspiring the references to “Valyrian steel” in the popular *Game of Thrones* saga. The secrets of the long-lost art were presumably re-discovered in 1981 by Oleg D. Sherby and Jeffrey Wadsworth at Stanford University. A key factor for the effectiveness of annealing is the heating process which enables the material to escape metastable atomic configurations (corresponding to local optima of the energy) and to find the most stable configuration that corresponds to the global minimum of the energy.

Simulated annealing (SA) is a stochastic optimization algorithm developed by Scott Kirkpatrick and co-workers. It uses randomness in the search for the global minimum of $\mathcal{H}(\mathbf{x})$. The randomness is introduced through (i) a proposal distribution which generates new proposal states $\mathbf{x}_{\text{new}}$ of the parameter vector and (ii) through probabilistic decisions whether to accept or reject the proposal states. In contrast, in deterministic optimization methods, every move is fully determined from the current state and the properties of the objective function.

Simulated annealing belongs in a class of models inspired by physical or biological processes; these algorithms are known as *meta-heuristics* (see Ingber (2012) and references therein.) Physical annealing involves a heat treatment process which in SA is simulated using the Metropolis Monte Carlo algorithm (Metropolis et al. 1953). At every iteration of the SA algorithm, a random solution (state) $\mathbf{x}_{\text{new}}$ is generated for the parameters by local perturbation of the current state $\mathbf{x}_{\text{cur}}$. The proposed state is accepted and becomes the new current state according to an acceptance probability; the latter depends on the value of $\mathcal{H}(\mathbf{x}_{\text{new}})$ compared to $\mathcal{H}(\mathbf{x}_{\text{cur}})$ and an internal SA “temperature” variable T . The new state is accepted if the proposal lowers the objective function. On the other hand, even states $\mathbf{x}_{\text{new}}$ such that $\mathcal{H}(\mathbf{x}_{\text{new}}) > \mathcal{H}(\mathbf{x}_{\text{cur}})$ are assigned a non-zero acceptance probability which is higher for higher temperatures.

The SA temperature is initially set to a problem-specific, user-defined high value which allows higher acceptance rates for sub-optimal states. This choice helps the algorithm to effectively explore the parameter space because it allows escaping from local minima of the objective function. The temperature is then gradually lowered in order to “freeze” the system in the minimum energy state. The function $T(k)$, where k is an integer index, describes the evolution of temperature as a function of the “annealing time” k and defines the *SA cooling schedule*. This should be carefully designed to allow convergence of the algorithm to the global minimum (Geman and Geman 1984; Salamon et al. 2002).

SA can be applied to objective functions with arbitrary nonlinearities, discontinuities, and noise. It does not require the evaluation of derivatives of the objective function, and thus it does not get stuck in local minima. In addition, it can handle arbitrary boundary conditions and constraints imposed on the objective function. On the other hand, SA requires

tuning various parameters, and the quality of the solutions practically achieved by SA depends on the computational time spent. SA provides a statistical guarantee for finding the (global) optimal solution, provided the correct combination of *state proposal (generating) distribution* and cooling schedule are used (Ingber 2012). The convergence of the algorithm is based on the weak ergodic property of SA which ensures that almost every possible state of the system is visited. Achieving the statistical guarantee can in practice significantly slow down classical SA, since a very gradual cooling schedule is required. To reduce the computational time, often a fast temperature reduction schedule is applied, i.e., $T(k+1) = cT(k)$, where $0 < c < 1$, which amounts to *simulated quenching (SQ)* of the temperature. However, this exponential cooling schedule does not guarantee convergence to the global optimum.

Very fast simulated reannealing (VFSR) is also known as adaptive simulated annealing (ASA) (Ingber 1989, 2012). ASA is the currently preferred term, while VFSR was used initially to emphasize the fast convergence of the method compared to the standard Boltzmann annealing approach. ASA employs a generating probability distribution for proposal states which allows an optimal cooling schedule. This setup enables the algorithm to explore efficiently the parameter space. *Reannealing* refers to a periodical resetting of the temperature to higher (than the current) value after a number of proposal states have been accepted. Then, the search begins again at the higher temperature. This strategy helps the algorithm to avoid getting trapped at local minima. Finally, ASA allows for different temperatures in each direction of the parameter space. The parameter temperatures determine the width of the generating distribution for each parameter, thus enabling *the cooling schedule to be adapted* according to the sensitivity of the objective function in each direction. ASA maintains SA's statistical guarantee of finding the global minimum for a given objective function, but it also features significantly improved convergence speed. It has been demonstrated that ASA is competitive with respect to other non-gradient-based global optimization methods such as genetic algorithms and Taboo search (Chen and Luk 1999; Ingber 2012).

Methodology

The SA algorithm is a method for global – possibly constrained – optimization of general, nonlinear, real-valued objective functions $\mathcal{H}(\mathbf{x})$ where $\mathbf{x} \in \chi \subset \mathbb{R}^D$ is a vector of parameters that takes values in the space χ :

$$\mathbf{x}^* = \arg \min_{\mathbf{x} \in \chi} \mathcal{H}(\mathbf{x}), \text{ potentially with constraints on } \mathbf{x}.$$

The standard SA algorithm is based on the Markov chain Monte Carlo method (Geman and Geman 1984). It involves homogeneous Markov chains of finite length which are generated at progressively lower temperatures. The following parameters should thus be specified: (i) A sufficiently high initial temperature T_0 ; (ii) a final “freezing” temperature T_f (alternatively a different stopping criterion); (iii) the length of the Markov chains; (iv) the procedure for generating a proposal state $\mathbf{x}_{\text{new}}$ “neighboring” the current state $\mathbf{x}_{\text{cur}}$; (v) the acceptance criterion which determines if the proposal state $\mathbf{x}_{\text{new}}$ is admitted; and (vi) a rule for temperature reduction (annealing schedule).

Boltzmann Simulated Annealing

The SA algorithm is initialized with a guess for the parameter vector $\mathbf{x}_0$. The simulation proceeds by iteratively proposing new states $\mathbf{x}_{\text{new}}$ based on the *Metropolis algorithm* (Metropolis et al. 1953). These proposal states are generated by perturbing the current state $\mathbf{x}_{\text{cur}}$. For example, in Boltzmann annealing, this can be done by drawing the proposal state $\mathbf{x}_{\text{new}}$ from the proposal distribution

$$g(\mathbf{x}_{\text{new}}|\mathbf{x}_{\text{cur}}) = \frac{1}{(2\pi T)^{D/2}} \exp\left[-\|\mathbf{x}_{\text{new}} - \mathbf{x}_{\text{cur}}\|^2 / 2T\right].$$

For every proposed move, the difference $\Delta\mathcal{H} = \mathcal{H}(\mathbf{x}_{\text{new}}) - \mathcal{H}(\mathbf{x}_{\text{cur}})$ between the current state, $\mathbf{x}_{\text{cur}}$, and the proposed state, $\mathbf{x}_{\text{new}}$, is evaluated. If $\Delta\mathcal{H} < 0$, the proposed state is accepted as the current state. On the other hand, if $\Delta\mathcal{H} > 0$, the acceptance probability is given by the exponential expression $P_{\text{acc}} = 1/[1 + \exp \Delta\mathcal{H}/T]$ (Salamon et al. 2002; Ingber 2012). The decision whether to accept $\mathbf{x}_{\text{new}}$ is implemented by generating a random number $r \sim U(0, 1)$ from the uniform distribution between 0 and 1; if $r \leq P_{\text{acc}}$, the proposed state is accepted ($\mathbf{x}_{\text{cur}}$ is updated to $\mathbf{x}_{\text{new}}$), otherwise the present state is retained as the current state.

The above procedure is repeated a number of times before the temperature is lowered. The *acceptance rate* is equal to the percentage of proposal states that are accepted. The initial temperature is selected to allow high acceptance rate (e.g., $\approx 80\%$) so that the algorithm can move between different local optima. The temperature reduction is determined by the *annealing schedule* which is a key factor for the performance of the SA algorithm.

The annealing schedule depends on the form of the generating function used: to ensure that the global minimum of $\mathcal{H}(\mathbf{x})$ is reached, it must be guaranteed that all states $\mathbf{x} \in \chi$ can be visited an infinite number of times during the annealing process. In the case of Boltzmann annealing, this condition

requires a cooling schedule no faster than logarithmic, i.e., $T_k = T_0 \ln k_0 / \ln k$, where $k_{\max} \geq k \geq k_0$. The integer index k counts the simulation *annealing time* and $k_0 > 1$ is an arbitrary initial counter value (Ingber 2012). The final temperature, $T_{k_{\max}}$, should be low enough to trap the objective function in the global optimum state.

Fast (Cauchy) Simulated Annealing

Boltzmann SA requires a slow temperature reduction schedule as determined by the logarithm function. In practical applications, an *exponential schedule* $T_{k+1} = cT_k$ where $0 < c < 1$ is often followed. However, this fast temperature reduction enforces *simulated quenching* which drives the system too fast toward the final temperature. The exponential annealing schedule offers computational gains, but it does not fulfill the requirement of weak ergodicity (infinite number of visits to each state). Hence, it does not guarantee convergence of SA to the global minimum.

A *fast annealing* schedule which lowers the temperature according to $T_k = T_0/k$ is suitable for the *Cauchy generating distribution*:

$$g(\mathbf{x}_{\text{new}}|\mathbf{x}_{\text{cur}}) = \frac{T}{(\|\mathbf{x}_{\text{new}} - \mathbf{x}_{\text{cur}}\|^2 + T^2)^{(D+1)/2}}.$$

Fast cooling works in this case due to the “heavy tail” of the Cauchy distribution which carries more weight in the tail than the Gaussian distribution used in Boltzmann SA (Ingber 2012; Salamon et al. 2002). The resulting higher probability density for proposal states considerably different than the current state allows the algorithm to visit efficiently all the probable states in the parameter space.

Very Fast Simulated Reannealing

ASA includes three main ingredients similar to classical annealing: the generating distribution of proposal states, the acceptance probability, and the annealing schedule. ASA uses an *acceptance temperature* $T_{\text{acc}}(k_a)$ which controls the acceptance rate of proposed moves and a set of D temperatures $\{T_i(k_i)\}_{i=1}^D$ which control the width of the generating distribution for each parameter individually.

In addition to these three ingredients, ASA includes a *reannealing scheme* which rescales all the temperatures after a certain number of steps so that they adapt to the current state of $\mathcal{H}(\mathbf{x})$. Reannealing adjusts the rate of change of the annealing schedule independently for each parameter. This helps the algorithm adapt to changing sensitivities of the objective function as

it explores the parameter space, encountering points (in D -dimensional space) with very different local geometry, where $\mathcal{H}$ may change rapidly with respect to some parameters but considerably more slowly with respect to others. The main steps of ASA are outlined in the text below and in Algorithm 1.

Generation of proposal states: If the parameters x_i are constrained within $[A_i, B_i]$, where A_i and B_i are, respectively, the lower and upper bounds, the ASA proposal states are generated by means of the following steps (Ingber 1989; Chen and Luk 1999; Ingber 2012):

$$\mathbf{x}_{\text{new}} = \mathbf{x}_{\text{cur}} + \Delta \mathbf{x}, \quad (1a)$$

$$\Delta x_i = y_i(B_i - A_i), \quad i = 1, \dots, D, \quad (1b)$$

$$y_i = \text{sign}\left(u_i - \frac{1}{2}\right) T_i(k_i) \left[\left(1 + \frac{1}{T_i(k_i)}\right)^{|2u_i-1|} - 1 \right] \quad (1c)$$

$$\text{where } u_i \sim U[0, 1], \quad (1d)$$

is a random variable uniformly distributed between zero and one. Consequently, the random variable y_i is centered around zero and takes values in the interval $[-1, 1]$. The integer indices k_i are used to count time. Certain values of y_i can yield proposed parameters outside the range $[A_i, B_i]$; these values should be discarded (Ingber 1989). Lower values of temperature force y_i to concentrate around zero, while high temperature values lead to an almost uniform distribution of y_i between -1 and 1 .

Acceptance probability: The *acceptance probability* of a proposed state is given by (Chen and Luk 1999)

$$P_{\text{acc}} = \frac{1}{1 + e^{\Delta \mathcal{H}/T_{\text{acc}}(k_a)}}. \quad (2)$$

In Eqs. (1) and (2), the set of integer indices $\{k_i\}_{i=1}^D \cup \{k_a\}$ represents different *annealing times*. ASA uses one time index per parameter, so that the reannealing process can adjust the annealing time differently for each parameter. This multi-dimensional annealing schedule enables adaptation of the proposal states to different sensitivities of $\mathcal{H}(\mathbf{x})$ with respect to the parameters x_i , $i = 1, \dots, D$.

Reannealing: Every time a number N_{acc} of proposal states have been accepted, *reannealing* is performed. This procedure adjusts the annealing temperatures and times to the local geometry of the parameter space. The *sensitivities* $\{s_i\}_{i=1}^D$ are calculated as follows:

$$s_i = \left| \frac{\partial \mathcal{H}(\tilde{\mathbf{x}})}{\partial x_i} \right| \approx \left| \frac{\mathcal{H}(\tilde{\mathbf{x}} + a\hat{\mathbf{e}}_i) - \mathcal{H}(\tilde{\mathbf{x}})}{a} \right|, \quad i = 1, \dots, D, \quad (3)$$

Very Fast Simulated Reannealing,

Algorithm 1 ASA main steps.

The integers n and N count the accepted and generated states, respectively. N_s : # proposed states between annealing steps. N_{acc} : # accepted proposals between reannealing steps.

Input: Initialize: $\mathbf{x}_0 \in \mathcal{X}$, $T_{acc}(0) \leftarrow \mathcal{H}(\mathbf{x}_0)$, $k_i \leftarrow 1$, $k_a \leftarrow 1$, $\mathbf{x}_{cur} \leftarrow \mathbf{x}_0$, $T_i(0) \leftarrow 1$, $n \leftarrow 0$, $N \leftarrow 0$

```

1 while termination condition is not met do Annealing loop
2   Generate proposal state  $\mathbf{x}_{new}$  using Eq. (1) ;
3    $N \leftarrow N + 1$  (increase generated states counter) ;
4   if  $\mathcal{H}(\mathbf{x}_{new}) < \mathcal{H}(\mathbf{x}_{cur})$  then Update current state
5     |  $\mathbf{x}_{cur} \leftarrow \mathbf{x}_{new}$ ;  $n \leftarrow n + 1$  (increase accepted states counter)
6   else
7     | Calculate acceptance probability  $P_{acc}$  from Eq. (2) ;
8     | if  $P_{acc} > r \sim U(0, 1)$  then Update current state
9     | |  $\mathbf{x}_{cur} \leftarrow \mathbf{x}_{new}$ ;  $n \leftarrow n + 1$  (increase accepted states counter)
10    | end
11  end
12  if  $n > N_{acc}$  then
13    | Reannealing based on Eqs. (3)-(5) ;
14    |  $n \leftarrow 0$  (reset accepted states counter)
15  else
16    | Go to Step 2
17  end
18  if  $N > N_s$  then
19    | Annealing based on Eqs. (6) ;
20    |  $N \leftarrow 0$  (reset generated states counter)
21  end
22 end

```

where $\tilde{\mathbf{x}}$ is the *current optimal parameter vector*, a is a small increment, and $\hat{\mathbf{e}}_i$ is a D -dimensional unit vector in the i -th direction of parameter space, i.e., $\{\hat{\mathbf{e}}_i\}_k = \delta_{i,k}$ ($\delta_{i,k} = 1$ for $i = k$ and $\delta_{i,k} = 0$ for $i \neq k$ being the Kronecker delta).

Reannealing adjusts the temperatures and annealing times as follows ($\leftarrow$ implies assignment of the value on the right side of $\leftarrow$ to the variable appearing on the left side)

$$T_i(k_i) \leftarrow \frac{s_{\max}}{s_i} T_i(k_i), \quad s_{\max} = \max_{i=1, \dots, D} s_i, \quad (4a)$$

$$k_i \leftarrow \left\lceil \frac{1}{c} \log \left(\frac{T_i(0)}{T_i(k_i)} \right) \right\rceil^D, \quad i = 1, \dots, D, \quad (4b)$$

where $c > 0$ is a user-defined parameter that adjusts the rate of reannealing and $T_i(0)$ is usually set to unity (Chen and Luk 1999).

Similarly, the acceptance temperature is rescaled according to

$$T_{acc}(k_a) \leftarrow \mathcal{H}(\tilde{\mathbf{x}}), \quad T_{acc}(0) \leftarrow \mathcal{H}(\mathbf{x}_a), \quad (5a)$$

$$k_a \leftarrow \left\lceil \frac{1}{c} \log \left(\frac{T_{acc}(0)}{T_{acc}(k_a)} \right) \right\rceil^D, \quad (5b)$$

where $\mathbf{x}_a$ is the *last accepted state*.

The reannealing procedure enables ASA to decrease the temperatures T_i along the high-sensitivity directions of $\mathcal{H}(\mathbf{x})$, thus allowing smaller steps in these directions, and to increase T_i along the low-sensitivity directions, thus allowing larger jumps.

Annealing: After a number of steps N_s have been completed, the *annealing* procedure takes place. The annealing times increase by one, and the annealing temperatures are modified according to Eq. (6). The annealing schedules for the temperature parameters follow the stretched exponential expression

$$\begin{aligned} T_{\text{acc}}(k_a) &= T_{\text{acc}}(0) \exp\left(-ck_a^{1/D}\right), \\ T_i(k_i) &= T_i(0) \exp\left(-ck_i^{1/D}\right), i = 1, \dots, D, \end{aligned} \quad (6)$$

where $T_i(k_i)$ is the current temperature for the i -th parameter (Ingber 1989).

Termination: The above operations are repeated until the algorithm terminates. Different termination criteria can be used in ASA depending on the available computational resources and a priori knowledge of $\mathcal{H}$. Such criteria include the following: (i) an average change of $\mathcal{H}(\mathbf{x})$ (over a number of accepted proposal states) lower than a specified tolerance; (ii) exceedance of a maximum number of iterations (generated proposal states); (iii) exceedance of a maximum number (usually proportional to D) of $\mathcal{H}$ evaluations; (iv) exceedance of a maximum computational running time; and (v) attainment of a certain target minimum value for $\mathcal{H}$.

Initialization: An initial value, $T_{\text{acc}}(0)$, should be assigned to the acceptance parameter. One possible choice is to set $T_{\text{acc}}(0) = \mathcal{H}(\mathbf{x}_0)$ where $\mathbf{x}_0$ is the random initial parameter vector (Chen and Luk 1999). A more flexible approach matches $T_{\text{acc}}(0)$ with a selected acceptance probability, e.g., $P_{\text{acc}} = 0.25$ (Iglesias-Marzoa et al. 2015). The optimal values of N_s and N_{acc} do not seem to depend crucially on the specific problem (Chen and Luk 1999). Often an adequate choice for N_{acc} is on the order of tens or hundreds and for N_s on the order of hundreds or thousands. Higher values may be necessary to adequately explore high-dimensional and topologically complex parameter spaces. A good choice for the annealing rate control parameter is $c \approx 1 - 10$. In general, higher values lead to faster temperature convergence at the risk that the algorithm may get stuck near a local minimum. Lower values lead to slower temperature reduction and therefore increase the computational time. Given ASA's adaptive ability, the choices for c , N_{acc} , N_s do not critically influence the algorithm's performance but they have an impact on the computational time. For problems where ASA is used for the first time, it is a good idea to experiment with different choices (Iglesias-Marzoa et al. 2015).

Applications

An introduction to SA for spatial data applications is found in Hristopulos (2020). Geostatistical applications are reviewed in Pyrcz and Deutsch (2014). An informative review of ASA (VFSR) is given in Ingber (2012), while a mathematical treatment of SA is presented in Salamon et al. (2002). ASA has been successfully used for geophysical inversion (Sen and Stoffa 1996). The ASA algorithm and its application in various signal processing problems are reviewed in Chen and Luk (1999). A detailed investigation of ASA's application for

determining the radial velocities of binary star systems and exoplanets is given in Iglesias-Marzoa et al. (2015).

Software Implementations

In Matlab, SA is implemented by means of the command `simulannealbnd`. In R, the packages `optimization` and `optim_sa` provide SA capabilities. In Python, this functionality can be found in the `scipy.optimize` module. A C-language code for ASA (VFSR) developed by Lester Ingber is available from his personal website.

Summary and Conclusions

ASA (VFSR) is a computationally efficient implementation of the simulated annealing algorithm which employs a fast (stretched-exponential) annealing schedule in combination with a reannealing program which periodically resets the temperature. ASA is an adaptive algorithm: both the annealing and reannealing schedules take into account the sensitivity of the objective function in different directions of parameter space. Thanks to this adaptive ability, ASA is an effective stochastic global optimization method which has been successfully applied to complex, nonlinear objective functions that may involve many local minima. Estimates of parameter uncertainty can be obtained by means of the Fisher matrix and Markov chain Monte Carlo methods.

ASA converges faster than the classical SA algorithm, it allows each parameter to adapt individually to the local topology of the objective function, and it involves a faster schedule for temperature reduction (Iglesias-Marzoa et al. 2015). The counterbalance of these advantages is that ASA is more complex than the classical SA algorithm, it involves more internal parameters, and the stretched exponential expression for temperature reduction requires using double precision arithmetic and checking of the exponents to avoid numerical problems. Overall, ASA has found several applications in the geosciences (Sen and Stoffa 1996; Pyrcz and Deutsch 2014; Hristopulos 2020).

Cross-References

- Markov Chain Monte Carlo
- Optimization in Geosciences
- Simulated Annealing
- Simulation

Bibliography

- Chen S, Luk BL (1999) Adaptive simulated annealing for optimization in signal processing applications. *Signal Process* 79(1):117–128. [https://doi.org/10.1016/S0165-1684\(99\)00084-5](https://doi.org/10.1016/S0165-1684(99)00084-5)

- Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell* 6(6):721–741. <https://doi.org/10.1109/TPAMI.1984.4767596>
- Hristopulos DT (2020) Random fields for spatial data modeling: a primer for scientists and engineers. Springer, Dordrecht. <https://doi.org/10.1007/978-94-024-1918-4>
- Iglesias-Marzoa R, López-Morales M, Arévalo Morales MJ (2015) The rvfit code: a detailed adaptive simulated annealing code for fitting binaries and exoplanets radial velocities. *Publ Astron Soc Pac* 127(952):567–582. <https://doi.org/10.1086/682056>
- Ingber L (1989) Very fast simulated re-annealing. *Math Comput Model* 12(8):967–973. [https://doi.org/10.1016/0895-7177\(89\)90202-1](https://doi.org/10.1016/0895-7177(89)90202-1)
- Ingber L (2012) Adaptive simulated annealing. In: Hime A, Ingber L, Petraglia A, Petraglia MR, Machado MAS (eds) *Stochastic global optimization and its applications with fuzzy adaptive simulated annealing*. Springer, Heidelberg, pp 33–62. <https://doi.org/10.1007/978-3-642-27479-4>
- Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. *J Chem Phys* 21(6):1087–1092. <https://doi.org/10.1063/1.1699114>
- Pyrce MJ, Deutsch CV (2014) *Geostatistical reservoir modeling*, 2nd edn. Oxford University Press, New York
- Salamon P, Sibani P, Frost R (2002) Facts, conjectures, and improvements for simulated annealing. *SIAM monographs on mathematical modeling and computation*, SIAM, Philadelphia. <https://doi.org/10.1137/1.9780898718300>
- Sen MK, Stoffa PL (1996) Bayesian inference, Gibbs' sampler and uncertainty estimation in geophysical inversion. *Geophys Prospect* 44(2):313–350. <https://doi.org/10.1111/j.1365-2478.1996.tb00152.x>

Virtual Globe

Jaya Sreevalsan-Nair
Graphics-Visualization-Computing Lab, International
Institute of Information Technology Bangalore,
Electronic City, Bangalore, Karnataka, India

Definition

While data visualization is widely used for exploring and visually comprehending the data, virtual reality (VR) takes it to a higher level of immersive experience. VR provides an artificially simulated environment in which the user can be part of the environment and interact with it, as permitted by the VR application. VR has been used as an extension to several graphical user interface (GUI) applications, but one of the most relevant applications is representing the three-dimensional (3D) model of the planet Earth itself. This 3D representation of the Earth extends to a four-dimensional (4D) one whenever spatiotemporal data is used. Such an application allows the user to explore and experience the environment and to edit the data loaded to visualize the world, representing the natural and man-made artifacts. Such a VR application is referred to as the *virtual world*

(Yu and Gong 2012). An example of a virtual globe that has popularized this technology in recent times is the Google Earth VR.¹

Overview

The history of virtual globes is not very old. The Geoscope was built as a large spherical display that entailed modeling and simulating a virtual environment of the Earth using computers, by the architect Buckminster Fuller in the 1960s (Yu and Gong 2012). This is considered as the early beginnings of the virtual globe. This has been further conceptualized by Al Gore in the early 1990s as the “Digital Earth” to serve as a digital access point for massive scientific as well as human-centric data. After a hiatus, by the late 1990s, the advent of high-performance graphics processor unit (GPU) technology, availability of high-resolution satellite images, and high-speed Internet connectivity paved the way to actual implementations of the virtual globes. The popular virtual globes released by the prominent technology as well as space organizations include the Encarta Virtual Globe 98 and Bing Maps Platform (Microsoft), 3D World Atlas (Cosmi), WorldWind (NASA), Google Earth (Google), and ArcGIS Explorer (ESRI). There are also several free virtual globe software, such as Earth3D and Marble, where the latter is open source too.

The virtual globe must be distinguished from a data visualization tool with a GUI involving 3D rendering of the globe. An example of such a browser-based visualization tool with 3D animated map for visualization of global weather conditions is the Earth Nullschool.² While a virtual globe is a special visualization tool of the planet, all visualization tools depicting the Earth cannot be considered as virtual globes. This is because virtual globes involve VR where the user is *immersed* in the virtual environment, which is achieved by using simulations of the camera orientation and position in the graphical application. The virtual globes have also been referred to as “geobrowsers” (Yu and Gong 2012) owing to their function of exploring the Earth, which is analogous to the exploration of the Internet by the Web browsers.

In a virtual globe, the worldwide geographic maps can be graphically rendered using different viewpoints and geographical projection systems (Yang et al. 2018). It can be set up in one of the four ways – (i) the conventionally used exocentric 3D globe, where the viewer sees from outside of the globe; (ii) a 2D flat map of the globe projected onto a plane; (iii) an egocentric 3D globe, where the viewer is immersed in the environment; or (iv) a curved map of the

¹<https://arvr.google.com/earth/>

²<https://earth.nullschool.net/about.html>

globe. The curved map is a projection on a section of a sphere, giving the appearance of an opera stage or a theater. Extensive user study has shown that the most accurate responses in the analytical tasks using the virtual globe, as well as in direction estimation, have been obtained in the case of the exocentric virtual globe. Motion sickness is a known issue users face in VR applications. The user study also revealed that the motion sickness was least felt in the cases of exocentric globe and flat maps. Exocentric globe, where the viewer sees the world as from space, is the most intuitive, and egocentric globe, where the viewer is in the center of the sphere, which renders the globe in its inner concave walls, is the least intuitive. The egocentric view provided the best immersive experience, but the perceptual distortion of the familiar convex (spherical) globe to concave walls made it a poor choice for this specific VR application. Similarly, the flat map using Hammer map projection, i.e., equal-area projection, has been found to be a better fit for VR application compared to the spherical projection used in curved map. This study has been conducted using head-mounted displays (HMDs).

HMDs and controllers are important accessories for virtual globes, and the technologies supported by the VR application are critical for its usability, e.g., Google Earth VR supports HMDs from HTC Vive and Oculus Rift, which are widely used HMDs. Figure 1 shows how an HMD and a hand controller are used for different settings of the Google Earth VR application.

Applications

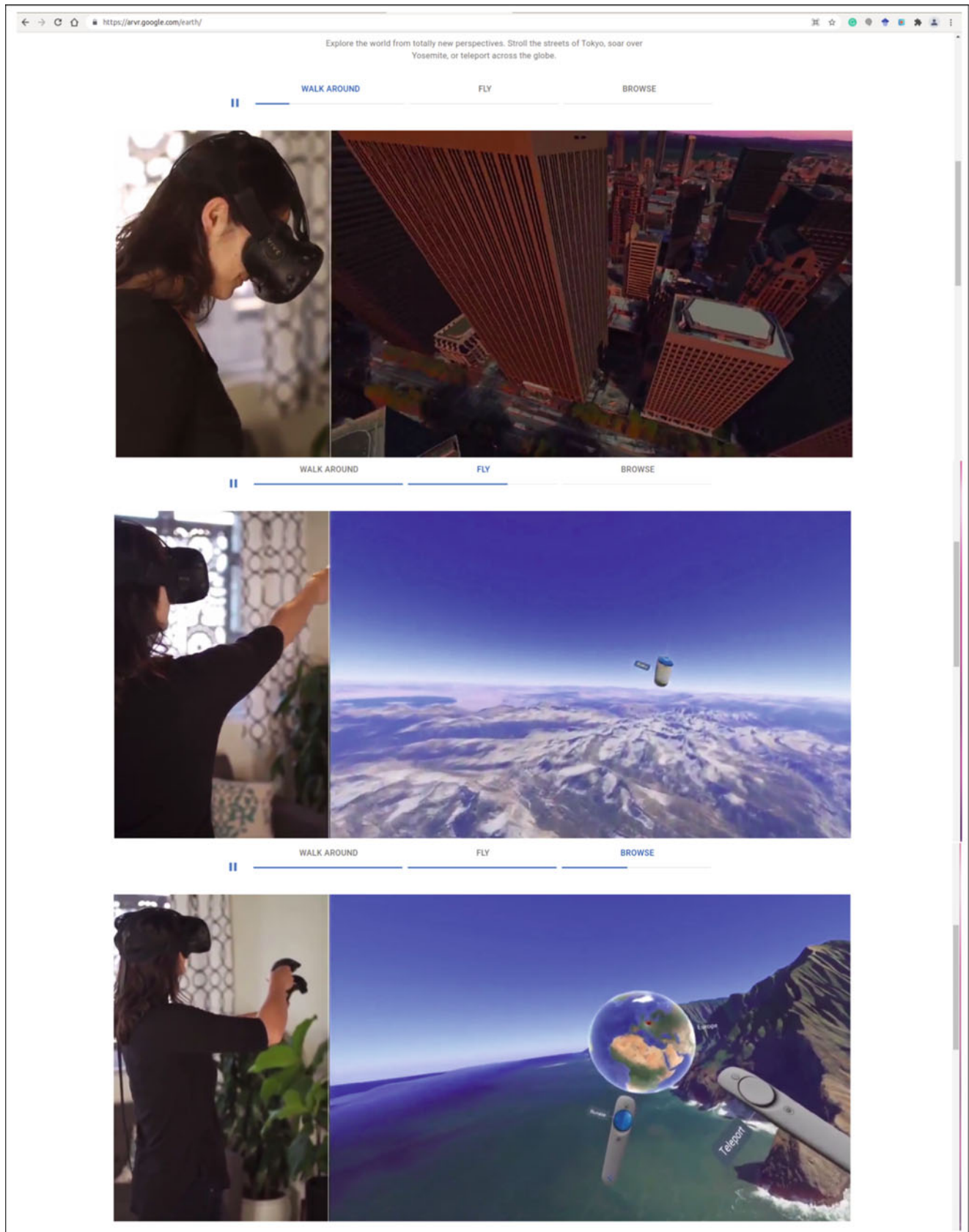
Google Earth itself has served several purposes since its release in 2005 (Yu and Gong 2012). It has been instrumental in supporting and advancing scientific research in the different domains, such as earth observations, natural hazards, human geography, etc. It has provided services for data visualization, collection, validation, integration, dissemination, modeling, and exploration and also as a decision support system. Its merits include the following: (i) support for global-scale research owing to the access to global data acquired from high-resolution satellite imaging, aerial imagery, etc., through the use of a single coordinate system, i.e., the World Geodetic System of 1984 data, and (ii) browser-based easy-to-use visualization through the use of WebGL. Studying the evolution of Google Earth points to a few pertinent gaps to be addressed to improve the uptake in the usability of virtual globes. These include the data inconsistency issue of nonuniform precision of satellite imaging and other Earth observation data; limited interoperability of the digital elevation model with other high-resolution topography data, e.g., acquired using Light Detection and Ranging (LiDAR) sensors (Bernardin et al. 2011); and limited flexibility in analytical functionality when using specific file

format, i.e., Keyhole Markup Language (KML), Web services, and API or software development kit (SDK), e.g., NASA WorldWind.

The learnings from the Google Earth project have been incorporated in the making of the Google Earth VR, which is a relatively modern virtual globe, launched in late 2016 (Käser et al. 2017). In addition to the features to scale-and-fly, search, rotate, and teleport using the HMD and controllers (Fig. 1), Google Earth VR also provides haptics and sound effects, e.g., when using the 3D Cone Drag gesture to grab the planet itself, the user can feel and hear the static friction that the Earth exhibits. In terms of performance, despite the requirement to render trillions of triangles in the data, Google Earth VR with both hardware (GPU) and software optimizations provides 90 fps (frames per second), which is a requirement for VR.

The improved user interactivity of the virtual globe has found itself an application in facilitating public participation in an urban planning project (Wu et al. 2010). In an example where Google Earth is used as the virtual globe application, a distributed system has been designed to display urban planning information gathered from the virtual globe and to facilitate discussion forums between urban planning designers and the general public. The general public can explore the relevant geospatial data from a macro- to a micro-spatial scale, where the latter is the street scale. The architecture for such a distributed system has been possible by using the Web Service technology, namely, Service-Oriented Architecture (SOA) integrated with a virtual globe, which is now treated as a Web Service. It has three layers, namely, the support, the service, and the client layers. Such a system has enabled the public to inspect an urban planning project at their own convenience and provide feedback to the planners. Additionally, the application uses interoperable CityGML, a standard descriptive language for city models, which improves the system extensibility. However, this also requires appropriate system capability to reduce network latency issues incurred owing to transfer of large XML files in CityGML. Overall, this example shows how virtual globes can be relevant when integrated with other systems for pertinent applications.

Demonstrating the diversity in applications, Crusta is an example of virtual globe to visualize planetary scale topography for geological studies (Bernardin et al. 2011). Crusta has been proposed to address the challenges in visualization of digital elevation models (DEMs) of high spatial resolution of $0.5\text{--}1\text{ m}^2$ per sample and large extent ($>2000\text{ km}^2$). The globe is geometrically rendered using a 30-sided polyhedron, defined by base patches, that can be subdivided into different resolutions in a pyramidal format. The subdivision is allowed to accommodate input data of varying resolutions, including global (BlueMarble by NASA) and local (acquired using tripod LiDAR) spatial scales. The use of the polyhedron resolves the issue of “pinching” or distortion at the poles.



Virtual Globe, Fig. 1 Examples of the usage of head-mounted display (HMD) and hand controller by a viewer in a virtual globe application, namely, the Google Earth VR, and its corresponding visualizations. (Image courtesy: <https://arvr.google.com/earth/>)

Crusta implements dynamic shading for its graphical rendering, which has clearly revealed the previous ruptures along the Owens Valley fault in Eastern California. One of the strengths of Crusta has been in providing visualizations with high-resolution topographic detail along with the shading and exaggeration of the topography, in addition to the vertical exaggeration that all virtual globes provide. This is a use case of virtual globes for geological exploration and knowledge discovery.

Archaeology is an area where virtual globes has practical applications, as they help to preserve fragile artifacts by its remote exploration without the physical presence of human beings. Archaeological data involves spatiotemporal information and hence is 4D (De Roo et al. 2017). The usability test of extending the application of a 3D virtual globe to a 4D archaeological geographical information system (GIS) has revealed its feasibility for the same. Testing the application on five archaeologists with a scenario of excavation site of Kerkhove, Belgium, has shown that the 4D virtual globe can be used for fieldwork preparation, report generation, and communication by knowledge sharing on the Internet. While the spatial and temporal dimensions have been decoupled to a larger extent in its treatment in this specific application, there is scope for tighter coupling between the two to improve querying as well as simulation in the combined 4D space. Altogether, virtual globes can be used in archaeology for site exploration, as well as for gaining insights to the site evolution using 4D GIS.

A similar use case of 4D geospace in virtual globe is in visualizing and assessing geohazards (Havenith et al. 2019). The requirement here for assessment and analysis of spatiotemporal phenomena, namely, geohazards, is to have a combination of GIS, and geological modeling, supplemented by interactive tools, such as the freely available Google Earth. Virtual geographic environments (VGEs) have been originally conceptualized and implemented to provide multi-dimensional spatiotemporal geographic analysis along with real-time collaboration between geoscientists. However, the relatively low contributions by geoscientists to build VGEs had not led to the intended fluid interactions and results sharing, despite the presence of collaborative virtual environments (CVEs). In the past, CVEs had been effectively used for collaborative visualization and analysis of geospatial data, using mapping elements, interactive cooperative work, and semiotics. In this backdrop, the development of VR has now allowed immersive visualization, thus adding a new dimension to the VGEs. The increased commoditization and advent of low-cost VR technology has promoted the use of virtual globes as a VGE. In addition to geological applications, e.g., Crusta, the virtual globes can now be extended with numerical models to simulate, visualize, and analyze natural hazards, such as dam breaks, rock falls, etc. Such a system requires combining large-scale high-resolution geophysical-

geological data, DEMs, textures from remote imagery, etc. and visualization support, such as 3D stereo-visualization with HMD for VR. While there is scope for extending Google Earth VR to studying geohazards, there is now the need for integrating more compute resources to solve numerical models, along with maintaining the frame rate of 90 fps for rendering in VR. As an application dealing with risk assessment, handling the epistemic and total uncertainty in geohazard assessment is a desirable feature in the 4D geospace depicted in virtual globe. Thus, handling uncertainty has to be considered as a future development. Another critical feature of a fully integrated 3D geohazard modeling application is the seamless interoperability of the virtual globe with several decision support systems.

Apart from geospatial applications, virtual environments also play an instrumental role in geo-education (Shakirova et al. 2020). VR is considered one of the effective emerging technologies that can be used in STEM education. An example is the use of virtual globes in geoscience education. A specific study testing the usability of Google Earth VR, Apple Maps (for “visiting” cities), My Way VR (for a virtual journey through countries and cultures), and VR Museum of Fine Arts (for visiting museums), for teaching/learning a physical geography course, has revealed that the experience of using the virtual globes had shifted the aspects of e-learning, such as gamification, interaction, professional activity imitation, etc., from being “desirable” to “expected,” by the students. This shows that the impact of virtual globes has been significant, leading to an increase in demand of its application in e-learning. In addition to developing a pedagogy involving experiences of physical phenomena (climatic changes, natural phenomena) and world tourism, using VR in geo-education provides opportunities for creating content through international educational groups, thus improving the quality of global education. Altogether, the use of more advanced GVEs, through virtual globes, in geo-education promotes geographic research.

Future Scope

While there are specific challenges to be addressed in specialized applications, the future of virtual world technology is in its transition to open-source software (OSS) (Coetzee et al. 2020). Several technologies of virtual world are already available as free applications, for which Google Earth is an example. One step further, NASA WorldWind³ is an example of an open-source virtual globe (Pirotti et al. 2017). WorldWind was originally developed for visualizing the NASA MODIS (Moderate Resolution Imaging Spectroradiometer) land

³<https://worldwind.arc.nasa.gov/about/>

products. It enables visualization of multidimensional data with the help of a framework that uses a digital terrain model (DTM) as its base. It supports several applications involving local communities through GeoWeb 2.0, e.g., exploration of cultural heritage sites, virtual cities to promote world tourism, etc. Just like Google Earth, it was initially freely available, but then it slowly transitioned to open source in the early 2000s. The open source of WorldWind has been available in a variety of programming languages and operating systems – in C# since 2003, in Java since 2006, in Android since 2012, and in JavaScript since 2014. This has been made possible by publishing a SDK that developers can use for building their software for GUIs. These GUIs are developed on the browser increasingly, thus using HTML5 and JavaScript. The SDK supports several file formats for input data, e.g., GeoTIFF, JPG, and PNG formats for raster data, ESRI Shapefiles, KML, and Collada for vector models. It also supports REST, WMS, and Bing Maps services for adding Web Service layers. Releasing the WorldWind software as open source has encouraged the developer community to build several relevant applications, e.g., the Environment Space and Time Web Analyzer (EST-WA) based on netCDF file format, PoliCrowd as a collaborative tourism/cultural heritage platform based on Java, etc. Thus, the evolution of WorldWind in itself and the engagement of larger developer community have extensively promoted the collaborative use of virtual globes in several innovative applications.

One of the near-future goals in terms of improving the usability of virtual globes is to include them in OSGeo (open-source geospatial) software. This can open up further opportunities in benchmarking, technology maturing, and increased uptake in applications worldwide (Coetzee et al. 2020). Overall, the virtual globe continues to be a useful geoscientific exploratory and analytic tool through immersive experience.

Bibliography

- Bernardin T, Cowgill E, Kreylos O, Bowles C, Gold P, Hamann B, Kellogg L (2011) Crusta: a new virtual globe for real-time visualization of sub-meter digital topography at planetary scales. *Comput Geosci* 37(1):75–85
- Coetzee S, Ivánová I, Mitasova H, Brovelli MA (2020) Open geospatial software and data: a review of the current state and a perspective into the future. *ISPRS Int J Geo Inf* 9(2):90
- De Roo B, Bourgeois J, De Maeyer P (2017) Usability assessment of a virtual globe-based 4D archaeological GIS. In: *Advances in 3D geoinformation*. Springer, pp 323–335
- Havenith HB, Cerfontaine P, Mreyen AS (2019) How virtual reality can help visualise and assess geohazards. *Int J Digital Earth* 12(2): 173–189
- Käser DP, Parker E, Glazier A, Podwal M, Seegmiller M, Wang CP, Karlsson P, Ashkenazi N, Kim J, Le A et al (2017) The making of Google Earth VR. In: *ACM SIGGRAPH 2017 talks*, ACM, pp 1–2
- Pirotti F, Brovelli MA, Prestifilippo G, Zamboni G, Kilsedar CE, Piragnolo M, Hogan P (2017) An open source virtual globe rendering

engine for 3D applications: NASA WorldWind. *Open Geospat Data Softw Stand* 2(1):1–14

- Shakirova N, Said N, Konyushenko S (2020) The use of virtual reality in geo-education. *Int J Emerg Technol Learn* 15(20):59–70
- Wu H, He Z, Gong J (2010) A virtual globe-based 3D visualization and interactive framework for public participation in urban planning processes. *Comput Environ Urban Syst* 34(4):291–298
- Yang Y, Jenny B, Dwyer T, Marriott K, Chen H, Cordeil M (2018) Maps and globes in virtual reality. *Comput Graph Forum* 37:427–438
- Yu L, Gong P (2012) Google Earth as a virtual globe tool for earth science applications at the global scale: progress and perspectives. *Int J Remote Sens* 33(12):3966–3986

Vistelius, Andrey Borisovich

Stephen Henley

Resources Computing International Ltd, Matlock, UK



Fig. 1 Andrey Borisovich Vistelius (1915–1995)

Biography

Andrey Borisovich Vistelius was born in St. Petersburg, Russia, on 7 December 1915. His father Boris Vistelius was a lawyer before the October Revolution of 1917. Boris's father (Andrey's grandfather) was a senior civil servant in the Russian Empire. Relatives of Andrey's mother (the Bogaevsky family) included distinguished academics.

There is no published information on Vistelius' early childhood and how he and his family fared during the

Modified after Henley S. (2018) Andrey Borisovich VISTELIUS. In: Daya Sagar B., Cheng Q., Agterberg F. (eds) *Handbook of Mathematical Geosciences*. Springer, Cham; direct link to the article (<https://link.springer.com/article/10.1007/s13202-015-0213-7>). According terms of Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>)

revolution and civil war. However, in 1933 Andrey entered Leningrad State University as a student. But in 1935, the family was exiled from Leningrad like many other intellectuals; first to a remote village in middle Russia, later to the city of Samara. As a result, Andrey's education was interrupted. His studies in Leningrad were eventually resumed and in 1939 he graduated brilliantly from the Department of Mineralogy.

Geology and mathematics were his overwhelming passions, with emphasis on practical applications. He was committed to evidence-based science, which in the prevailing atmosphere of "lysenkoism" was a barrier to his career progress.

During World War II, Andrey Vistelius was trapped in besieged Leningrad. He was not enlisted into the army because of poor eyesight. However, his studies continued, with award of his "Candidacy" (equivalent to PhD) in 1941, and Doctor of Science in 1948. He then worked in several state geological organizations, including as director of "expeditions" (responsible for regional geological mapping).

However, to overcome continuing political obstacles, with the support of mathematical Academicians, in 1961 a group headed by Vistelius was set up as the Laboratory of Mathematical Geology within the Leningrad Branch of the Steklov Institute of Mathematics (LOMI) of the USSR Academy of Sciences.

In 1968, he was instrumental, with others, in founding the International Association for Mathematical Geology (IAMG), and was elected its first president. Although the Cold War prevented him from participating in many of IAMG's activities, he continued work as a prolific researcher in Leningrad.

Vistelius' scientific achievements, notably in process modeling and statistical testing of hypotheses against real data were recognized by IAMG, awarding him the Krumbein Medal in 1980 and naming a new award (for young scientists) after him. The Soviet Union authorities rejected this latter proposal on the grounds that such an honor should not be conferred upon a living person, so it was initially designated the President's Award, and changed to the Vistelius Award, as originally intended, after his death in 1995.

In 1987 the Laboratory of Mathematical Geology was transferred to the Institute of Precambrian Geology and Geochronology, as a prelude to its conversion to an Institute in 1991 when the Russian Academy of Natural Sciences (RANS) was founded, and Andrey Vistelius was named an Honorary Member of this Academy.

Andrey Borisovich Vistelius died on 12 September, 1995. He continued to work, lucid and inventive, to the end despite serious illness. In 1992, an English translation of his life's work "Principles of Mathematical Geology" was published (Vistelius 1992) – a reworked version of his Russian monograph published in 1980. Altogether he had more than 200 published works (Henley 2018; Dech and Henley 2003; Romanova and Sarmanov 1970). Dech and Glebovitsky (2000) give a detailed account of the many fields in which the work of Vistelius advanced geological knowledge. A special issue of the *Journal of Mathematical Geology* (volume 35, number 4) in memory of Vistelius was published in 2003 with papers by many of his former colleagues, as well as one previously unpublished paper by Vistelius himself. The breadth of geoscientific subject matter and mathematical approaches shown by this collection of papers is ample illustration of the scientific legacy of Andrey Borisovich Vistelius.

Bibliography

- Dech VN, Glebovitsky VA (2000) Establishment of mathematical geology in Russia: some brush strokes in the portrait of the founder of mathematical geology, Prof. A.B. Vistelius. *Math Geol* 32(8):897–918
- Dech VN, Henley S (2003) On the scientific heritage of Prof. A.B. Vistelius. *Math Geol* 35(4):363–379
- Henley S (2018) Andrey Borisovich Vistelius. In Sagar BSD, Cheng Q, Agterberg F (eds) *Handbook of mathematical geosciences: fifty years of IAMG*. SpringerOpen, pp 793–812, 914pp. <https://doi.org/10.1007/978-3-319-78999-6>
- Romanova MA, Sarmanov OV (1970) Chapter 2: The published works of A. B. Vistelius. In: Romanova MA, Sarmanov OV (eds) *Topics in mathematical geology*. Springer, New York, pp 6–12
- Vistelius AB (1992) *Principles of mathematical geology*. Kluwer Academic Publishers, Dordrecht/Boston/London, 477 p

Watson, Geoffrey S.

Noel Cressie¹ and Carol A. Gotway Crawford²

¹School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW, Australia

²Albuquerque, NM, USA

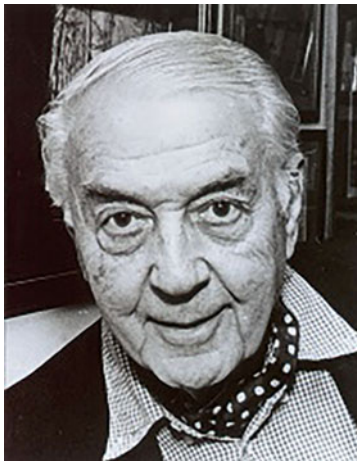


Fig. 1 By Grasso Luigi – Own work, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=68766582>

Biography

Geoffrey S. Watson was born in 1921 in Bendigo, a rural community in the state of Victoria, Australia. He was an Australian statistician who spent the majority of his career in North America (USA and Canada), with shorter periods in Australia and the UK. He received a Bachelor of Arts with Honours from the University of Melbourne (1942) and a PhD in statistics from North Carolina State University (1951). He wrote his PhD thesis while visiting Cambridge University in the UK and, while there, he worked with James Durbin of the

London School of Economics. Together they published companion papers in 1950 and 1951 in the journal *Biometrika*, on what is now called the Durbin-Watson statistic. It is used to test the presence of serial correlation in data observed at equal time intervals and is a fundamental component of most time series software packages. Generalizing from time series to spatial data brings attention to the variogram at its closest spatial lag: If this value is small compared to its sill, a null hypothesis of spatial independence is rejected. With fast computing, it is now possible to test for temporal (spatial) independence by computing the statistic for all permutations of the times (locations) of the data; from this null distribution, the percentile of the observed statistic can be obtained.

Geof (as he was known and as he signed his letters) was a quintessential statistical scientist, curious about new ideas in science, with strong mathematical statistical skills and a gift for clear writing in the collaborating discipline. In the geosciences, his most influential work was to help resolve the controversy between “fixism” and “mobilism,” leading up to mobile-plate tectonic theory (continental drift). His research on statistics for directional data on the circle and the sphere allowed hypothesis testing on paleomagnetism of rocks on different sides of the oceans. He is the author of an important monograph, *Statistics on Spheres*, published in 1983. He was also influential in bringing the work of Georges Matheron and his centers of geostatistics and mathematical morphology, located in Fontainebleau, France, to the English-speaking world. Watson spent a summer visiting Matheron in 1972.

Much of his working life was spent at Princeton University; he arrived in 1970, to head the Department of Statistics, a position he held until 1985, and he retired from the university in 1992 as emeritus professor. While at Princeton, Geof became involved in energy and environmental issues such as estimating US oil and gas reserves, climate trends, and assessing the effects of air pollution on public health and the ozone layer.

An autobiographical account of his life and research is given in two chapters of the edited book, *The Art of Statistical Science: A Tribute to G.S. Watson*, published in 1992. Geof was an accomplished watercolor painter; during his retirement until his death in 1998, he had shown his work in several solo shows in Princeton art galleries. His personality was unique, coming from a palette of Australian, British, and US influences.

Cross-References

- [Geostatistics](#)
- [Mathematical Morphology](#)
- [Matheron, Georges](#)
- [Spatial Statistics](#)
- [Variogram](#)

Bibliography

- Durbin J, Watson GS (1950) Testing for serial correlation in least squares regression, I. *Biometrika* 37:409–428
- Durbin J, Watson GS (1951) Testing for serial correlation in least squares regression, II. *Biometrika* 38:159–178
- Mardia KV (ed) (1992) *The Art of Statistical Science: A Tribute to G.S. Watson*. Wiley, Chichester
- Watson GS (1983) *Statistics on Spheres*. Wiley, New York

Wavelets in Geosciences

Guoxiong Chen¹ and Henglei Zhang²

¹State Key Laboratory of Geological Processes and Mineral Resources, China University of Geosciences, Wuhan, China

²Institution of Geophysics and Geomatics, China University of Geosciences, Wuhan, China

Definition

Wavelets (or wavelet transform) are a series of mathematical functions which can analyze nonstationary signal with both time and frequency resolutions. The fundamental idea behind wavelet transform is to analyze signals according to scale, by using the scaling and shifting transform of wavelet function. They decompose a signal into different frequency components that make it up, just like the Fourier transform, but also identify where a certain frequency or wavelength exists in the temporal or spatial domain. Wavelets have been endowed with many excellent properties for signal analyzing, including the time-frequency localization, multiresolution, compression (or sparsity), decorrelation, and so

on. Wavelets have therefore enjoyed a tremendous attention and success in Geosciences community, including, but not limited to, the applications of time-frequency analysis, multi-scale decomposition, filtering/denoising, and fractal/multifractal analysis.

Historic Wavelets Analysis

Wavelets have greatly fascinated the research in engineering, mathematics and nature science communities since 1980s because of the versatility in numerical analysis and function approximation. Its origin should be traced back to the Fourier transform theory that was pioneered by Joseph Fourier in 1807 to express signal as the sum of a, possibly infinite, series of sines. Fourier transform provides a powerful tool for analyzing stationary signals by converting them into frequency-domain, in order to explain the characteristics of frequency (or wavelength) of signals. As the most classical and successful signal processing and analyzing method, Fourier transform has achieved great success in both theory and applications in the past 200 years. However, it is theoretically only suitable for analyzing stationary signal without resolution of timescale, making them unavailable to realize the time-frequency localization analysis of nonstationary signal. For this reason, Gabor (1946) proposed the windowed Fourier transform, which has made substantial progresses in the aspect of time-frequency localization of signals, whereas it is often limited in the practical applications for analyzing complicated signals due to the lack of self-adaptability of time-frequency window. As a milestone in the development history of Fourier transform, wavelet transform (WT) has been a perfect achievement of functional analysis, Fourier analysis, harmonic analysis, and numerical analysis. WT overcomes many shortcomings of classical Fourier analysis, such as the limitation for analyzing nonstationary signal, the lack of time-frequency positioning ability and thus absence of time-frequency resolution. As such, wavelets have the reputation of “mathematical microscope” which can analyze nonstationary signals and capture many details of signals at multiscale time-frequency resolutions.

The initial concept of wavelet transform was first proposed by French geophysicists, Morlet and his colleagues (Morlet et al. 1982), with incentive to analyze and process seismic reflection signals in oil and gas exploration. They proposed the rudiment of wavelet analysis that uses shape-invariance property to overcome the shortcoming of windowed Fourier transform. Immediately, the powerful numerical analysis ability of Morlet’s method inspired many scholars, including Morlet himself, French theoretical physicist Grossman, French mathematician Meyer, and many others, to promote the rapid development of wavelet transform/analysis (e.g., Daubechies 1988; Grossmann and Morlet 1984;

Lemarié-Rieusset and Meyer 1986; Mallat 1989). In 1986, Meyer and his student Lemarie firstly proposed the fundamental idea of wavelet multiscale analysis. In 1988, Daubechies proposed the smooth orthogonal wavelet with compact support set-Daubechies basis which became one of the most widely used wavelet basis in WT. In 1989, Mallat proposed the concept of multiresolution analysis, the general method to construct orthogonal wavelet basis, and the famous Mallat algorithm of multiscale decomposition based on discrete wavelet transform. Sweldens (1995) developed the “promotion method” to construct the second-generation wavelets and made the construction of wavelets get rid of its conventional dependence on Fourier Transform. The fundamental theory of wavelet transform and analysis have been maturely developed in the last decade of the 20 century (Mallat 1999).

As wavelet analysis have many advantages over Fourier analysis in many aspects, including the time-frequency localization, multiresolution, compression, and decorrelation, its applications have achieved tremendous popularity and success in wide range of scientific fields such as mathematics, physics, chemistry, biology, earth science, and computer science. In the field of earth science, wavelet transform, as an advanced data and signal analysis tool, has remarkable achievements in various case studies, such as signal denoising/filtering and time-frequency analysis in seismic exploration data processing, multiresolution analysis of geophysical field (e.g., gravity and magnetic fields) and remote sensing image, hydrological or geological time series analysis, turbulence analysis, to name but few examples. It is worth mentioning that wavelet analysis has been used as powerful tool for multiscale analysis of fractal/multifractals which show emphasis on scale-invariant properties across scales, such as singularity detection (Chen and Cheng 2016; Mallat and Hwang 1992) and multifractal spectrum analysis (Arneodo et al. 1988; Chen and Cheng 2017). Overall, wavelets are essentially used to study non-stationary processes and signals in geoscience in two ways: (1) as a signal analysis tool to extract multiscale information or components of signals and (2) as a basis for representation or characterization of the signal/process.

Since wavelets is not a new theory or methodology anymore, and the theoretical developments of wavelets have been largely completed over the last two decades of the twentieth century, we will not present a full and in-depth theory description here. Several good handbooks (e.g., Mallat (1999), Percival and Walden (2000)) on wavelet theory are available and many readable papers (e.g., Kumar and Foufoula (1997)) with a good review of wavelet theory have been published. We do however describe some mathematical background of wavelet analysis in this paper, and also present some important applications of using wavelet analysis in geoscience research.

Concepts

Fourier Transform

The Fourier transform of a function $f(x)$ is formulated as

$$F(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(x) e^{-i\omega x} dx, \quad (1)$$

where ω is the angular frequency of the periodic function $e^{-i\omega x}$, and therefore $F(\omega)$ can represent the frequency content of the function $f(x)$. Using the inverse Fourier transform, the original function can be recovered by

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{i\omega x} d\omega. \quad (2)$$

According to the form of Eq. (2), $f(x)$ can be regarded as the weighted sum of the simple waveforms $e^{i\omega x}$, where the weight at particular frequency ω is given by $F(\omega)$.

Although the Fourier transform built a bridge between the time and frequency domains, it has significant limitations in analyzing and processing nonstationary signals in several aspects. Specifically, the Fourier transform integrates the entire time domain and lacks local analysis capabilities; on the other hand, the instantaneous change of the signal (such as singularity) often affects the entire spectrum analysis. The short-time Fourier Transform (STFT) was developed to overcome above issues, and it is defined as (Gabor 1946)

$$STFT(\omega, x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(u) w(u-x) e^{-i\omega u} du, \quad (3)$$

where $w(x)$ is the windowing function (e.g., Gaussian window), ω and u are frequency and translation (time) parameters, respectively. The Fourier transform of the segmented signal through $w(x)$ centered at $x = u$ then provides the position of frequency and extract local-frequency information of signal.

However, STFT cannot address the problem of adaptive adjustment of resolution units in practical applications because of the fixed resolution units defined in Eq. (3). There is an urgent need for the time-frequency window or resolution units that have an adaptive adjustment function when analyzing real signals. In detail, when analyzing the low-frequency components of the signal, the resolution unit or time-frequency window is automatically widened, and the ability of frequency localization becomes relatively weak. On the other hand, when analyzing the high frequency components of the signal, the resolution unit or time-frequency window is automatically narrowed, and the ability of frequency localization is improved. In short, an ideal method of time-frequency localization should have the function of “microscopy” that is scale independent.

Continuous Wavelet Transform

A wavelet is a wave-like oscillation with an amplitude that begins at zero, increases, and then decreases back to zero. A mother wavelet function is defined as

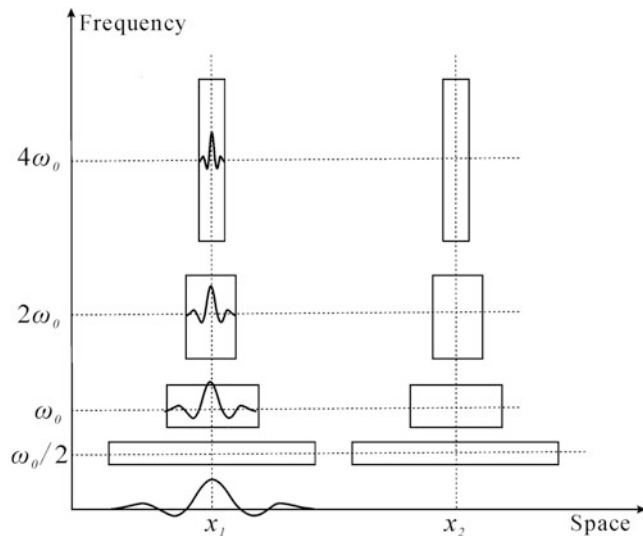
$$\psi(x) = \frac{1}{s^n} \psi\left(\frac{x-b}{s}\right), s > 0, x \in \mathbb{Z}^n, b \in \mathbb{Z}^n \quad (4)$$

where s is the scaling factor and b is a timescale like translation parameter. Note that L^1 – norm ($1/s$) is used in Eq. (4) for normalization. The continuous wavelet transformation (CWT) of a function $f(x) \in L(\mathbb{Z}^n)$ at scale a (related to frequency) and position b is given by (Grossmann and Morlet 1984)

$$Wf(s, b) = \frac{1}{s^n} \int_{-\infty}^{+\infty} f(x) \psi\left(\frac{x-b}{s}\right) dx \quad (5)$$

where $\psi_{s, b}$ defines a family of isotropic wavelets through translation and dilation of the mother wavelet $\psi(x)$. Eq. (5) can be further considered as a convolution product: $Wf(s, b) = f * \psi_s(b)$, and thereby the function $f(x)$ can be decomposed into elementary space–scale contribution by convolving it with a suit of mother wavelets, which are well localized in time and frequency space.

Wavelet transform provides a good solution for addressing the above-mentioned practical problems facing Fourier transform or STFT. Figure 1 shows a tiling of space–frequency plane defined by a 1D wavelet basis, which illustrates the shape of a wavelet resolution adaptively depends on the scale. As the scale decreases, the space support is reduced but the frequency spread increases. It is becoming clear that the



Wavelets in Geosciences, Fig. 1 Multiresolution space-frequency tiling of wavelet analysis, illustrating the shape of a wavelet resolution adaptively depends on scale. (Chen and Cheng 2016)

resolution of wavelet transform can be improved adaptively for analyzing high-frequency signal, while it becomes poor for analyzing low-frequency signal. The zooming capability of the wavelet transform not only can help to locate isolated singularity events, but can also characterize more complex multifractal signals when possessing non-isolated singularities (Mallat 1999; Mallat and Hwang 1992).

Discrete Wavelet Transform

Like the discrete Fourier transform, the discrete wavelet transform (DWT) decomposes the signal into mutually orthogonal set of wavelets using a discrete set of wavelet basis with scales a and translations b . This specificity represents the main difference between CWT and DWT. In practical, the DWT often employs a dyadic grid (Mallat 1999), where the mother wavelet is scaled by power two ($a = 2^j$) and translated by an integer ($b = k2^j$), where k is a location index running from 1 to $2^j N$ (N is the number of observations) and j runs from 0 to J (J is the total number of scales).

Through defining the dyadic mother wavelet as:

$$\psi_{j,k}(x) = 2^{-j/n} \varphi(2^{-j}x - k) \quad (6)$$

we can obtain the DWT coefficients using the following expression

$$W_{j,k} = W(2^j, 2^j k) = 2^{-j/2} \int_{-\infty}^{+\infty} f(x) \psi(2^{-j}x - k) dx \quad (7)$$

On the other hand, the original signal (or its parts) can be reconstructed from the wavelet coefficients by means of the Inverse Discrete Wavelet Transform (IDWT), which is formulated as

$$f(x) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} W_{j,k} \psi_{j,k}(x) \quad (8)$$

The DWT supports different mother wavelets, such as Harr, Daubechies, Biorthogonal, Symlet, Meyer, and Coiflets. The Haar wavelet is the first and simplest mother wavelet, and the Daubechies (Db) wavelet is a family of orthogonal wavelets, where Db1 is similar to Haar wavelet. A proper wavelet should be chosen for real signal analysis, which often depends on the similarity between wavelets and signal.

Applications Review of Wavelets in Geosciences

Wavelet transform or analysis has been widely used in many aspects of geoscience research for the past decades. In this section, we only present several interesting applications of using wavelets in geoscience (e.g., multiscale decomposition,

multifractal analysis, time-frequency analysis, filtering/denoising, etc.) as reviewing all of them in a couple of pages is certainly not possible, and hopefully will be a source of inspiration for new research.

Multiresolution Analysis/Decomposition

An essence of WT is “to see the wood and the trees,” in other words, to obtain multiscale natures of signals through the wavelet-based multiscale decomposition (WMD) algorithm. The classic framework of WMD is introduced by Mallat(1989) to project signal $f(\mathbf{x})$ onto a set of basis $\left\{ \varphi_{J,k}(\mathbf{x}), \left\{ \psi_{J,k}^{(i)}(\mathbf{x}) \right\}_{j \leq J} \right\}, j \in \mathbb{Z}^n, \mathbf{k} \in \mathbb{Z}^n, i = 1, 2, \dots, 2^n - 1$, where

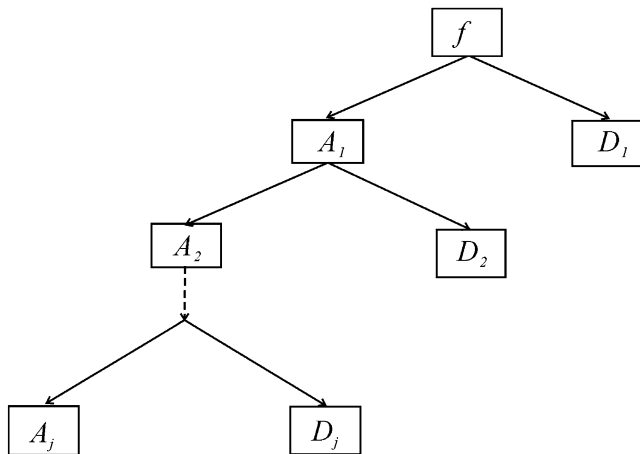
$$\varphi_{J,k}(\mathbf{x}) = 2^{-j/n} \varphi(2^{-j}\mathbf{x} - \mathbf{k}) \quad (9)$$

and

$$\psi_{J,k}^{(i)}(\mathbf{x}) = 2^{-j/n} \psi^{(i)}(2^{-j}\mathbf{x} - \mathbf{k}) \quad (10)$$

are obtained by dilating and translating the mother wavelet $\psi(\mathbf{x})$ and father wavelet (scaling function) $\varphi(\mathbf{x})$ with the special choice $a = 2^j$ and $b = 2^j \mathbf{k}$ for discretization, respectively. The WMD is realized using a simple scheme, the so-called pyramid algorithm (Fig. 2). More details about the WMD algorithm can be found in Mallat (1989). Hence, the signal $f(\mathbf{x})$ can be expressed as

$$f(\mathbf{x}) = \sum_{\mathbf{k} \in \mathbb{Z}^n} A_{\mathbf{k}} \varphi_{J,k}(\mathbf{x}) + \sum_{j=1}^J \sum_{\mathbf{k} \in \mathbb{Z}^n} \sum_i d_{j,k}^{(i)} \psi_{J,k}^{(i)}(\mathbf{x}) \quad (11)$$



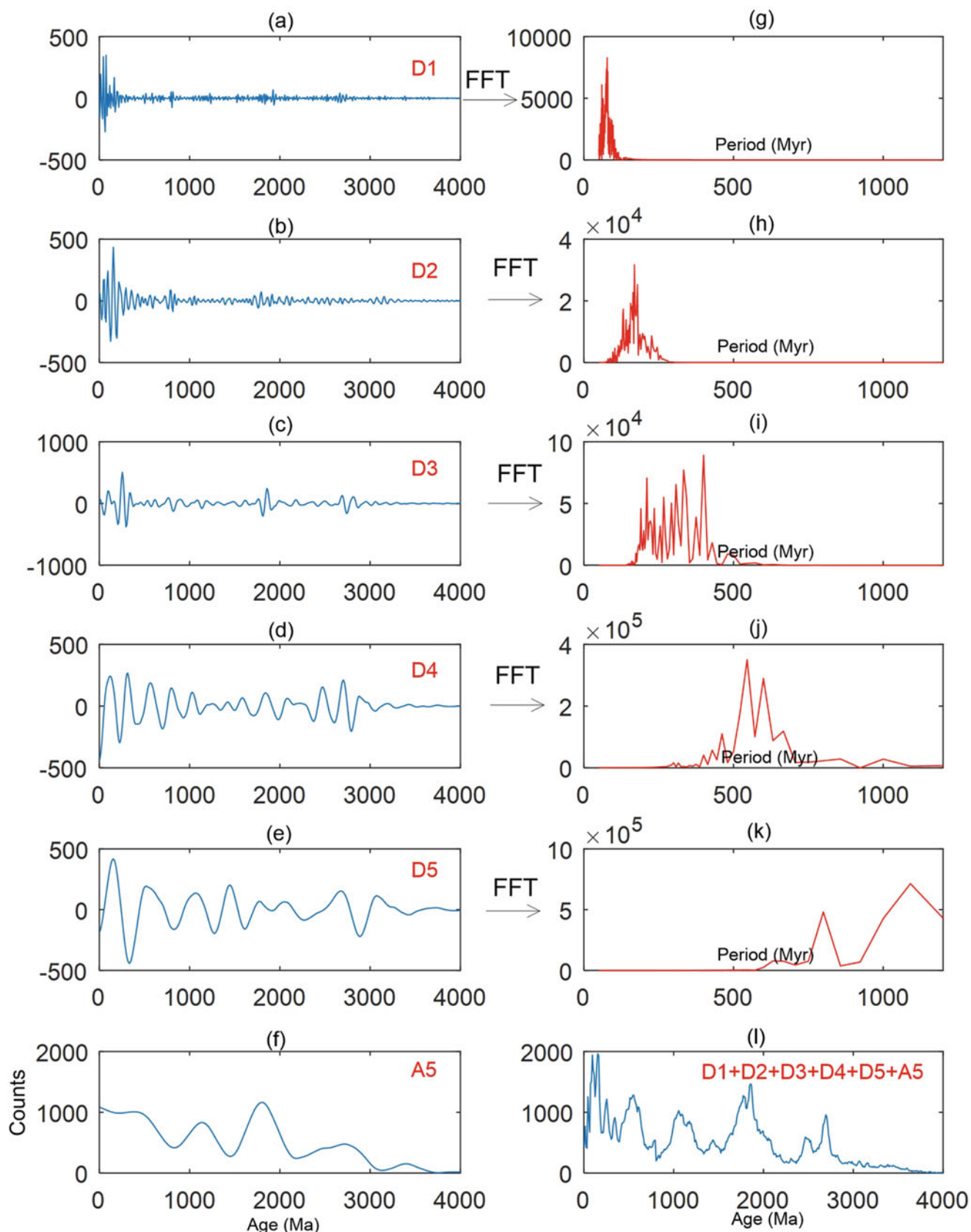
Wavelets in Geosciences, Fig. 2 Pyramid architecture of the multi-resolution analysis. Details and approximations are progressively built in pyramid from up to scale j (Chen and Cheng 2016)

with $A_{\mathbf{k}} = \int f(\mathbf{x}) \varphi_{j,k}(\mathbf{x}) d\mathbf{x}$ and $d_{j,k}^{(i)} = \int f(\mathbf{x}) \psi_{j,k}^{(i)}(\mathbf{x}) d\mathbf{x}$ termed as approximation coefficient and detail/wavelet coefficient, respectively. For instance, the multiscale representation of 2D signal $f(\mathbf{x})$ up to scale J consists of a low-frequency approximation A_J and three high-frequency details $D_{j,k}^{(i)}, i = 1, 2, 3$ in horizontal-, vertical-, and diagonal direction.

In this entry, we present an application of wavelet analysis of geological time series data to determine the cyclicity of Earth's long-term evolution, such as Wilson cycle of plate tectonics - supercontinent assembly and breakup. The zircon U-Pb age peaks and evolved $\delta^{18}\text{O}$ isotopes in deep geological time are often linked to supercontinent assembly. The recently improved global database of U-Pb ages and $\delta^{18}\text{O}$ isotopes from zircon grains allows the time series periodicity analysis by a popular method of wavelet transform (see Chen and Cheng (2018), and references therein). In this section, wavelet-based multiscale decomposition is undertaken to decompose one dimensional (1D) time series signal of U-Pb age distribution into multiscale components including approximations and details. In practical, there exist a variety of wavelet families in WMD algorithm (e.g., Haar, Daubechies (db), symlets (sym), and Biorthogonal, etc.), and a proper wavelet should be determined for processing geochemical time series data. Three criteria may influence the choice of wavelets including number of vanishing moments, support size, and regularity (Mallat 1999). Considering the fact that the practical geochemical time series signal often presents complex or fractal natures, the regular orthonormal wavelets with higher compact support like db8 may be better choices for obtaining reasonable decompositions. Using the five levels in WMD, the raw U-Pb age time series signal was decomposed into multiscale components involving wavelet details at scale $j = 1, 2, 3, 4$, and 5, as shown in Fig. 3a-e, and wavelet approximation at scale $j = 5$ in Fig. 3f. The wavelet approximation signals could be regarded as smoothing (averaging) results of original signal because of the low-pass filtering of father wavelet in WMD. Superposition of Fig. 3a-f time series can recover the raw sedimentary record, while superposition of Fig. 3a-d could construct the new signal without background interference, i.e., detrended time series. Also shown in Fig. 3g-k are Fourier spectrum analysis of each wavelet detail components, and they identified the periodicity of each detail components of U-Pb age distribution, ranging from 100 Myr to 800 Myr, related to global and continental-scale plate tectonics. In addition, WMD is often used to extract multiscale natures (including components and edges) of two-dimensional image or fields (e.g., Chen et al. 2015; Fedi and Quarta 1998).

Fractal/Multifractal Analysis

Wavelet transform acts as a natural tool for investigating the interscale relationship of fractal measures (e.g., scale-free,



Wavelets in Geosciences, Fig. 3 (continued)

self-similarity and singularity) because of the inherent scaling property of wavelet basis (Arneodo et al. 1988). Using wavelet or detail coefficients, the scaling behavior (i.e., Hurst exponent h) of singular increment function: $|f(x + \Delta I) - f(x)| \propto |\Delta I|^{h(x)}$, can be expressed as (Argoul et al. 1989)

$$W_\psi(\lambda s, x) \propto \lambda^{h-n+1} W(s, x) \quad (12)$$

On the other hand, the scaling behavior (i.e., singularity index α) of fractal clustering measure: $\mu(\epsilon) = \int_{B(\epsilon)} d\mu \propto \epsilon^{\alpha(x)}$, can be mirrored by scaling or approximation coefficients as (Chen and Cheng 2016)

$$W_\varphi(\lambda s, x) \propto \lambda^{\alpha-n} W(s, x) \quad (13)$$

where s represents the scale, λ represents the scalar number and n is the dimension of dataset. Notably, within the context of fractals, the Hurst index h characterizes the persistence or intermittency of time series, while the singularity index α characterizes the clustering or accumulation of mass or energy. Therefore, the wavelet transform provides an appropriate framework for both local Hurst detection (LHD) and local singularity analysis (LSA) of fractal signals or fields.

Moving forward, utilizing the inherent scaling property of the wavelet basis, the classic multifractal spectrum analysis based on box counting also can be revisited in wavelet forms. Accordingly, the power-law scaling relation defined in Eq (13) can be written as fractal density model when using approximate coefficients of DWT (Chen and Cheng 2016), namely $A(j) \propto (2^j)^\alpha - n$. As such, the mass-partition function of the q -th moment routinely used for multifractal spectrum calculation can be revisited as (Chen and Cheng 2017)

$$\chi_p(j, q) = \frac{1}{N_j} \sum_{k=1}^{N_j} [A(j, k)]^q, (-\infty < q < +\infty) \quad (14)$$

where N is the total number of scaling coefficients at level j . The DWT is generally implemented using a dyadic decimation (or down-sampling) scheme, the so-called Mallat's algorithm. These functions are built recursively in the space-scale half-plane (i.e., dyadic tree) of the orthogonal wavelet transform (see Fig. 1), cascading from an arbitrary given large scale toward small scales. Above wavelet-based strategy for implementing multifractal spectrum analysis employs the fast multiscale decomposition algorithm and the sparsity of wavelet representation; therefore, it would be less time-consuming

compared to conventional method especially for analyzing large datasets.

Figures 4a, b show the local singularity sequence of the zircon U-Pb age distribution and $\delta^{18}\text{O}$ isotope time series, respectively. Also shown in Figs. 4e, f are the point-wise Hurst sequences. Both indices were estimated using the multi-scale wavelet method. The local singularity indices representing the clustering of zircon ages identify the peaks and troughs in detail, while the point-wise Hurst index characterizes the persistence of adjacent values within the series. Six statistically significant peaks at approximately 4.1, 3.4, 2.6, 1.8, 1.0, and 0.2 Ga were observed in both Fig. 4a, b for the U-Pb age spectra and $\delta^{18}\text{O}$ analyses, and relatively smaller peaks at approximately 3.7, 2.5, 2.1, 1.4, 0.8, 0.6, and 0.1 Ga were detected in the zircon age spectra. In particular, the peak around 2.7-2.5 Ga, which was subdued in the raw $\delta^{18}\text{O}$ pattern was significantly amplified in Fig. 4b. Consequently, the estimated peaks of the age clusters and $\delta^{18}\text{O}$ values and their extents correspond well with supercontinent assembly periods at approximately 2.7-2.5 Ga (Superia and Sclavia), 2.1-1.7 Ga (Nuna), 1.3-0.95 (Rodinia), 0.7-0.18 (Gondwana and Pangea). In addition, the Hurst sequences (Fig. 4e, f) obtained from the zircon age distribution and $\delta^{18}\text{O}$ time series also reveal periodic patterns over 4.4 Gyrs. The higher Hurst values indicate the strong persistence of the time series (for U-Pb age signal it may imply a continuous growth of continent), while the lower values indicate a weaker persistence, where $h = 0$ implies an interrupted time series value. In particular, the timing of valleys within the Hurst sequence mark the beginning and end of supercontinent development well (see Fig. 4e, f). Both the singularity and Hurst series indicate the geological time series have a highly episodic nature from a fractal perspective, but the time series periodicity requires further quantitative evaluation.

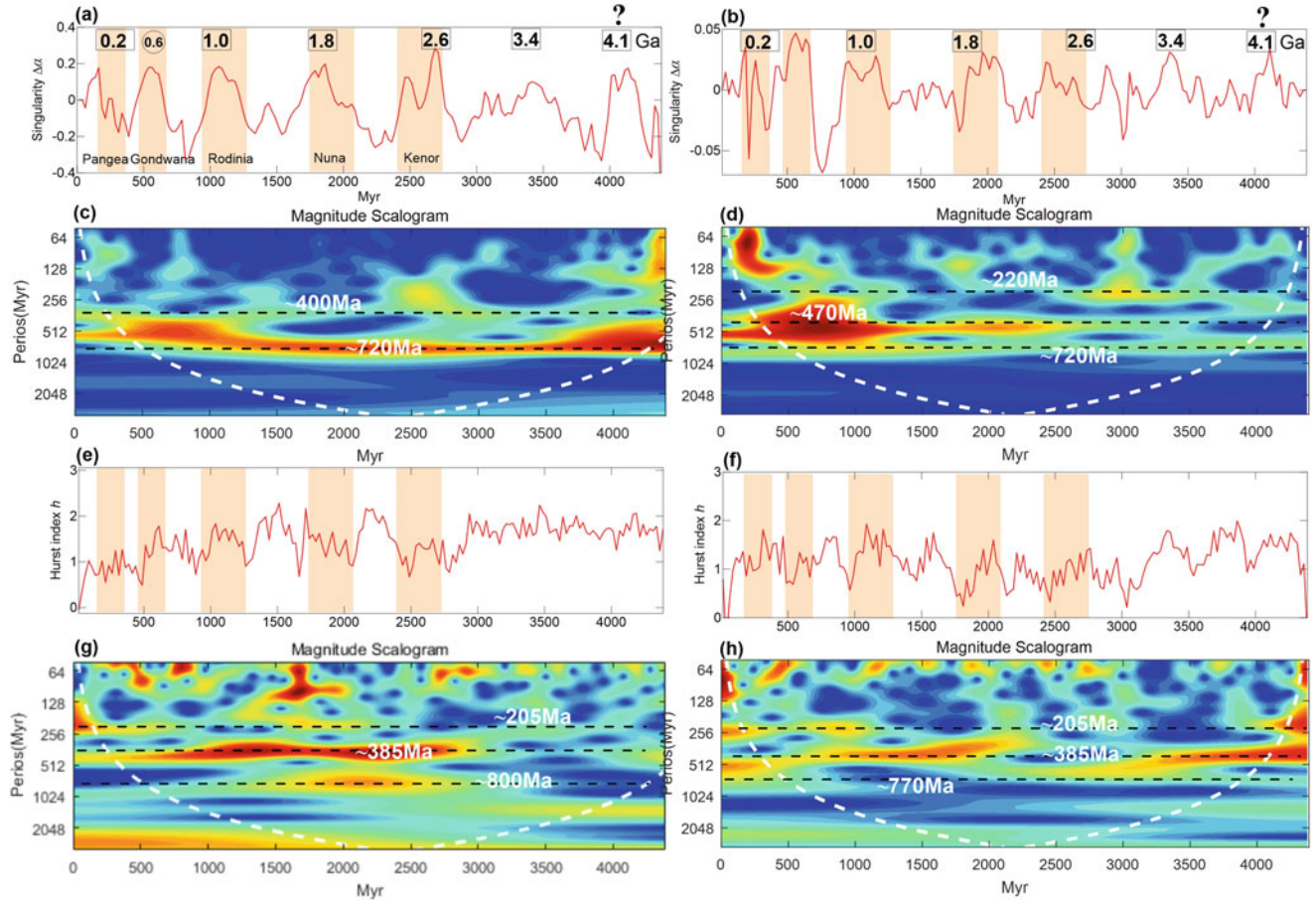
Time-Frequency Analysis

Wavelet analysis allows a rapid detection of periodic patterns and the determination of persistence of cycle across time. According to the form of Eq. (5), a discrete time series $x(t_i)$ of length N with sample spacing Δt can be transformed into wavelet coefficients in both time and frequency domains by

$$W(s) = \frac{\Delta t}{\sqrt{s}} \sum_{j=0}^{N-1} x(t_j) \psi(\Delta t(j-i)/s), \quad (15)$$

Wavelets in Geosciences, Fig. 3 Multiscale decomposition of U-Pb age time series using wavelet-based multiscale decomposition algorithm. (a-e) represent multiscale wavelet details at scale $j = 1, 2, 3, 4, 5$, and (f) represents approximation or residual component.

Superposition of (a-f) time series can recover the raw sedimentary record (L). Also shown in (g-k) are Fourier spectrum analysis of each wavelet detailed components, and this identifies a dominant Wilson cycle of 600-800 Myr with minor cycles (~300, ~200, ~100, Myr)



Wavelets in Geosciences, Fig. 4 Local singularity analysis of (a) zircon U-Pb age database and (b) $\delta^{18}\text{O}$ isotope signal, and point-wise Hurst detection of (e) zircon U-Pb age and (f) $\delta^{18}\text{O}$ series (Chen and Cheng 2018a). Also shown in (c), (d), (g), and (h) are the continuous

wavelet spectrum of the singularity logs and Hurst logs of zircon U-Pb age distribution and $\delta^{18}\text{O}$ signal, respectively. The black dotted line represents the estimates of cyclicity determined on the basis of global wavelet power spectrum

where i represents the position shifting and s is the scale expansion factor. In fact, the above equation can be regarded as an enhanced version of the discrete Fourier transform in Eq. (2): $F(\omega) = \sum x(t_j)e^{i\omega t_j}$. The difference is that the periodic exponential $e^{i\omega t_j}$ is replaced by a wavelet function $\psi(t, s)$, and the advantage of wavelet functions is that they are well localized in both time and frequency space. There exists a series of wavelet function as diverse as Haar, Daubichies, Morlet, Mexican Hat, and so on. Among them, the Morlet wavelet is widely used because of the non-orthogonal and optimal joint time-frequency concentration, and its form is defined as

$$\psi(t) = \pi^{-1/4} e^{i\omega t} e^{-1/2t^2}, \quad (16)$$

where ω is dimensionless frequency. Although WT suffers from the Heisenberg uncertainty principle, the Morlet wavelet is a good choice for achieving a balance between time and frequency.

Wavelet coherence analysis (WCA) (also noted as wavelet cross-correlation analysis) is a method for quantifying similarity or coherency between two nonstationary series by using the localized correlation coefficients in time and frequency space (Torrence and Compo 1997). Given two time series X and Y , firstly, the wavelet cross-correlation transform is defined as

$$W_i^{XY} = W_i^X(s) \cdot W_i^Y(s). \quad (17)$$

Then wavelet coherence can be defined by

$$R^2(s) = |S(W_i^{XY}(s))/s|^2 / |S(|W_i^X(s)|^2/s) \cdot S(|W_i^Y(s)|^2/s)|, \quad (18)$$

in which S is a smoothing operator defined by the wavelet type used. R^2 takes a value between 0 and 1, which resembles the implication of traditional correlation coefficient.

Here, the CWT was used to determine the periodicity of the global zircon age distribution and $\delta^{18}\text{O}$ time series. When processing raw data, the trend in zircon preservation potential interferes with the continuous wavelet spectra which poses the significant problem for periodicity determination (Chen and Cheng 2018a). This tendency has been eliminated by LSA and LHD, as shown in Figs. 4(a) and (b), and the resultant singularity or Hurst consequences representing isolated fluctuations geological time series signals can be employed to explore the underlying periodicity quantitatively. As expected, the application of CWT to singularity and Hurst sequences (Figs. 4c, d, g, and h) yields clearer patterns with persistent power for period estimation. The results reveal three prominent cycles, i.e., ~ 200 Ma, ~ 400 Ma, and ~ 750 Ma for both the U-Pb age distribution and $\delta^{18}\text{O}$ signal. Specifically, the ~ 750 Ma cycle is persistent throughout geological time, indicating a secular change, while the less persistent cycles of ~ 400 Ma and ~ 200 Ma are mainly identified for the period < 3.0 Ga. Furthermore, we utilized the WCA method to investigate the correlation between the identified cycles of the two singularity sequences. The results shown in Fig. 5 indicate significant in-phase coherence for an ~ 750 Ma cycle between the two series throughout Earth's history. The in-phase coherence found here indicates synchronization between processes resulting in age peaks (e.g., intensive magmatism or supercontinent assembly) and $\delta^{18}\text{O}$ highs (e.g., continent reworking). Additionally, in-phase coherence was observed with 400–500 Ma periodicity after 1.5 Ga. The present time series analysis findings suggest that the Earth continent evolved with a main periodicity of ~ 750 Ma along a series of minor cycles (~ 200 and ~ 400 Ma) linked to Wilson cycle of plate tectonics (supercontinent assembly and breakup).

Filtering/Denoising

The compression (or sparsity) property of WT allows that, for most signals, a large fraction of signal energy can be captured by a few very large wavelet coefficients. Motivated by and capitalizing this property, wavelet thresholding filtering is designed by a way of thresholding coefficients via a pre-defined threshold (Donoho and Johnstone 1994). The general process of wavelet thresholding filtering can be expressed as (Chen and Cheng 2018b).

$$W = W(f), W_T = F(W, T), A = W^{-1}(W_T) \quad (19)$$

where W and W^{-1} are the wavelet transform and inverse transform, respectively, F is the thresholding function, and T is the threshold value. More specifically, the above wavelet filtering model can be summarized in three steps: (I) implement DWT on the data and obtain its wavelet coefficients, (II) determine a threshold and perform thresholding

(or shrinkage) on these coefficients, (III) implement the IDWT of these selected coefficients to obtain the filtered signal in the space or time domain. Therefore, the determination of the threshold value and the choice of threshold function are two key steps for wavelet thresholding filtering.

Threshold determination is a very important question as well as a crucial step in wavelet shrinkage. For example, in signal denoising, a small threshold may preserve too much noise, whereas a large threshold leads to a loss of signal, causing artifacts and blurred edges. While the idea of thresholding is simple and effective, finding an optimal threshold is not trivial. Wavelet statistics can provide critic information for addressing above question, and recent study suggests that wavelet coefficients yield fractal /multifractal distribution and offer a new scheme for threshold determination based on different self-similarities (Chen and Cheng 2018b). A vast literature has emerged over decades on signal denoising or image compression using nonlinear techniques, in the setting of additive white Gaussian noise, such as VisuShrink, SureShrink, and BayesShrink. In order to separate coefficients of different class, thresholding functions (e.g., hard-thresholding and soft-thresholding) are often used in wavelet denoising. Hard thresholding sets any coefficients less than the threshold to zero, which can be expressed as:

$$\hat{C} = \begin{cases} C & \text{if } |C| \geq T \\ 0 & \text{if } |C| < T \end{cases} \quad (20)$$

While in soft thresholding, the threshold is subtracted from any coefficient which is larger than the threshold, and the coefficients less than the threshold is threshold to zero. It is expressed as:

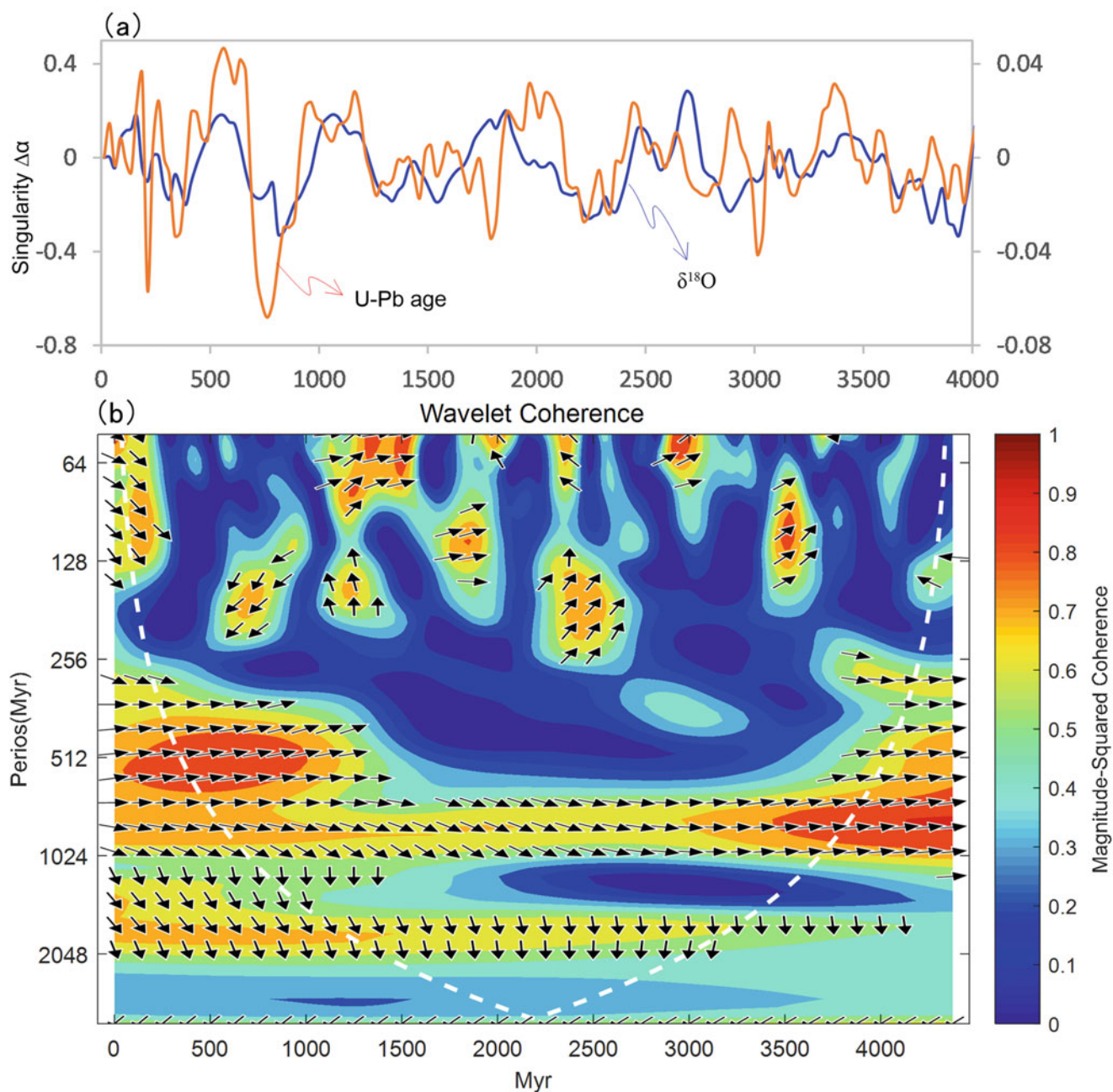
$$\hat{C} = \begin{cases} C - t & \text{if } |C| \geq T \\ 0 & \text{if } |C| < T \end{cases} \quad (21)$$

To avoid the discontinuity caused by using the hard-thresholding and the biased estimation caused by using the soft-thresholding, one could apply nonlinear thresholding given by:

$$\hat{C} = \begin{cases} C & \text{if } |C| \geq T \\ C \sin\left(\frac{\pi(|C| - 0.5T)}{t}\right) & 0.5T \leq \text{if } |C| < T \\ 0 & \text{if } |C| < 0.5T \end{cases} \quad (22)$$

where C is the wavelet coefficient, $\hat{C}$ is after thresholding, and T is the threshold value.

Here we simply present an application of using wavelet denoising to improving the resolution of seismic reflection

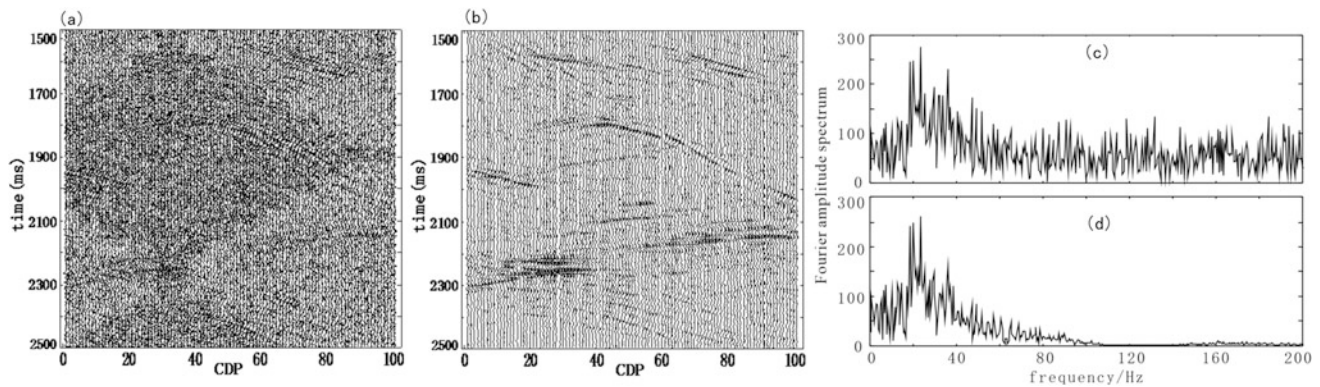


Wavelets in Geosciences, Fig. 5 Wavelet coherence analysis of singularity sequences between zircon U-Pb age distribution and $\delta^{18}\text{O}$ time series (Chen and Cheng 2018a). Right-pointing arrows indicate that the

two signals are in-phase while left-pointing arrows are for anti-phase signals. The white dotted line indicates possible distorted results due to edge effects of the time series

image in oil/gas exploration. Wave field separation is the premise and foundation for high-resolution seismic processing, imaging, inversion of lithology parameters, and attributes analysis. In most land seismic surveys, the complicated surface and subsurface geologic conditions, together with ambient noise interferences, would produce strongly random noise in raw seismic reflection data, leading to the low signal to noise ratio, which could obscure important reflector information. Figure 6a shows the original seismic section from the

survey area, which has strong interference and low signal-to-noise ratio. Figure 6b shows the section processed by using the wavelet denoising. It can be observed that the wavelet thresholding method can reduce noise to some extent. Figure 6a, b show the spectrum analysis results of original data and the wavelet denoised data on the 50th trace. The result indicates that the denoising using wavelet transform can suppress the random noise drastically with the precondition of maintenance for the effective seismic wave. Although the



Wavelets in Geosciences, Fig. 6 Wavelet denoising. (a) The original seismic reflection image, (b) clean seismic image after wavelet denoising, (c) the power spectrum of seismic signal on the 50th trace, and (d) that after wavelet denoising

wavelet transform has been widely used for seismic data denoising, it can only detect in horizontal, vertical, diagonal, and several other limited directions when 2-D wavelet transform is used to process image data; thus most of the linear and even anisotropic characteristics in the signal cannot be well described. Recently, many advanced and powerful wavelet-derived techniques have been developed to improve denoising, e.g., the ridgelet transform, curvelet transform, and contourlet transform.

Others

Wavelet Neural Networks: WNNs are a new class of networks that combine the classic sigmoid neural networks and the wavelet analysis (Zhang et al. 1995). Wavelet transform is good at time-frequency localization analysis, while neural network has good self-learning function and fault-tolerant ability. In WNNs, the classic sigmoidal family of neural networks was replaced by wavelets as activation functions. Therefore, neural networks using wavelets as basis functions are particularly efficient in learning from sparse data, because WNNs adapt the scheme of multi-resolution analysis. Overall, WNNs have more flexible function approximation, stronger model recognition ability and higher tolerance of errors, while it will be less time-consuming compared to the classic ANN. More details of WNN methodologies can be referred to excellent reviews (Alexandridis and Zaprani 2013). WNNs have been used with great success in a wide range of applications in geosciences, including time series prediction, signal classification, and data compression. More recently, the wavelets have been utilized to develop new architecture of deep learning (deep neural network), such as wavelet convolutional neural network, which can capture both spatial and spectral features.

Wavelet Compression: Apart from its original intention of analyzing nonstationary signals, wavelets have been most successful in data and image compression applications (Shapiro 1993). The idea behind wavelet compression is based on the sparsity property of the wavelet coefficients,

and regular signal component can be recovered using a small number of wavelet approximation coefficients (at a suitably chosen level) and small part of the wavelet detail coefficients. In practical, wavelet compression offers an approach that allows reducing the size of data/signal while at the same time improving its quality through denoising the high-frequency components.

Conclusions

Wavelet analysis has enjoyed a tremendous success in geoscience research over the last decades, as wavelets provide a very simple and efficient mean of analyzing nonstationary signals that are often encountered in wide range disciplines of geosciences. So what is next? Theoretically, wavelets are not optimal for nonpoint singularity in higher dimensions since they have isotropic properties and only provide optimal approximations for (zero dimensional) point singularity. In the last decade, multiscale geometric analysis, as the development of wavelet theory, can deal with the singularity and anisotropy of high-dimensional data more effectively. Up to now, there are many new achievements in multiscale geometric analysis, including ridgelet transform (Do and Vetterli 2003), curvelet transform (Starck et al. 2002), contourlet transform (Do and Vetterli 2005), and shearlet transform (Easley et al. 2008), to name but few examples. All in all, considering the endless variety of nonstationary and/or anisotropic signals and processes that are commonly encountered in engineering and nature science, it is not difficult to estimate that there will be wider search for new fields of wavelet theories and applications.

Bibliography

Alexandridis AK, Zaprani AD (2013) Wavelet neural networks: A practical guide. *Neural Networks* 45(42):1–27

- Argoul F, Arneodo A, Elezgaray J, Grasseau G, Murenzi R (1989) Wavelet transform of fractal aggregates. *Phys Lett A* 135:327–336
- Arneodo A, Grasseau G, Holschneider M (1988) Wavelet transform of multifractals. *Phys Rev Lett* 61:2281
- Chen G, Cheng Q (2016) Singularity analysis based on wavelet transform of fractal measures for identifying geochemical anomaly in mineral exploration. *Comput Geosci* 87:56–66
- Chen G, Cheng Q (2017) Fractal density modeling of crustal heterogeneity from the KTB deep hole. *J Geophys Res Solid Earth* 122: 1919–1933
- Chen G, Cheng Q (2018a) Cyclicity and persistence of Earth's evolution over time: wavelet and fractal analysis. *Geophys Res Lett* 45: 8223–8230
- Chen G, Cheng Q (2018b) Fractal-based wavelet filter for separating geophysical or geochemical anomalies from background. *Mathematical Geosciences* 50(3):249–272
- Chen G, Liu T, Sun J, (2015) Gravity method for investigating the geological structures associated with W–Sn polymetallic deposits in the Nanling Range, China. *Journal of Applied Geophysics* 120:14–25
- Daubechies I (1988) Orthonormal bases of compactly supported wavelets. *Commun Pure Appl Math* 41:909–996
- Do MN, Vetterli M (2003) The finite ridgelet transform for image representation[J]. *IEEE Trans Image Process* 12(1):16–28
- Do MN, Vetterli M (2005) The contourlet transform: an efficient directional multiresolution image representation. *IEEE Trans Image Process* 14(12):2091–2106
- Donoho DL, Johnstone JM (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81:425–455
- Easley G, Labate D, Lim WQ (2008) Sparse directional image representations using the discrete shearlet transform. *Appl Comput Harmon Anal* 25(1):25–46
- Fedi M, Quarta T (1998) Wavelet analysis for the regional-residual and local separation of potential field anomalies. *Geophys Prospect* 46 (5):507–525
- Gabor D (1946) Theory of communication. Part 1: the analysis of information. *J Instit Elect Eng Part III Radio Commun Eng* 93: 429–441
- Grossmann A, Morlet J (1984) Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM J Math Anal* 15: 723–736
- Kumar P, Foufoula E (1997) Wavelet analysis for geophysical applications. *Rev Geophys* 35:385–412
- Lemarié-Rieusset PG, Meyer Y (1986) Ondelettes et bases hilbertiennes. *Revista Matematica Iberoamericana* 2:1–18
- Mallat S (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans Pattern Anal Mach Intell* 11: 674–693
- Mallat S (1999) A wavelet tour of signal processing. Academic, San Diego
- Mallat S, Hwang WL (1992) Singularity detection and processing with wavelets. *IEEE Trans Inf Theory* 38:617–643
- Morlet J, Arens G, Fourgeau E, Glard D (1982) Wave propagation and sampling theory-Part I: complex signal and scattering in multilayered media. *Geophysics* 47:203–221
- Percival DB, Walden AT (2000) Wavelet methods for time series analysis. Cambridge university press, Cambridge
- Shapiro JM (1993) Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans Signal Process* 41:3445–3462
- Starck JL, Candès EJ, Donoho DL (2002) The curvelet transform for image denoising. *IEEE Trans Image Process* 11(6):670–684
- Sweldens W (1995) Lifting scheme: a new philosophy in biorthogonal wavelet constructions. SPIE's 1995 international symposium on optical science, engineering, and instrumentation. International Society for Optics and Photonics, pp 68–79
- Torrence C, Compo GP (1997) A practical guide to wavelet analysis. *Bull Am Meteorol Soc* 79:61–78
- Zhang J, Walter GG, Miao Y, Lee WNW (1995) Wavelet neural networks for function learning. *IEEE Trans Signal Process* 43:1485–1497

Wave-Number Filtering

Sid-Ali Ouadfeul¹ and Leila Aliouane²

¹Algerian Petroleum Institute, Sonatrach, Boumerdes, Algeria

²LABOPHT, Faculty of Hydrocarbons and Chemistry, University of Boumerdes, Boumerdes, Algeria

Definition of the Wave-Number

In physics, the wave-number also known as propagation number or angular wave-number is defined as the number of wavelengths per unit distance the spatial wave frequency and is known as spatial frequency. It is a **scalar quantity** represented by k and the mathematical representation is given as follows:

$$k = 1/\lambda$$

where

- k is the wave-number
- λ is the wavelength

Wave-number equation is mathematically expressed as the number of the complete cycle of a wave over its wavelength, given by –

$$k = 1/\lambda$$

where

- k is the wave-number, and its unit is (1/m)
- λ is the wavelength of the wave in meter

The 2D Wave-Number

In the space (x,y) , the vector wave-number (k) has two components K_x and K_y :

$$k = k_x u_{k_x} + k_y u_{k_y} \quad (1)$$

where u_{k_x} and u_{k_y} are the vector units in k_x and k_y directions.

The modulus of the vector wave-number is

$$|k| = \sqrt{k_x^2 + k_y^2} \quad (2)$$

Wave-Number Calculation

Let $S(x)$, $x \in \mathbf{R}$ is signal, x presents the space (abscissa), $S(x)$ content on wave-number $F(k)$ is calculated using the 1D Fourier transform:

$$F(k) = \int_{-\infty}^{+\infty} S(x)e^{-2\pi j k x} dx = A(k) + jB(K) \quad (3)$$

The spectrum of amplitude is:

$$|F(k)| = \sqrt{A(k)^2 + B(k)^2} \quad (4)$$

In the 2D domain, the wave-number content is calculated using the 2D Fourier transform, let $S(x, y) \in \mathbf{R}$ a signal; x, y present the space, the wave-number content is obtained using the 2D Fourier transform:

$$\begin{aligned} \mathbf{F}(k_x, k_y) &= \iint_{-\infty}^{+\infty} s(x, y)e^{-2\pi j x k_x - 2\pi j y k_y} dx dy \\ &= A(k_x, k_y) + jB(k_x, k_y) \end{aligned} \quad (5)$$

The spectrum of amplitude is the modulus of $\mathbf{F}(k_x, k_y)$:

$$|F(k_x, k_y)| = \sqrt{A(k_x, k_y)^2 + B(k_x, k_y)^2} \quad (6)$$

Wave-Number Filtering

The wave-number filtering of a given signal $S(x)$, $x \in \mathbf{R}^n$, $n = 1, 2$ is an elimination of wave-number components of the signal; this can be done by a multiplication of the signal in the wave-number domain by a transfer function $F(k)$, $k \in \mathbf{R}^n$, $n = 1, 2$. In the space domain, the wave-number filtering is a convolution of the signal $S(x)$ with the transfer function $F(x)$.

Let $F(x)$, $x \in \mathbf{R}^n$, $n = 1, 2$ the transfer function of the filter, the wave-number filtering is:

$$FS(x) = \iint_{\mathbf{R}^n} S(\theta)F(x - \theta)d\theta \quad (7)$$

Where $\theta \in \mathbf{R}^n$, $n = 1, 2$.

$FS(x)$: is the filtered $S(x)$

In the wave-number domain, the spectrum of the filtered $FS(x)$, $FFS(K)$ is given as follows:

$$FFS(K) = FT[S(x)]^* FT[F(x)] \quad (8)$$

Where FT is the Fourier transform.

*Denotes the product.

Examples of Wave-Number Filter

The Butterworth Filter

The Butterworth filter is a type of signal processing filter designed to have a frequency or wave-number response as flat as possible in the passband. It is also referred to as a maximally flat magnitude filter. It was first described in 1930 by the British engineer and physicist Stephen Butterworth in his paper entitled “On the Theory of Filter Amplifiers” (Weinberg and Slepian 1960).

$$G(k) = \frac{1}{\left(1 + \left(\frac{k}{k_c}\right)^n\right)} \quad (9)$$

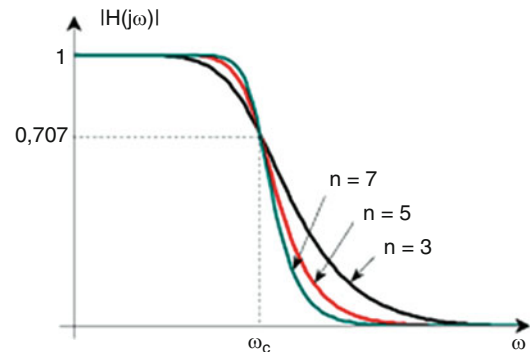
n is the number of poles in the filter.

k_c : is cutoff wave-number.

Figure 1 shows the graph of the transfer function of the Butterworth with different number of poles.

The Chebyshev Filters

Chebyshev filters are analog or digital filters having a steeper roll-off than Butterworth filters and have passband ripple (type I) or stopband ripple (type II). Chebyshev filters have the property that they minimize the error between the idealized and the actual filter characteristic over the range of the filter, but with ripples in the passband. This type of filter is named after Chebyshev because its mathematical



Wave-Number Filtering, Fig. 1 Graph of the transfer function of the Butterworth filter with the number of poles ($n = 3, 5$, and 7)

characteristics are derived from Chebyshev polynomials. Type I Chebyshev filters are usually referred to as “Chebyshev filters,” while type II filters are usually called “inverse Chebyshev filters” (Weinberg and Slepian 1960).

The bandpass ripple transfer function is (Daniels 1974):

$$G(k) = \frac{1}{\sqrt{1 + \epsilon^2 T_n^2\left(\frac{k}{k_0}\right)}} \quad (10)$$

where ϵ is the ripple factor, k_0 is the wave-number cutoff, and T_n is a Chebyshev of the n order.

The transfer function of inverse Chebyshev filters is (Williams and Taylors 1988) (Figs. 2 and 3):

$$G(k) = \frac{1}{\sqrt{1 + \epsilon^{-2} T_n^{-2}\left(\frac{k}{k_0}\right)}} \quad (11)$$

The Gaussian Filter

A Gaussian filter is a ► [filter](#) whose impulse response is a [Gaussian function](#) (or an approximation to it, since a true Gaussian response would have [infinite impulse response](#)). Gaussian filters have the properties of having no [overshoot](#) to a step function input while minimizing the rise and fall time. This behavior is closely connected to the fact that the Gaussian filter has the minimum possible [group delay](#) (Lutovac et al. 2001).

$$G(x) = \frac{1}{\sigma\sqrt{2}} e^{-\left(\frac{x}{\sqrt{2}\sigma}\right)^2} \quad (12)$$

In the wave-number domain, the transfer function of the Gaussian Filter is:

$$FG(k) = e^{-\left(\frac{k}{\sqrt{2}}\right)^2} \quad (13)$$

where σ is the standard deviation.

In the 2D domain the Gaussian filter is defined as (Smith 2002):

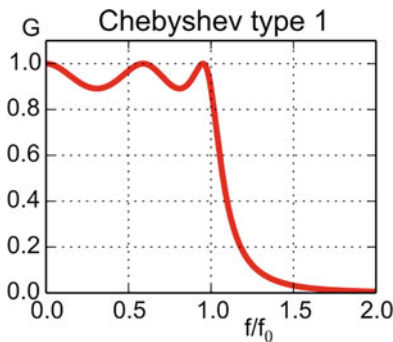
$$G(x, y) = \frac{1}{\sigma\sqrt{2}} e^{-\left(\frac{x+y}{\sqrt{2}\sigma}\right)^2} \quad (14)$$

In the wave-number domain, the transfer function of the Gaussian Filter is (Fig. 4):

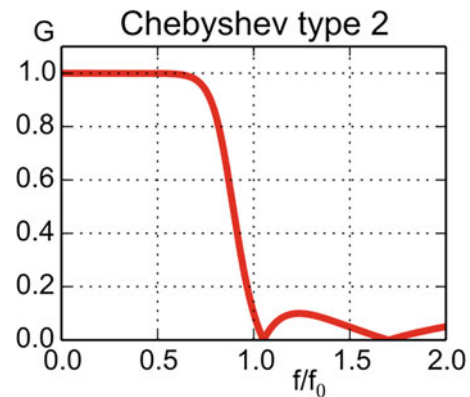
$$FG(k_x + k_y) = e^{-\left(\frac{k_x + k_y}{\sqrt{2}}\right)^2} \quad (15)$$

Application to Gravity Data

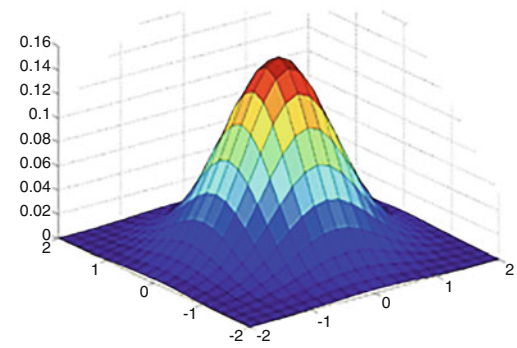
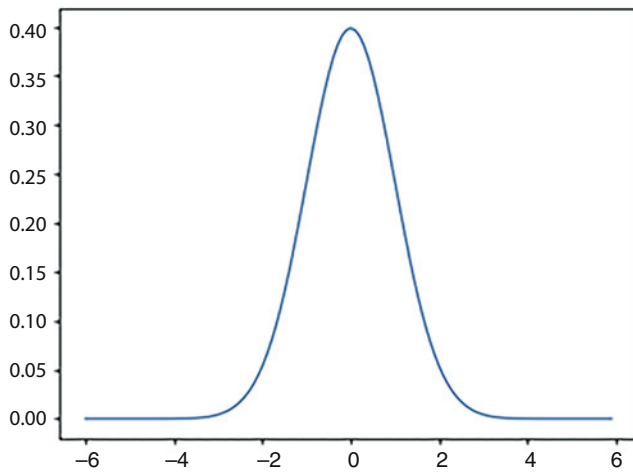
In this section, we show an example of 1D gravity Bouguer anomaly data filtering in the wave-number domain using a Gaussian filter. Figure 5 shows the graph of these data versus the longitude; the gravity profile is located in Algeria, and data of the International Gravimetric Bureau (BGI, <https://bgi.obs-mip.fr/>) are used. The 1D Fourier transform is used to calculate the wave-number content of the signal. Figure 6 shows the graph of the spectrum of amplitude versus the wave-number, while Fig. 7 shows the graph of the spectrum of amplitude in the wave-number interval $[0, 3 \text{ deg}^{-1}]$. We observe two slopes: the first is in the wave-number interval $[0, 0.3 \text{ deg}^{-1}]$, while the second is in the wave-number interval $[0.3-10 \text{ deg}^{-1}]$. The first Bouguer anomaly is called



Wave-Number Filtering, Fig. 2 Transfer function of the Chebyshev filter (Type I)



Wave-Number Filtering, Fig. 3 Transfer function of the Chebyshev filter (Type II)

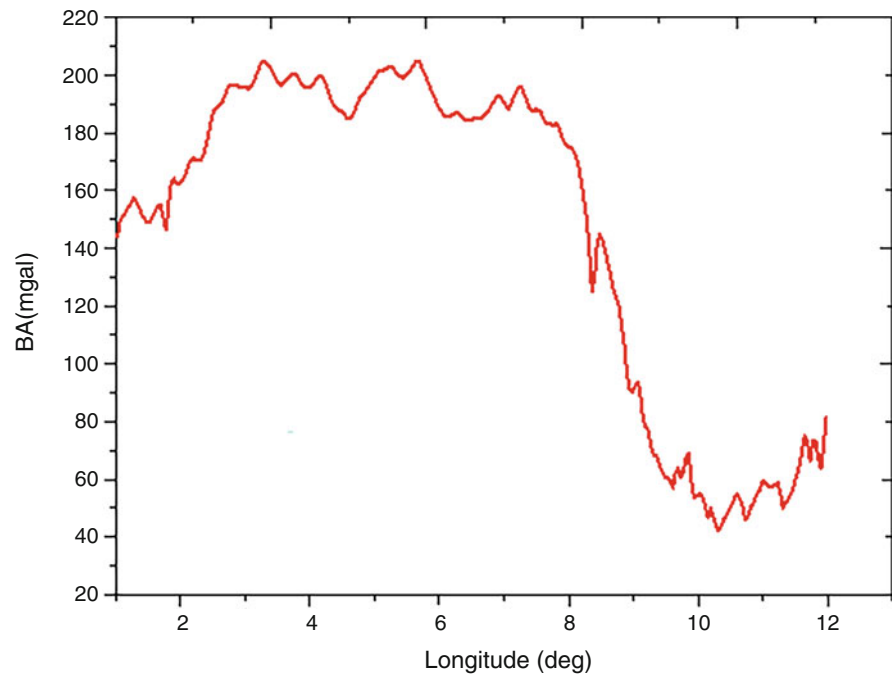


Gaussian kernel

$$G_{\sigma} = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}}$$

Wave-Number Filtering, Fig. 4 Transfer function of the Gaussian, left (1D) and right (2D) (Smith 2002)

Wave-Number Filtering, Fig. 5 Graph of the gravity Bouguer Anomaly (BA) of a profile located in Algeria



the regional anomaly; it is due to deep causative sources and is characterized by low wave-numbers, while the second is called the residual anomaly, which is due to shallow causative sources and is characterized by high wave-numbers.

A Gaussian low filter with cutoff wave-number $k_c = 0.3 \text{ deg}^{-1}$ used to attenuate high wave-number content is the Bouguer gravity anomaly. The filtered anomaly is presented in Fig. 8; we observe that the obtained anomaly is

smoothed, which demonstrates the absence of the high wave-number content in the signal.

Figure 9 shows the graph of the spectrum of amplitude in the wave-number interval $[0, 3 \text{ deg}^{-1}]$; it is clear that the spectrum of amplitude has no content of wave-numbers higher than 0.3 deg^{-1} , which demonstrates the friability of the filter.

Wave-Number Filtering,

Fig. 6 Graph of the spectrum of amplitude of the Bouguer gravity anomaly shown in Fig. 5

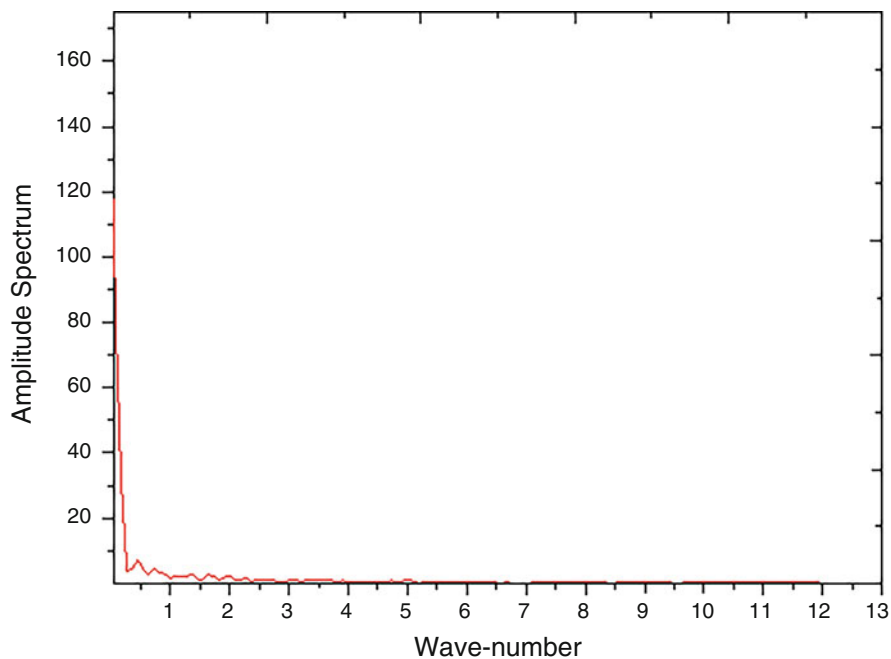
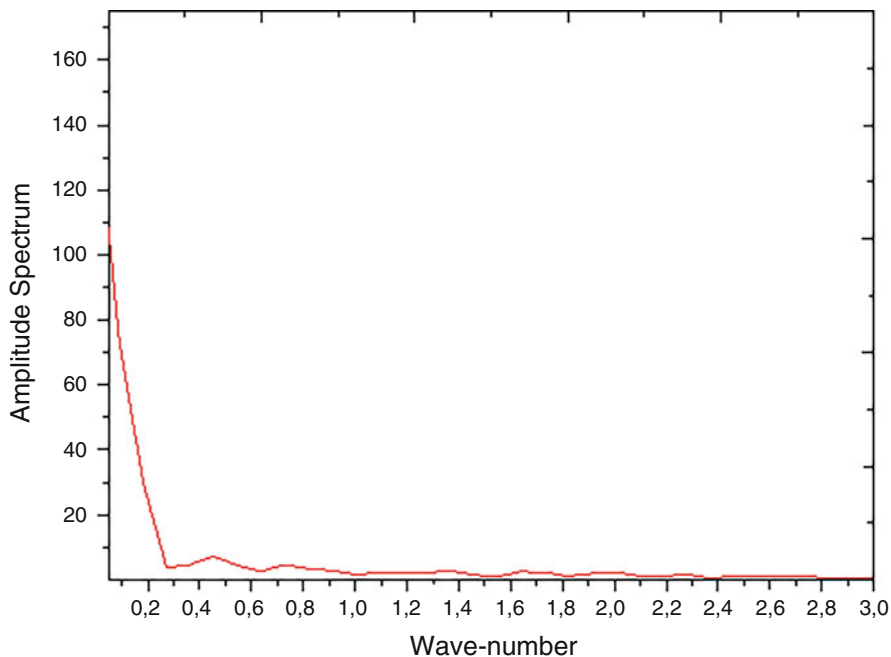
**Wave-Number Filtering,**

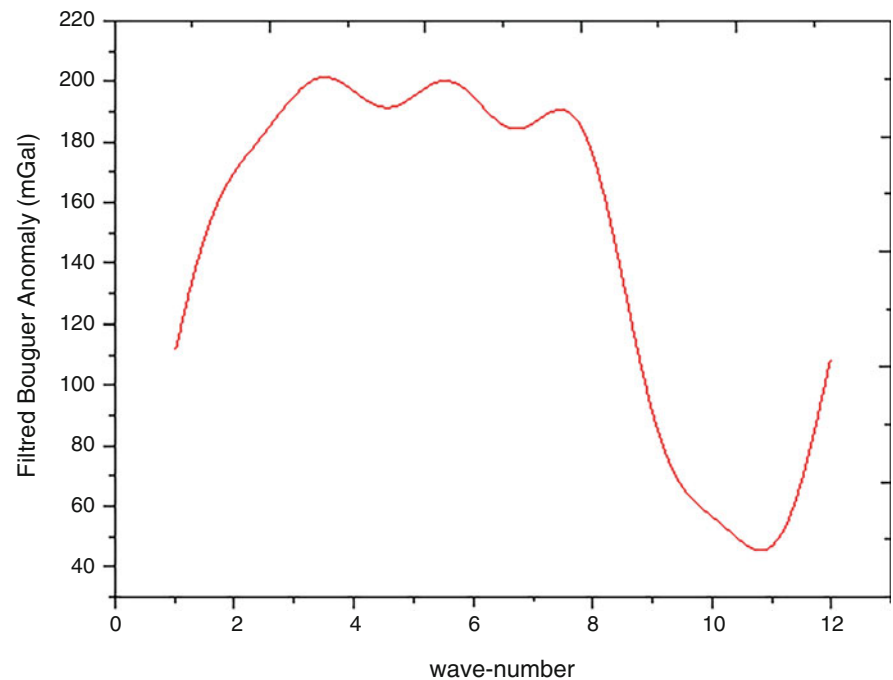
Fig. 7 Graph of the spectrum of amplitude of the Bouguer gravity anomaly presented in the wave-number interval $[0, 3 \text{ deg}^{-1}]$

**Conclusions**

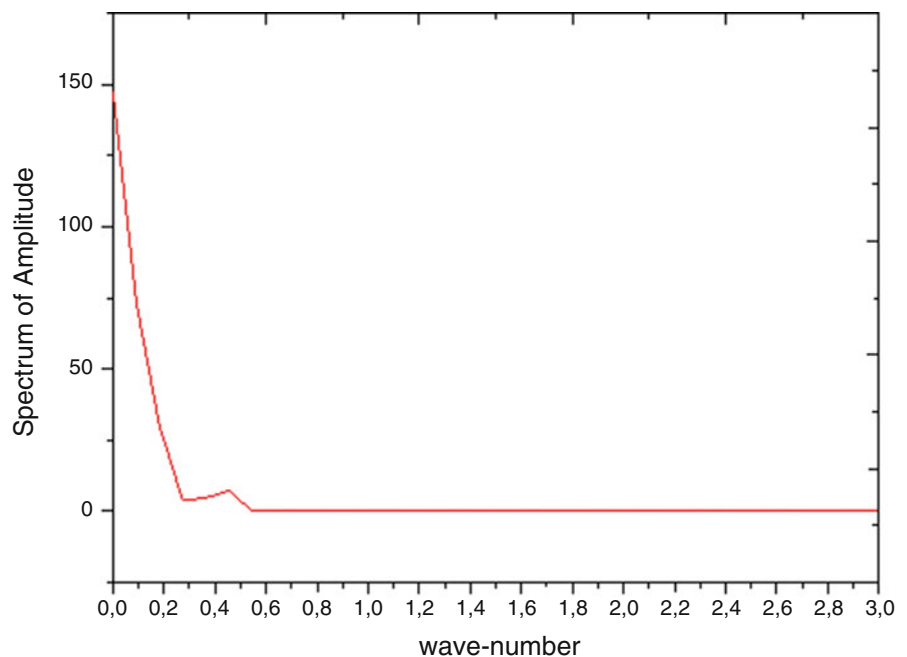
Wave-number known as propagation number or angular wave-number is defined as the number of wavelengths per unit distance; 1D and 2D Fourier transforms are used to calculate the wave-number content of a given signal. Many

filters can be used for wave-number filtering, which consists of attenuating some wave-number components of a signal. In this chapter, we show an example of wave-number filtering of gravity Bouguer anomaly data; the filtering can be used to separate between shallow and deep causative sources.

Wave-Number Filtering,
Fig. 8 Graph of the filtered
 gravity Bouguer anomaly



Wave-Number Filtering,
Fig. 9 Graph of the spectrum of
 the filtered gravity Bouguer
 anomaly



Bibliography

- Daniels R-W (1974) Approximation methods for electronic filter design. McGraw-Hill, New York. ISBN 0-07-015308-6
- Lutovac M, D. et al (2001) Filter design for signal processing. Prentice Hall
- Smith SW (2002) Digital signal processing. <https://doi.org/10.1016/B978-0-7506-7444-7.X5036-5>
- Weinberg L, Slepian P (1960) Takahasi's results on Tchebycheff and Butterworth ladder networks. IRE Trans Circuit Theory 7(2):88–101. <https://doi.org/10.1109/TCT.1960.1086643>
- Williams A-B, Taylors FJ (1988) Electronic filter design handbook. McGraw-Hill, New York. ISBN 0-07-070434-1

Whitten, Eric Harold Timothy

John Cubitt
Wrexham, UK



Fig. 1 Eric Harold Timothy Whitten, courtesy of Prof. Whitten

Biography

Eric Whitten, known to his colleagues as Tim, started his career as a hard-rock geologist, specializing in the tectonics, petrology, petrography, and geochemistry of granite batholiths. Out of the mass of data generated in his studies arose an interest in the use of statistics and 3-dimensional modeling for analyzing and displaying geological data. Following the introduction of computers in the late 1950s, Tim saw the significant potential of this modern technology and he became one of the founding and preeminent specialists in the application of computers to the Earth Sciences. These two topics then formed the main strands of his subsequent multi-faceted career.

Tim was born on 26 July 1927 in Ilford and spent most of his formative years in SE England. After his schooling, Tim worked as a clerk in London during WWII. He then went to University and earned a BSc in Geology from Kings College, London, in 1948 and a PhD in Geology from the University of London in 1952. Subsequently, he was awarded a DSc in Geology from the same university in 1968.

His career in the geological sciences started as a Lecturer in Geology at Queen Mary College, London (1948–1958) and, after emigration to the USA, continued as Professor of Geology at Northwestern University in Illinois (1958–1981) where he was Chairman of the Department from 1977 to 1981.

Moving to the Michigan Technological University in 1981, he joined the administrative arena as, Vice-President for Academic Affairs and Professor of Geology, from 1981 to 1988, and, after the role was renamed, Provost from 1988 to 1990.

He moved back to the UK in 1990 to the village of Widecombe-in-the-Moor on the granite batholith that forms the National Park of Dartmoor in the County of Devon. Here, apart from undertaking a few projects and writing technical articles, he has indulged his interests in sailing, ancient map collecting, and conducting research on local history, especially of Widecombe and Devonshire. Notably, Tim has been Treasurer of Widecombe Fair Committee Co. Ltd., for 14 years (on their Board for 25 years), Commodore of the Royal Western Yacht Club of England, and recently Vice-Chair of the Widecombe History Group.

By the time of his retirement, Tim was widely recognized as an authority on granite batholiths and the application of statistics and 3-dimensional visualization to the Earth Sciences. As a result, he has received a number of medals and awards including the Daniel Pidgeon Fund of the Geological Society of London in 1954, the William Christian Krumbein Medal of the International Association for Mathematical Geology in 1989 and the Geologists Association's Fullerton Award in 1996.

He was instrumental in the establishment of the International Association of Mathematical Geologists, as a Founding Member in 1968, and played an active role for many years including Secretary-General from 1976 to 1980, and President, from 1980 to 1984.

He has published prolifically with 114 professional contributions in the form of books, journal articles, abstracts, and reports.

Tim was married and has five children. Sadly, his wife died in 2019.

Zadeh, Lotfi A.

Behnam Sadeghi
EarthByte Group, School of Geosciences, University of
Sydney, Sydney, NSW, Australia
Earth and Sustainability Science Research Centre, University
of New South Wales, Sydney, NSW, Australia



Fig. 1 Lotfi A. Zadeh. (From IEEE Engineering and Technology History Wiki (ETHW) – Copyright held by IEEE)

Biography

Lotfi A. Zadeh was an electrical engineer, well known for his seminal fuzzy sets theory (Zadeh 1965, see the entry “Fuzzy Set theory in Geosciences”) and fuzzy logic (Zadeh 1975). Zadeh was born on 4 February 1921, in Baku, Azerbaijan. In 1931, his family moved to Tehran, Iran, his father’s birthplace. After graduating from Alborz College, one of the renowned colleges in Iran, he joined the University of Tehran (UT). A year after he graduated from UT as an Electrical Engineer in 1942, he immigrated to the US, and in 1944 he entered the Massachusetts Institute of Technology (MIT). Two years later, in 1946, he received his M.S. degree in Electrical Engineering and then moved to the Columbia University in New York City as a Ph.D. candidate as well as an instructor in Electrical Engineering (Trillas 2011). He would teach several courses such as electromagnetic theory, system theory, information theory, circuit analysis, and sequential machines (IEEE History Center).

In his Ph.D. dissertation, he worked on a novel approach in frequency analysis of time-varying networks (IEEE History Center). In 1949, after graduation and receiving his Ph.D. in Electrical Engineering, he started a new position as an Assistant Professor at the same university and became an Associate Professor in 1953, then a Full Professor, in 1957. The years 1950 and 1952 are significant in his early scientific life, since he published two seminal papers with Prof. J. R. Ragazzini on Wiener’s theory of prediction, and sampled-data systems, respectively.

In 1959, Zadeh joined the University of California, Berkeley as a Professor at the faculty of Electrical Engineering Department. During this time, he mainly focused on linear systems and automata theory. In 1963 when he was named the chair of the department, the name of the department was changed to Electrical Engineering and Computer Sciences

(EECS). In 1991, when he became a Professor Emeritus, Zadeh continued his teaching and research activities and also served as the director of Berkeley Initiative in Soft Computing (BISC). Simultaneously, he held several other visiting positions at renowned research institutes such as MIT, Princeton, IBM Research Laboratory, and Stanford University (Zadeh's Berkely | EECS faculty page).

As a Fellow of the IEEE, AAAS, ACM, AAAI, and IFSA, Zadeh received many prestigious awards such as the IEEE Education Medal, the IEEE Richard W. Hamming Medal, the IEEE Medal of Honor, the IEEE Millennium Medal, the ASME Rufus Oldenburger Medal, the B. Bolzano Medal of the Czech Academy of Sciences, the Kampe de Fariet Medal, the Nicolaus Copernicus Medal of the Polish Academy of Sciences, the Egleston Medal, the Franklin Institute Medal, and the Silicon Valley Engineering Hall of Fame, in addition to 24 honorary doctorates (Zadeh's Berkely | EECS faculty page).

Zadeh was serving on 70 journals' editorial boards. He also published a large number of peer-reviewed papers, including over 200 single-authored papers, in a variety of disciplines in information/intelligent systems. His main fields of interest were systems analysis, decision analysis and information systems, possibility theory, computational theory of perceptions, and computing with words.

He published his groundbreaking theory on fuzzy sets in 1965 (Zadeh 1965), and afterward, he proposed the fuzzy logic (Zadeh 1975, 1996, 2004). Fuzzy logic has been extensively applied in a variety of fields even in geosciences such as mineral prospectivity mapping (Sadeghi and Khalajmasoumi 2015, see the entry "Fuzzy Set theory in Geosciences").

Zadeh passed away in Berkeley, California on 6 September 2017, and was buried in Baku.

Cross-References

► [Fuzzy Set Theory in Geosciences](#)

Bibliography

- IEEE History Center. https://ethw.org/Lotfi_A._Zadeh
- Sadeghi B, Khalajmasoumi M (2015) A futuristic review for evaluation of geothermal potentials using fuzzy logic and binary index overlay in GIS environment. *Renew Sust Energ Rev* 43C:818–831
- Trillas E (2011) Lotfi A. Zadeh: on the man and his work. *Sci Iranica* 18(3):574–579
- Zadeh LA (1965) Fuzzy sets. *Inf Control* 8:338–353
- Zadeh LA (1975) Fuzzy logic and approximate reasoning. *D Reidel Publ Co* 30(i):407–428
- Zadeh LA (1996) Fuzzy logic=computing with words. *IEEE T Fuzzy Syst* 4(2):103–111

Zadeh LA (2004) Fuzzy logic systems: origin, concepts, and trends. *Science* 80:16–18

Zadeh LA. Berkeley Electrical Engineering and Computer Sciences faculty webpage. <https://www2.eecs.berkeley.edu/Faculty/Homepages/zadeh.html/>

Z-Transform

Sathishkumar Samiappan¹, Rajendra Mohan Panda¹ and B. S. Daya Sagar²

¹Geosystems Research Institute, Mississippi State University, Starkville, MS, USA

²Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, Karnataka, India

Definition

The Z-Transform is a technique commonly used to convert a discrete-time (DT) signal into its frequency domain representation. The Z-transform is useful to analyze discrete systems and their stability.

Z-Transform of a DT signal $x[n]$ is defined as

$$X(z) = Z\{x[n]\} = \sum_{n=-\infty}^{\infty} x[n] z^{-n} \quad (1)$$

Inverse Z-Transform

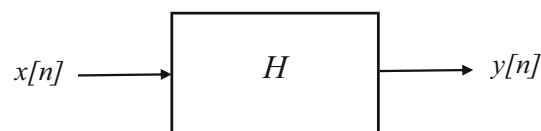
$$x[n] = Z^{-1}\{X(z)\} = \frac{1}{2\pi j} \oint_C X(z) z^{n-1} dz \quad (2)$$

where n is an integer, and z is a complex number represented by $Ae^{j\theta}$.

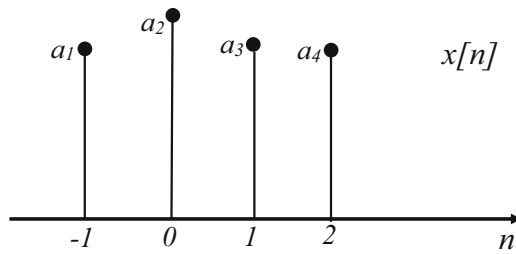
Introduction

Let us consider a discrete-time system with a transfer function H that takes a series of numbers as inputs ($x[n]$), one at a time. Based on the values of $x[n]$, H produces a series of numbers as output ($y[n]$) based on current and past inputs (Fig. 1).

Systems such as H are common in many areas such as signal processing and population dynamics. Using the



Z-Transform, Fig. 1 A typical discrete-time system



Z-Transform, Fig. 2 Plot of a DT signal $x[n]$

Z-Transform, $y[n]$ can be determined from $x[n]$ and H or the system H can be modeled based on $y[n]$ and $x[n]$. The Z-Transform will also help determine the stability and behavior of the system.

Z-Transform of a DT Signal

Consider a DT signal with a first non-zero value starting at $n = -1$, which is denoted by $x[n] = [a_1, a_2, a_3, a_4]$. By applying Eq. 1, the Z-Transform of the sequence $x[n]$ can be written as

$$X(z) = a_1 z + a_2 + a_3 z^{-1} + a_4 z^{-2} \quad (3)$$

A careful look at Eq. 3 reveals that the value of the sequence $x[n]$ at every instance n is multiplied by z^{-n} resulting in a polynomial (Fig. 2).

Z-Transform of an Impulse Signal

A unit impulse signal $\delta[n]$ has a value of 1 at $n = 0$ and has a value of zero at all other times (Fig. 3).

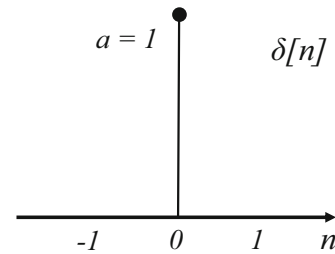
By applying Eq. 1, the Z-Transform of $\delta[n]$ can be written as $Z\{\delta[n]\} = Z^0 = 1$. The delayed version of $\delta[n]$ by N units is represented by $\delta[n-N]$ and corresponding Z-Transform $Z\delta[n-N] = Z^{-N}$. A delay $\delta[n]$ by N steps is equivalent to multiplying the Z-transform of $\delta[n]$ by Z^{-N} . This property is both simple and powerful because any sequence $x[n]$ can be written in terms of time-shifted $\delta[n]$. For example, the sequence in Fig. 2 can be written as

$$X[n] = a_1 \delta[n+1] + a_2 \delta[n] + a_3 \delta[n-1] + a_4 \delta[n-2] \quad (4)$$

Because of Eq. 4, the output of any DT system H can be computed by only knowing the output of the same system for $\delta[n]$. The output of the system H for the input of $\delta[n]$ is known as the impulse response ($h[n]$) of the system. The transfer function of the system H can be obtained by computing the Z-Transform of impulse response $h[n]$. i.e., $Z\{h[n]\} = H(z)$.

Stability of a System from Transfer Function $H(z)$

Since the Z-Transform of a sequence produces a polynomial, it is reasonable to assume that $H(z)$ will be a rational function



Z-Transform, Fig. 3 Plot of an impulse function

with a polynomial in the numerator and denominator. For example,

$$H(z) = \frac{q_1 Z + q_2}{P_1 Z^2 + p_2 Z + p_3} \quad (5)$$

Where q_s and p_s are coefficients of numerator and denominator polynomials, respectively. Using factorization and partial fraction decomposition, Eq. 5 can be written in the form of

$$\frac{b_i}{(z - c_i)^j} \quad (6)$$

Where c_i is called a pole and j is the highest order of denominator polynomial in Eq. 5. At pole values, the denominator of $H(z)$ becomes zero. Using geometric series, the system can be tested for stability. Detailed steps can be found on (Oppenheim and Willsky 1996).

Z-Transform in Geosciences

The Z-Transform has multiple applications in the fields of engineering and geosciences with low computational complexity. The Z-Transform technique has implications in studying rainfall-runoff processes in complex hilly watersheds (Rai and Upadhyay 2009) and in improving accuracy for time-discrete sequences (Sakrison et al. 1967). The Z-Transform-based Genetic Algorithm is used to image authentication technique in the frequency domain (Mandal and Khamrui 2012). The Z-Transform is also helpful in the prediction of periodical variable structure system behavior using minimum data acquisition time (Dobrucky et al. 2007). Weighted wavelet the Z-transform is shown to be effective for period analysis of unevenly sampled time-series data (Foster 1996). The Z-Transform is used in the numerical modeling of poroelastic and viscoacoustic wave equations (Jian-Guo et al. 2014; Xu 2019). The Z-transform is used to extract features from seismic wave data (Robinson 1968; Hanming et al. 2009; Shi et al. 2012).

Summary

The Z-Transform is a generalization of a discrete-time Fourier transform commonly used to evaluate digital filters with feedback. This transfer function of a discrete-time system can analyze discrete-time signals using simple hardware architecture. The Z-Transform can be applied to study signals and discrete-time systems in geophysics.

Bibliography

- Dobrucky B, Marcokova M, Pokorny M, Sul R (2007) Prediction of periodical variable structure system behavior using minimum data acquisition time. In 26th IASTED international conference on modeling, identification, and control (MIC): proceedings. ACTA Press, Calgary, pp 440–445
- Foster G (1996) Wavelets for period analysis of unevenly sampled time series. *Astron J* 112:1709–1729
- Hanming H, Rui L, Jun LS, Ju BY (2009) Discrimination of earthquakes and explosions using chirp-z transform spectrum features. In: WRI World Congress on computer science and information engineering 7, pp 210–214. IEEE
- Jian-Guo Z, Rui-Qi SHI, Jing-Yi C, Jian-Guo P, Hong-Bin W (2014) A matched Z-transform perfectly matched layer absorbing boundary in the numerical modeling of viscoacoustic wave equations. *Chin J Geophys* 57(4):1284–1291. <https://doi.org/10.6038/cjg20140425>
- Mandal JK, Khamrui A (2012) An image authentication technique in frequency domain using genetic algorithm (IAFDGA). *Int J Software Eng Appl* 3(5):39
- Oppenheim A, Willsky A (1996) Signals and systems, 2nd edn. Pearson, London
- Rai RK, Upadhyay A (2009) Z-transform based instantaneous unit hydrograph for hilly watersheds. *J Water Resource Protect* 01(06): 381–390. <https://doi.org/10.4236/JWARP.2009.16046>
- Robinson EA (1968) Basic equations for synthetic seismograms using the Z-transform approach. *Geophysics* 33(3):521–523. <https://doi.org/10.1190/1.1439949>
- Sakrison DJ, Ford WT, Hearne JH (1967) The z transform of a realizable time function relation of the real and imaginary parts and applications to deconvolution. *IEEE Trans Geosci Electron* 5(2):33–41
- Shi R, Wang S, Zhao J (2012) An unsplit complex-frequency-shifted PML based on matched Z-transform for FDTD modeling of seismic wave equations. *J Geophys Eng* 9(2):218–229. <https://doi.org/10.1088/1742-2132/9/2/218>
- Xu Z (2019) Application of matched Z-transform perfectly matched layer in the numerical modeling of poroelastic wave equations. In SEG technical program expanded abstracts 2019. Society of Exploration Geophysicists, pp 3909–3913. <https://doi.org/10.1190/segam2019-3214696.1>

Author Index

A

Abdalla, R., 519
Abdelfattah, R., 1079
Abioui, M., 582
Aggarwal, D., 175
Aghighi, M.A., 208
Agterberg, F., 107, 114, 115, 193, 372, 387, 407, 437, 502, 512, 526, 587, 644, 702, 881, 934, 1066, 1095, 1152, 1332, 1486, 1567, 1575, 1576, 1591
Ali, M.A., 1199
Ali, T., 1079
Alimoradi, A., 489
Aliouane, L., 620, 1264, 1288, 1636
Alowairdhi, A., 369
Ansari, R.A., 1297
Ardejani, F.D., 489
Arroyo, D., 613
Artemenko, I., 582
Arun Kumar, D., 307, 465, 756, 1008, 1203, 1513, 1537
Ashok, B., 751
Attanasi, E.D., 1182
Awange, J., 1047
Azevedo, L., 539

B

Bacon-Shone, J., 17
Balamurali, M., 262, 1273, 1443, 1527
Balasubramanian, N., 1358
Bansal, A.R., 1092
Baranwal, E., 299
Batty, M., 1559
Bengtson, P., 1218
Benndorf, J., 1451
Benssaou, M., 582
Beven, K., 618
Bhattacharjee, S., 1382, 1386
Bilal, M., 1199
Boisvert, J.B., 1166
Boodala, J., 1358
Bourgault, G., 1414
Bovolo, F., 625
Bruzzone, L., 625
Buccianti, A., 127
Burger, H., 1539

C

Caciagli, N., 860, 1126, 1130
Caers, J., 960
Cappello, C., 1353, 1605
Carranza, E.J.M., 364, 862, 1166
Cathles, L., 1578
Chai, T., 1236
Chakraborty, T., 230, 774
Challa, A., 92, 94, 222, 600, 901, 902, 906, 921
Chandrasekhar, E., 1297
Chave, A.D., 1428
Chen, G., 1626
Chen, J., 1382, 1386
Chen, Q., 15, 1342, 1390, 1540
Chen, Z., 744
Cheng, Q., 107, 407, 801, 990, 994, 1066
Christakos, G., 71
Chudasama, B., 449, 1057
Ciriello, V., 1271
Coburn, T.C., 1182
Cohen, D.R., 358
Collins, D., 857
Cox, S.J.D., 858
Craglia, M., 295
Cressie, N., 1362, 1625
Cubitt, J., 1642

D

Damm, A., 404
Danda, S., 92, 94, 222, 600, 901, 902, 906, 921
Dasgupta, A., 1239
Davis, J.C., 317
Daya Sagar, B.S., 92, 94, 222, 226, 247, 257, 404, 586, 600, 703, 901, 902, 906, 921, 923, 1135, 1235, 1513, 1537, 1644
De Iaco, S., 184, 1345, 1373
Desassis, N., 705
Deutsch, C.V., 96
Di, L., 123
Dikshit, O., 1358
Dimitrakopoulos, R., 605
Dimri, V.P., 678, 1092
Doghmane, M.Z., 1264, 1288
Doveton, J.H., 339, 818
Dowd, P.A., 1

E

Emanuel, K., 773
 Engle, M.A., 724, 731
 Erten, O., 96
 Ertunc, G., 1589

F

Filzmoser, P., 1225
 Florinsky, I.V., 928
 Fornberg, B., 1403
 Fouedjio, F., 938, 1410
 Fox, P., 1209
 Fraga Alves, M.I., 883
 Frank, A.U., 100
 French, R.H., 494
 Frery, A.C., 47, 664, 766, 1192, 1339, 1408, 1593
 Freulon, X., 705

G

Gabriel, A.-A., 468
 Gamon, J., 404
 Ganguli, S.S., 678
 Gao, F., 267
 Gastner, M.T., 103
 Gautier, D., 1259
 Giao, P.H., 1086
 Giungato, G., 897
 Gokceoglu, C., 854
 Gómez-Hernández, J.J., 342, 1185, 1276
 Gotway Crawford, C.A., 1625
 Gozzi, C., 127
 Gradstein, F., 1152
 Gradstein, F.M., 502
 Grana, D., 65
 Griffiths, C.M., 393, 568
 Grunsky, E., 95, 143, 1583
 Guha, S., 1135
 Gupta, U., 17, 55

H

Hahmann, T., 1013
 Harff, J., 985, 1510
 He, J., 71
 Hemalatha, G., 1203
 Henley, S., 1623
 Hillier, M., 1175
 Hohn, M.E., 853
 Hossain, T.M., 1600
 Howarth, R.J., 113, 1256
 Hristopulos, D.T., 303, 842, 1597, 1614
 Hron, K., 4, 10
 Huang, F., 235
 Huqqani, I.A., 1020

I

Irving, J., 28

J

Jain, V., 319, 1135
 Jia, Q., 43
 José Egozcue, J., 133

K

Kamiya, N., 1083, 1261
 Kanniah, K.D., 213
 Karacan, C.Ö., 951, 1012, 1196
 Kaufman, G.M., 1516
 Kawamura, Y., 875
 Kitanidis, P.K., 79
 Klichowicz, M., 588
 Klump, J., 656
 Koike, K., 575, 1229
 Korvin, G., 24, 49, 290, 1213, 1456
 Kostyuchenko, Y., 582
 Kotsakis, C., 1032
 Kreinovich, V., 986
 Krieger, V., 404
 Kroner, U., 350
 Kshirsagar, V., 1382, 1386
 Kumar, U., 165, 175, 180, 230, 376, 594, 774, 1053, 1239

L

Lehnert, K., 656
 Leung, J.Y., 954
 Leung, R., 1443
 Lieberwirth, H., 588
 Lim, S.L., 923
 Lin, W., 1083, 1261
 Liu, F., 1390
 Liu, G., 325, 1342, 1540
 Lovejoy, S., 1244

M

Ma, C., 1390
 Ma, X., 85, 293, 369, 1342, 1390, 1535, 1540
 MacLeod, N., 31, 311, 1119, 1219
 Madani, N., 1067
 Madl, J., 1382, 1386
 Maggio, S., 897, 974, 1353
 Maghsoudy, S., 489
 Manglik, A., 672, 1024, 1039
 Marchi, L., 633
 Marcotte, D., 250
 Mariethoz, G., 960
 Marinelli, D., 625
 Mateu-Figueras, M.G., 4, 10, 445, 769
 Mazzetti, P., 534
 McArthur, J.M., 619
 McKinley, J., 1090, 1260
 Menafoglio, A., 1108
 Meschede, M., 604
 Mhawish, A., 1199
 Minnitt, R., 704
 Mitra, A., 791
 Modis, K., 61
 Mohr, M., 468
 Monti, G.S., 445
 Moores, M.T., 1362
 Mortula, M.M., 1079
 Mosegaard, K., 65, 890
 Mueller, U., 870, 958
 Muller, O., 404
 Muralikrishnan, L., 1057
 Murthy, V.S.S., 1008
 Myers, D.E., 1373

N

Nascimento, A.D.C., 1112, 1545
 Nativi, S., 295, 534
 Navalgund, R., 1206
 Negri, R.G., 271, 630, 1063, 1520
 Newman, W.I., 328
 Nguyen, P.H., 1086
 Nichiwaki, N., 985
 Nichol, J.E., 1199
 Nordhausen, K., 688, 720, 728, 748, 1328, 1551

O

Olea, R.A., 251, 435, 769, 1065, 1188, 1566
 Omre, H., 65
 Ortiz, J.M., 346, 960, 1322
 Ouadfeul, S.-A., 620, 1264, 1288, 1636

P

Padmanabhan, S., 785
 Paláncz, B., 1047
 Palarea-Albaladejo, J., 637
 Palma, M., 974, 1605
 Pan, F., 781
 Pan, G., 201, 380
 Panda, R.M., 226, 247, 257, 1644
 Pandalai, H.S., 543
 Panigrahi, N., 666
 Panja, M., 165, 230
 Pardo-Igúzquiza, E., 845
 Pawlowsky-Glahn, V., 133, 1583
 Pedretti, D., 388
 Porwal, A., 449, 454, 1057
 Posa, D., 184, 1345, 1373
 Pour, A.B., 454
 Prabhu, A., 1209
 Prakasa Rao, B.L.S., 1194
 Profeta, L.R., 1017

Q

Qiu, Z., 1199
 Que, X., 485, 1342, 1390
 Quiros-Vargas, J., 404

R

Radojičić, U., 720, 1328
 Rahimzadegan, M., 454
 Ramdeen, S., 656
 Ranjan, U., 17, 55, 350, 795
 Rascher, U., 404
 Rodríguez-Tovar, F.J., 845
 Roshan, H., 208
 Rundle, J., 1578

S

Sadeghi, B., 169, 454, 1322, 1332, 1398, 1583, 1609, 1643
 Sahoo, R., 319, 430
 Salembier, P., 1279
 Samiappan, S., 1644
 Schaeben, H., 215, 350, 759, 1169
 Schuberth, B.S.A., 468
 Schuenemeyer, J.H., 981

Serafina Monti, G., 4
 Serra, J., 820, 835, 910
 Shafaei, F., 489
 Shahhosseiny, M., 489
 Shao, Y., 990
 Shiraz, F.A., 489
 Shokri, B.J., 489
 Shu, M., 43
 Siegmann, B., 404
 Silva, J.F., 346
 Silva Lomba, J., 883
 Silversides, K.L., 739, 1061, 1396, 1511
 Simmons, C.T., 799
 Singer, D.A., 601
 Singh, R.N., 319, 672, 1024, 1039
 Singh, R.P., 1206
 Singh, S., 277
 Smith, L.K., 225
 Soares, A., 539
 Solomon, I., 180, 376
 Sreenivasan, K.R., 784
 Sreevalsan-Nair, J., 241, 336, 355, 447, 642, 660, 695, 697, 700, 715, 735, 851, 868, 970, 999, 1116, 1320, 1619
 Srivastava, P., 277
 Srivastava, R.M., 693
 Stephan, T., 350
 Stoyan, D., 1073, 1501
 Su, S., 485
 Subramanyam, A., 543
 Sun, Z., 123

T

Tahmasebi, P., 523, 573
 Tartakovsky, D.M., 1271
 Taskinen, S., 688, 728, 748, 1328
 Tay, L.T., 1020
 Tercan, A.E., 717
 Thakur, S., 449, 1057
 Thiart, C., 741
 Thiergärtner, H., 1003, 1234
 Thottolil, R., 594
 Tian, Y., 43
 Tomita, S.A., 575
 Trevisani, S., 633, 928

V

Varouchakis, E.A., 842
 Venkatachalam, P., 473
 Venkatanarayana, M., 1008, 1203
 Ver Hoef, J.M., 214
 Verrecchia, E.P., 28
 Villanueva, J.S., 274
 Völgyesi, L., 1047

W

Wang C., 293, 1535
 Wang, R., 404
 Wang, W., 744, 990, 994
 Watada, J., 1600
 Webster, R., 1439, 1579
 Wellmann, F., 1561
 Wen, T., 238
 Wendebourg, J., 603

Whitten, E.H.T., 713
Wilson, B., 1166
Wojnar, L., 1490
Wu, J., 1229
Wyborn, L., 656

X

Xie, S., 252, 744

Y

Yahia, M., 1079
Yao, L., 605
Yarus, J.M., 494

Z

Zhang, H., 325, 1626
Zhang, J., 1540
Zhao, J., 990, 994
Zhuang, J., 1472
Zlotnik, S., 274
Zuo, R., 945

Subject Index

A

- Abbyy Fine Reader, 38
- Abduction, 644, 654
- Abraham-Gottlob-Werner-Medal, 1539
- Absolute ASC (AASC), 178
- Absolute Euler poles, 353
- Absolute loss, 778
- Absolute plate motions, 350
- Absolute porosity, 1091
- Absorption coefficient, 52
- Absorption law, 912
- Abstraction, 1054
- Abundance non-negativity constraint (ANC), 177
- Abundance sum-to-one constraint (ASC), 177
- Acceptance probability, 69
- Acceptance sampling, 1451
 - in environmental monitoring, 1453–1454
- Accept-reject method, 563
- Accessible, 370
- Accuracy, 1–4
- Accurate and relevant attributes, 371
- Acknowledged governance, 536
- Acoustic RS, 1206
- Acquisition, 49
- Across-scale, 319
- Activation function, 44, 757
- Active data dictionary, 232
- Active (Advance) dewatering, 491
- Active sensors, 1207
- Activity of neuron, 1266
- Actuarial risk analysis, 1516
- Actuopalaeontology, 1219
- AdaBoost algorithm, 777
- Adaptation, 1269
- Adaptive filters, 1079
- Adaptive learning, 725
- Adaptive linear (ADALINE), 727, 728
- Adaptive MMSE PolSAR filtering, 1080–1082
- Adaptive neural network, 453
- Adaptive neuro fuzzy inference system, 450
- Adaptive simulated annealing (ASA), 1321, 1615
- Additive identity, 838
- Additive logistic normal model, 5, 6, 12
 - compositional data, 5
 - geosciences, 7, 8
 - key properties, 7
 - logratio (alr) representation, 4
 - logratio methodology, 4
 - multivariate normal distribution, 4
- Additive logistic skewnormal model, 5, 12
- Additive logratio, 144
 - representation, 10
- Additivity relationship, 708
- Adiabatic atmosphere, 584
- Admissibility condition, 1304, 1607
- Advection, 320
- Advection-diffusion equation, 320, 1042
- Advection-dispersion-reaction (ADR) equation, 717
- Aeolian landscape, 1141
- Aeolian processes, 1141, 1142
- Aeolian transport, 320
- Aerial laser scanning (ALS), 736
 - point cloud analysis, 737–738
- Aerosol optical depth (AOD), 1199, 1201
- Affine change of support model, 556
- Aftershocks, 330
- Age of the Earth, 804, 805
- Agglomerative approach, 271, 982, 1505
- Aggregation, 1054
- Agterberg, F.P., 15, 16
- Air quality index (AQI), 487, 488
- Air Research and Development Command (ARDC) Model, 584
- Aitchison, J.
 - career, 17
 - distance, 10, 132
 - education, 17
 - geometry, 136, 1117
 - measure, 11, 12
 - principles, 135
 - statistical compositional data analysis, 17
- Akaike information criterion (AICc), 486, 1394, 1476
- Aki-Richards approximation, 69
- Aleatoric or irreducible uncertainty, 971
- Alexa, 34
- AlexNet, 45, 265
- Algae blooms, 1184
- Algebraic closing, 902
- Algebraic filter, 93
- Algebraic openings, 922–923
- Algebraic perspective, 1033
- Algebraic reconstruction techniques (ART)
 - advantages and disadvantages, 21
 - algorithm, 18–19
 - characteristics of the solution, 20–21
 - convergence of ART algorithm, 20
 - definition, 17–18
 - SART, 21–23
 - SIRT, 21
 - tomographic projection model, 18
- Algebraic topology, 1565

- Algorithm, 34, 685, 686, 688
- Algorithm-based methods, 963
 - nodes, 963
 - sequential framework, 963
- Algorithmically-defined random functions, 1187–1188
- Alias filter, 290
- Allocating agents, 657–659
- Allometric, 1222
- Allometric power laws
 - in geosciences, 25
 - modified power laws, 26
 - and self-similarity, 24–26
 - size-distribution, 26
- Allometry, 24
- Alpha-numeric data, 34
- Alternating sequential filters, 93, 596, 911–912
 - grey-scale images, 601
- Ambient noise level, 51
- American Association of Petroleum Geologists (AAPG), 252
- American Statistical Association, 1443
- Ammonites, 1219
- Amount, 1492
- Amplitude spectrum, 1298
- Amplitude variation with angle (AVA), 686
- Amplitude versus offset (AVO), 53
 - inversion, 67
- Anabranching channels, 1139
- Analog filter, 290
- Analogistic method, 1280
- Analog seismic recorders, 53
- Analog-to-digital (AD) converters, 53
- Analyses of variance (ANOVA), 1441
- Analysing wavelet, 1304
- Analysis, 1053
- Analysis of variance (ANOVA), 147, 151, 155, 632, 1579
 - balanced designs, 1580
 - geochemistry soil Jura, 1582
 - lead, 1582
 - multi-stage analysis, 1580
 - multi-stage survey, 1581
 - residual maximum likelihood estimation, 1582
 - single-stage, 1579
 - unbalanced design, 1579
 - variogram, 1581
- Analytical methods, 212
- Analytical reconstruction, 55
- Analytical solution, 726
- Anamorphosis function, 556
- Angular data, 115
- Angular dispersion, 626
- Angular frequency, 30
- Anhydrite, 818
- An Introduction to Statistical Models in Geology*, 714
- Anis-Lloyd corrected R/S Hurst exponent, 1216, 1217
- Anisotropic, 548
 - response functions, 1481
 - singularity estimation, 746
 - singularity index estimation methods, 747
- Anisotropy, 548, 1348
 - in space-time, 1379
- Annealing, 1466
- Annual mean temperature data, 1216
- Annual precipitation, 260
- Annual rainfall data, 1216
- Annual river flow data, 1216
- Anomalism, 359
- Antarctic Meteorite, 882–883
- Antecedent, 450
- Anti-alias filter, 290
- Anticlines, 372
- Anti-extensive operator, 909, 922
- Anti-persistence, 432, 1213, 1215, 1217
- Anti-persistent, 1311
 - random walk, 620
- Apache Hadoop, 125
- Apache Spark, 125
- Apollo, 343
- Appalachian ultramafic belt, 1569
- Apparent resistivity, 675
- Application ontology, 1016
- Application-programming-interface (API), 86
- Applied Geochemistry Research Group, 619
- Approx-Grid Top k algorithms, 1389
- Approximate coefficients, 1307
- Approximation, 1601–1602
 - of reality, 1542
- A priori variance, 1605–1607
- Arbitrary high-order DERivative (ADER), 472
- ArcGIS, 225
- Archaeology, 39, 1622
- ARCHEO-NET, 38
- Archie's equation, 1465
- Archie's law, 27
- Archimedes' principle, 1085
- Archive formats, 248
- AR coefficients, 846
- Area cartogram, 103
 - categories, 103
- Area-occurrence plot, 866, 867
- Area-to-point kriging, 709
- Argand diagram, 30
 - of complex numbers, 1170
 - definition, 28
 - geometric plot, 29
 - in geosciences, 29–30
 - historical origin, 28–29
- Arrangement, 1053, 1493
 - pattern, 1054
- Arrhenius equation, 1470
- Arrhenius law, 1470
- Arrival times, 51
- AR technology, 880
- Artificial intelligence (AI), 31, 124, 166, 262, 329, 498, 774, 814–816, 1055, 1345
 - computer vision, 36
 - definition, 31
 - development and activity domains, 32–38
 - Earth Science applications, 38–40
 - expert systems, 34
 - history, 32
 - hostile aircraft identification, 36
 - machine learning, 35–36
 - natural language processing, 34–35
 - planning systems, 36–37
 - and robotics, 37–38
- Artificial neural network (ANN), 34, 262, 323, 675, 819, 951–953
 - applications, 45–46
 - components of, 44
 - definition, 43

learning, 44
 models, 815, 816
 organization, 44–45
 Artificial neuron, 32
 AR visualization system for rock stress in an underground mine, 881
 Aspect ratio, 1460, 1464
 Association(s), 1054
 pattern, 1054
 Association and correlation (AC) analysis, 239
 Associative law, 838
 ASTER Global Digital Elevation Model (ASTER GDEM), 924
 Astrochronology, 647
 Astronomical data, 231
 Asymptotic distribution, 768–769
 Asymptotic normality, 852
 Asymptotic variance, 852
 Athy's law, 1461, 1462
 Atmosphere model, 583, 584
 Atmospheric physics, 443
 ATmospheric REMoval (ATREM), 628
 Atmospheric science, 38
 Atomic second, 505
 Attenuate, 1640
 Attenuation, 49, 51
 Atterberg limit test, 574
 Atterberg soil index, 574
 Attitude behavior, 280
 Attitude Orbit Control System (AOCS), 280
 Attributes, 1054
 Attribution, 1018
 Augmented semantic classification, 737
 Authigenesis, 1263
 Authorized linear combination, 546, 550
 Autocorrelation, 897, 1571, 1574
 definition, 47
 extensions, 48
 time, 1217
 time series, 47–49
 Autocorrelation function (ACF), 48, 292, 376, 526, 1456–1459
 Autocovariance, 526
 functions, 526
 Autoencoder (AE), 270
 Autogenic control, 434
 Automated character recognition, 924
 Automated correlation, 1164
 Automated information extraction and integration, 1535
 Automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE), 1529
 Automatic gain control, 49
 Automatic level control, 49
 Automatic volume control, 49
 Automating geoscience analysis
 data mining vs. machine learning, 982
 SEM, 982
 supervised vs. unsupervised learning, 983
 Autonomous computing, 500
 Autoregressive (AR) method, 844, 851
 Autoregressive moving average (ARMA) model, 1553
 Auxiliary variables, 547
 Average absolute amplitude, 53
 Average distance, 128
 AVIRIS sensor, 626
 Axially symmetric, 1348
 Axiomatic ontology, 1017
 Axons, 1264

B

Backprojection tomography, fan-beam image reconstruction, 58–60
 Back propagation algorithm, 953
 Back-propagation method, 45
 Back substitution, 1434
 Backup types, 249
 Backus-Gilbert method, 672, 674
 Backus-Naur-form (BNF), 569, 570
 Bagging, 258
 Banach space, 614
 Banded iron formations (BIF), 1513
 Band ratio, 302
 Barometric equation, 1461
 Barotropic atmosphere model, 584
 Barrier models, 1371
 Bartlett window, 291
 Basalt, 1566
 Base metal, 453
 Basin analysis, 1470
 Basin scale, 1140
 Basis contrast matrix, 137
 Basis functions, 546, 1047
 Basis pursuit inversion (BPI), 686
 Batch learning, 953
 Bath's law, 330, 331, 1474, 1483
 Bathymetry, 928
 Bayes generalized linear model (BayesGLM), 1322
 Bayesian analysis, 84
 Bayesian decision theory, 774
 Bayesian framework, 1023
 Bayesian inference, 795–796
 Bayesian inverse problems, 786
 Bayesian inverse theory, 65, 66, 69, 70
 Bayesian inversion, 674–675
 Bayesian linear formulation, 67
 Bayesian linearized AVO inversion, 69
 Bayesian linearized seismic inversion methods, 540
 Bayesian LISP-based systems, 38
 Bayesian maximum entropy (BME), 72
 chronotopologic domain, 73
 chronotopologic estimation, 75
 definition, 71
 knowledge bases, 72–73
 model, 1387
 in real word applications, 76
 in software, 76
 uncertainty assessment, 75
 versatility, 75–76
 workflow, 74–75
 Bayesian methods, 65, 147, 1059, 1381
 Bayesian networks, 36
 Bayesian statistics, 64
 Bayes' rule, 75, 645, 894
 binary map layer, 653
 single map layer, 651
 Bayes's theorem, 62–64, 649–650, 674
 definition, 61
 derivation, 63
 Baysian Monte Carlo Markov Chain (MCMC) approach, 400
 Behavior, 1053, 1638
 Benioff strain, 1478
 Berea sandstone, 1083, 1084
 Bernstein functions, 1377
 Bessel function, 716, 1349
 Best linear unbiased estimator (BLUE), 1415

- Best optimal predictor (BOP), 1365
 Beta distribution, 665, 1192
 Bhatnagar-Gross-Krook, 1466
 Bias, 437
 Bias-corrected and accelerated (BCa) approach, 1431
 Bias correction in goodness of fit tests, 1432
 Bicubic splines, 1406
 Bifurcation ratio, 25
 Big data, 166, 231, 233, 235, 237, 323, 498, 814–816, 1020, 1345
 context-aware data and resource for geoscience, 90
 data science approaches, 91
 definition, 85
 infrastructure and characteristics of, 87–88
 machine-readable FAIR metrics, 89
 management, 249
 open science and open data, 88
 smart data ecosystem for geoscience, 88–89
 Bilateral gamma distribution, 197
 Bilinear interpretation, 1242
 Bilinear unmixing, 739
 Binarization, 926
 Binary association indices, 1119
 Binary classification, 777
 Binary gain control (BGC), 49, 53
 Binary images, 92, 93, 903, 905, 907
 basic MM operators, 93
 filtering on, 93
 geodesic dilation/erosion, 93
 HMT, 93
 skeletonization, 93
 thinning/thickening, 93
 Binary partition tree (BPT), 95
 definition, 94
 in edge-weighted graphs, 94
 merging order, 94
 region model, 94
 Binary set theory, 454
 Binary similarity coefficients, 1120
 Binomial distribution, 882, 999, 1190
 Bins, 435
 Bioengineering, 37
 Biogeography, 1219
 Biography
 Agterberg, F.P., 15–16
 Aitchison, J., 17
 Bonham-Carter, G., 95–96
 Burrough, P.A., 100–101
 Chayes, F., 113–114
 Cheng, Q., 114–115
 Cracknell, A.P., 213–214
 Cressie, N.A.C., 214–215
 Dangermond, J., 225–226
 David, M., 250–251
 Davis, J.C., 251–252
 de Marsily, G., 799–800
 Doveton, J.H., 317–318
 Fisher, R.A., 387–388
 Goodchild, M.F., 586–587
 Gradstein, F., 587–588
 Griffiths, J.C., 601–602
 Harbaugh, J.W., 603
 Harff, J., 604
 Horton, R.E., 618
 Howarth, R.J., 619
 Journel, A., 693
 Kolmogorov, A.N., 702–703
 Korvin, G., 703–704
 Krige, D.G., 704–705
 Krumbein, W.C., 713–714
 Lorenz, E.N., 773
 Mandelbrot, B.B., 784–785
 Matheron, G., 835–836
 McCammon, R.B., 853–854
 Merriam, D.F., 857–858
 Nimec, V., 985–986
 Olea, R.A., 1012
 Pawłowsky-Glahn, V., 1065–1066
 Pengda, Z., 1066, 1067
 Rao, C.R., 1194–1195
 Rao, S.V.L.N., 1195
 Reyment, R.A., 1218–1219
 Rodionov, D.A., 1234–1235
 Rodriguez-Iturbe, I., 1235–1236
 Schuenemeyer, J.H., 1259
 Schwarzacher, W., 1260
 Serra, J., 1279–1280
 Stoyan, D., 1510–1511
 Thiergärtner, H., 1539–1540
 Tobler, W., 1559–1560
 Tukey, J.W., 1575–1576
 Turcotte, D.L., 1578
 Vistelius, A.B., 1623–1624
 Watson, G.S., 1625–1626
 Whitten, E.H.T., 1642
 Zadeh, L.A., 1643–1644
 Biological, 1224
 neuron, 1265
 Biology, 38, 1119
 Biometric applications, 1219
 Bi-orthogonal decomposition, 1110
 Biostratigraphic correlation, 198
 Biplots, 146
 Bi-square, 486
 Bivariate distributions, 554
 Bivariate Gamma, 555
 Bivariate Gaussian, 555–557
 Bivariate negative binomial, 555
 Bivariate stationarity, 545
 Black Hole Algorithm, 1048
 Blackman and Tukey method, 1093
 Blanks, 1130–1131
 Blasting Optimisation in Open-Pit Mine, 878
 Blind source separation (BSS), 145, 643, 1557
 Block kriging, 709
 Bochner-Khinchin-Wiener theorem, 305
 Bochner's Theorem, 547, 1346
 Body frame coordinate system, 281
 Bolgiano-Obukhov scaling, 1253
 Boltzmann, Ludwig, 842
 Boltzmann constant, 1206, 1461
 Boltzmann distribution, 964
 Boltzmann simulated annealing, 1615–1616
 Boltzmann's kinetic theory of gases, 1466
 Bombay offshore basin, 1314
 Bond-percolation, 1467
 Bonds, 1468
 Bonham-Carter, Graeme, 95
 Boolean complementation, 822
 Boolean lattices, 822
 Boolean methods, 967
 Boolean models, 566
 Boolean random function model, 566

- Boosted generalized linear model (GLMBoost), 1322
 - Boosting, 147, 258
 - Bootstrap, 323, 894, 1430, 1431, 1434
 - definition, 96
 - method, 1190
 - percentile method, 1431
 - realization, 97
 - sample, 1182
 - traditional, 96–97
 - Bootstrap-t confidence interval, 1431
 - Borderline area, 1601
 - Borehole profiles, 1004
 - Bouguer anomaly, 686, 1638–1641
 - Bouma sequences, 1577
 - Bounce-back' conditions, 1466
 - Boundary(ies)
 - conditions, 63, 275
 - disordered objects, 1004
 - hierarchy of boundaries, 1004
 - ordered objects, 1003
 - recognition, 36
 - region, 1601
 - Boundaries-only algorithms, 105
 - Boundary element method (BEM), 212
 - Bounded function, 1606
 - Bounded influence estimators, 1436
 - Bounds, 1463, 1517–1518
 - Boxcar function, 291
 - Box-Muller method, 891
 - Boyle's law, 1085
 - Branches of geosciences, 624
 - Branching aftershock sequence (BASS), 1484
 - Breakdown point, 1225
 - Bre-X, 1126
 - Brightness value (BV), 300
 - Brine-filled porous rocks, 1464
 - British petroleum, 619
 - Broad probability distribution, 1314
 - Brownian motion, 1321, 1464
 - Brownian passage-time (BPT), 1477, 1478
 - Brownian relaxation oscillator (BRO), 1477
 - Browser-based visualization tool, 1619
 - B-spline interpolation, 288
 - BSS/ICA model, 643
 - Buffer analysis, 1344
 - Bulk chemistry, 740
 - Bundle adjustment, 278
 - Bundle adjustment theory, 36
 - Buoyancy method, 1084
 - Burial after sedimentation, 1261
 - Burial-temperature-time history, 1470
 - Burn-in time, 893
 - Burrough, P.A., 100, 101
 - Burridge-Knopoff Model, 334, 335
 - Business management, 231
 - Butterworth filter, 1637
- C**
- Calcareous nannofossils, 194
 - Calcite, 1464
 - California's earthquake faults, 331
 - Camera orientation and 3D point position, 877
 - Canonical correlation analysis (CCA), 974–976
 - Canonical favorability model, 384, 386
 - Canonical variates analysis (CVA), 313–315
 - Cantor set, 431, 432
 - Capillary pressure, 955
 - Carbonate(s), 1465
 - rocks, 603
 - Carbon cycle, 1031
 - Carboniferous-Permian (CP), 508
 - Cardinal data, 1405
 - Cardinal polynomials, 1404
 - Cardinal sinus function, 1290
 - Career progression, 37
 - CARE Principles, 1018
 - Carrying capacity, 1026
 - Cartesian coordinates, 390
 - Cartesian coordinate space, 337
 - Cartesian coordinate system, 804
 - Cartesian plane, 29
 - Cartesian reference frame, 1308
 - Cartier's relation, 557
 - Cartographer, 1361
 - Cartography, 100, 521
 - Cascadia subduction zone, 328
 - Catalog relations, 231
 - Categorical random process, 562
 - Categorical random variable, 564
 - Categorical variables, 1325
 - Cauchy's functional equation, 24
 - Cauchy simulated annealing, 1616
 - Causal signals, 1297
 - Causative sources, 1639
 - Cause-and-effect relationship, 208
 - Cause-effect, 324
 - Caves, 25, 26
 - Celerity, 320
 - Cell declustering, 99
 - Cementation, 1263
 - Censored data, 144, 638
 - Censoring, 1430
 - Centered logratio (CLR), 1116, 1117
 - Central Geological Survey Berlin, 1539
 - Centralized repository, 231
 - Central limit theorem (CLT), 323, 787, 999, 1440
 - Central metadata repository, 232
 - Central-point cartogram, 105, 106
 - Centred logratio (clr) representation, 5, 10, 144
 - Centre of mathematical morphology of the École des Mines de Paris, 1280
 - Certain rules, 1600, 1602, 1603
 - Certified reference materials (CRMs), 1127
 - CFD-DEM coupling, 525
 - Change detection methods, 629
 - Change of support, 544, 556–559
 - Changjiang Scholar, 115
 - Channel erosion, 433
 - Channel networks, 26
 - Chaos, 430, 995–996
 - definition, 107
 - deterministic models, 108
 - Lorentz Attractor, 111, 112
 - model of de Wijs, 108, 109
 - multifractals, 108
 - till sampling, 109–111
 - Chaotic, 1469
 - geometries, 404
 - systems, 813
 - Chapman-Kolmogorov, 703

- Characteristic analysis, 383–384
- Characteristic function, 1457–1459
- Characteristic polynomial, 337
- Chayes, F., 113
 - database, 114
 - data-retrieval requests, 114
 - geochemical data, 114
 - major-element analyses, 114
- Chebyshev filters, 1637–1638
- Chebyshev's inequality, 1408
- Check assays, 1131
- Chemical composition, 740
- Chemical contents, rocks, 730
- Chemical weathering, 1137
 - index/indices of alteration, 1137
- Chemo-hydromechanical coupling, 211
- Chemo-thermo-hydromechanical coupling, 211
- Cheng, Qiuming
 - career achievement, 115
 - early life, 114
 - William Christian Krumbein medal, 115
- China University of Geosciences (CUG), 1067
- Chinese academy of sciences, 115
- Chi-square test, 387
- Cholesky decomposition, 561
- Cholesky method, 708
- Chord distance, 128
- Chromosome, 466
- Chronostratigraphic assignments, 1162
- Chronostratigraphic scale, 507
- Chronotopologic BME estimation, 75
- Chronotopologic domain, 73
- Chunking, 1525
- Circular cartograms, 104, 105
- Circular error probability
 - angles, 116, 117
 - azimuths and dips, 122
 - Bjorne Formation, 117
 - 2-D circular and 3-D spherical statistics, 116
 - definition, 115
 - directed and undirected unit vectors, 118, 119
 - 3-D space, 115
 - geoscience, 116
 - linear features, 115
 - microscopic level, 115
 - statistical analysis, 115
 - statistical theory, 116
 - TRANSALP profile, 119, 122
 - types of mean, 116
- Circularity, 1496
- CityGML, 1620
- Civil rights, 38
- Clamped spline, 1404
- CLARA (Clustering LARge Applications), 698, 699
- CLARANS (Clustering Large Applications based on RANdomized Search), 698, 699
- Class, 1004, 1053
 - boundaries, 758
- Class dependent LDA (CDLDA), 308
- Classical planning, 37
- Classical scaling, 939–940
- Classification, 147, 1053
 - problem, 1182
 - process, 312
- Classification analysis (CA), 240
- Classification and regression tree (CART), 147, 269
- Class independent LDA (CILDA), 308
- Class typicality, 148
- Class variance, 307, 309
- Clausius, Rudolph, 842
- Clay, 1461
 - minerals, 819
- ClearEarth project, 38
- ClearTK, 39
- Climate anomaly, 583
- Climate change, 1056
- Climate dynamics, 430
- Climate prediction, 583
- Climate projection, 583
- Climate sciences, 39
- Climate system, 583, 1397
- Closed geometric mean (CGM), 640
- Closed-number systems, 514
- Cloud, 25
 - service, 123, 165
- Cloud computing, 249, 1019
 - cloud-based data analysis, 125
 - cloud-based data storage, 125
 - cloud-native, 126
 - data center, 124
 - data security, 126
 - definition, 123
 - geosciences, 124
 - geoscientific use cases, 125
 - geoscientists, 125
 - governance, 126
 - instance, 124
 - institutes and commercial companies, 123
 - large-scale and high-performance computing, 123
 - managing cloud expenses, 126
 - privacy, 126
 - public clouds, 126
- CloudStack, 126
- Cloud storage, 249
- C–lower-approximation, 1601
- clr*-transformed data, 361
- Cluster, 1054
 - centroids, 131
 - classification, 127
 - intra-cluster structure, 1530
- Cluster analysis, 127, 240, 368, 974, 977, 1005, 1006, 1008, 1527
 - average linkage, 129
 - censored data, 132
 - complete linkage, 129
 - compositional data, 127
 - dendrogram, 127
 - dichotomic variables, 132
 - distance matrix, 128
 - fuzzy clustering, 128
 - hard clustering, 128
 - hierarchical procedure, 127
 - historical material, 127, 128
 - k-means algorithm, 131
 - log-ratio family, 132
 - multicollinearity, 133
 - outliers, 132
 - partitioning methods, 131
 - Q-mode, 128
 - representativeness of the sample, 133
 - R-mode, 128
 - single linkage, 129
 - weighted Euclidean distance, 128

- Clustering, 340, 447, 465, 981–982, 1053, 1060
 - algorithms, 447
 - technique, 466
- Clustering methods
 - agglomerative, 977
 - divisive, 977
- Cluster point patterns, 1077
- Cluster sampling, 1240
- CO₂ sequestration, 388
- Coal, 1512
- Coalition for Publishing Data in the Earth and Space Sciences (COPDESS), 656
- Coarse duplicates, 1132
- Coarse-graining, 1469
- Coastal regions, 1143, 1144
- Coast effect, 1314
- Cobweb diagram for the logistic map, 755
- CoDaPack, 142
- Code-based data science, 497
- Coefficient(s), 486, 1119
 - of dispersion, 416
 - of variation, 1
- Co-evolution, 535
- COGEODATA, 1219
- Cognition, 36
- Co-kriging, 550–552, 1179
- Cole-Cole model, 29, 30
- Collaborative governance, 536
- Collaborative virtual environments (CVEs), 1622
- Collaboratory for Studies of Earthquake Predictability (CSEP), 1480
- Collision, 1466
 - rules, 1466
- Collocated, 547
- Column matrix, 837
- Committee for Mineral Reserves International Reporting Standards, 1130
- Common logic, 1015
- Common shot gather (CSG), 869
- Commonwealth Scientific and Industrial Research Organisation (CSIRO), 387
- Communality, 1121, 1222
- Communication protocols, 370
- Community standards, 371
- Commutative law, 838
- COMOD, 1591, 1592
- Co-monotonicity, 1516, 1517
- Compaction, 27, 1461–1462
- Compactly supported, 1304
- Complete lattices, 820, 821, 905
 - anamorphosis, 821
 - atom, 822
 - complement, 822
 - co-prime, 822
 - sup-generating family, 821
- Completely heterotopic, 547
- Complete spatial randomness (CSR), 1058, 1076, 1368
- Complete symmetric, 1349
- Complex, 319
 - function, 1352
 - integration, 1299
 - numbers, 1352
 - plane, 30
 - RFs, 1346
 - valued RFs, 1346
- Complexity, 434, 928
 - measure, 405
- Complex-valued random functions, 1381
- Components, 1637, 1640
- Composite, 1463
 - material, 1464
- Compositional colinearity, 819
- Compositional data, 5, 144, 368, 639–642, 1125
 - additive log-ratio representation, 137
 - Aitchison geometry for, 136
 - Aitchison principles, 135
 - analysis, 513, 514, 983
 - books on, 142–143
 - center and variability, 139–140
 - centred log-ratio representation, 137
 - compositional biplot, 140–141
 - definition, 134
 - dendrogram, 141–142
 - entry, 11
 - linear functionals on, 136
 - orthogonal log-ratio representations, 137
 - sample space of, 135–136
 - sequential binary partition, 138
 - set, 14
 - software packages, 142
 - tools, 139–142
 - varanasi data set, 139
 - zeros, 138–139
- Compositional time series (CTS), 1558
- Compositions, 1189
- Compound pattern, 1055
- Compressibility, 1460
- Compression ratio, 247
- Computation, 623
 - cost, 1524, 1525
 - efficiency, 609
 - linguistics, 38
 - phase transitions, 1467–1470
- Computational fluid dynamic (CFD), 525, 575
- Computational geoscience
 - AOV, 151, 155
 - censored data, 144
 - coordinates and elemental associations, 146
 - fractal methods, 145, 146
 - geospatial coherence, 146, 147, 151, 152, 156
 - metric, 143, 144
 - MINFILE, 149
 - multivariate data analysis, 145
 - principal component analysis, 147
 - process discovery, 149–153
 - processes, 143
 - process validation, 147, 148, 152–155, 159, 160
- Computer-aided or computer-assisted instruction (CAI), 165
- Computer-aided or computer-assisted learning (CAL), 165
- Computer Applications in the Earth Sciences, 857
- Computer-assisted or computer-aided instruction (CAI)
 - advantages, 167
 - cognitive processes, 166
 - computer programs or games or online activities, 167
 - creating' knowledge, 166
 - features of, 166
 - in geospatial analysis, 168, 169
 - immediate feedback mechanism, 167
 - learner using, 167
 - pattern-recognition skills, 166
 - personalized information, 166
 - recognition type of interactions, 166
 - reconstruction type of interaction, 166

- Computer-assisted or computer-aided instruction (CAI) (*cont.*)
 - restrictions, 168
 - self-directed learning nature, 167
 - self-pacing, 167
 - trial sessions, 166
 - types of, 165, 166
 - types of student interaction, 166
- Computer-based instruction (CBI), 165
- Computer Contributions, 857
- Computers & Geosciences, 857, 1219
- Computer science, 34, 37
- Computer vision, 36, 262, 264, 265
- Computing probabilities, 999
- Concave function, 1349
- Concavity, 322
- Concentration-area (C-A)
 - fractal model, 170–174, 810
 - method, 407
 - multifractal model, 946
- Concentration-concentration (C-C) model, 173
- Concentration-volume (C-V) model, 173
- Concept map, 1014, 1016
- Conceptualization, 235, 1013
- Conceptual modeling, 1014
- Condensed matter physics, 1470
- Conditional autoregressive (CAR) model, 1366
- Conditional bias, 1416–1417, 1422–1427
- Conditional covariance, 343
- Conditional covariance matrix, 560
- Conditional distribution, 1278
- Conditional entropy, 347
- Conditional inference random forests, 1183
- Conditional intensity, 1480, 1481, 1484
 - definition, 1476
 - functions, 1475
 - likelihood, maximum likelihood estimate and AIC, 1476
 - transformed time sequence and residual analysis, 1476, 1477
 - waiting time, 1475
- Conditionally negative definite function, 545
- Conditionally positive definite (CPD), 1175, 1176
- Conditional negative definite function, 1347
- Conditional planning, 37
- Conditional probability, 63, 649–650, 1278
- Conditional random field (CRF), 1536
- Conditional realizations, 1278
- Conditional sampling method, 789
- Conditional simulation, 560, 1196
- Conductance, 1463
- Conduction, 577
- Conductivity, 1463, 1465
- Conductors, 1264
- Confidence interval, 437, 1240
 - estimation, 1431
- Confidence levels, 777
- Confocal microscopy, 1499
- Conformal stratigraphic, 1179
- Conformant planning, 37
- Confusion matrix (CM), 311, 316, 1010
- Conjugated quaternions, 1171
- Connection, 1265
- Connectivity graphs, 1563
- Connectivity of fracture, 1231
- Consequent, 450
- Conservation of mass, 319
- Conservative plate boundaries, 350
- Consolidation, 1263
- Constant dimension, 406
- Constant probability density, 891
- Constitutive relation, 320
- Constrained optimization, 37
 - cases, 178, 179
 - constraint function, 178
 - convexity, 176
 - definition, 175
 - economic decisions, 175
 - general form, 176
 - geospatial science, 177
 - Lagrange multiplier, 176
 - objective function, 178
 - problem, 1322
 - substitution method, 176
- Constrained simultaneous algebraic reconstruction technique (C-SART), 23
- Constraining models, 964
- Contact angle, 955, 1467
- Contact force, 1462–1463
- Contaminant source, 344
- Content, 1640
 - based video retrieval, 1056
- Contiguous cartograms, 104, 105
- Continental drift, 330
- Continuous covariance function, 1352
- Continuous/discrete wavelet transform, 1298
- Continuous integration, 498
- Continuous probability distribution, 224
- Continuous signals, 1289
- Continuous-time analog signals, 290
- Continuous variables, 1325
- Continuous wavelet transform (CWT), 377, 623, 1305, 1306, 1628
 - coefficients, 1305
 - lithologies, 1305
 - optimization, 1305
 - scaled (compressed/dilated), 1305
 - scalogram, 1305
 - translational invariance property, 1305
 - well logging, 1305
- Continuum mechanics, 274
- Contour extraction, 663
- Contouring, 480
- Contourlet transform, 1308
- Contour map, 289
- Contrast ratio (CR), 301
- Control area, 383
- Controlled vocabulary, 1014, 1016
- Convection, 577
- Convergence, 726
- Convex analysis, 176
 - affine functions, 180
 - criteria, 181
 - definition, 180
 - field of economics, 180
 - first order conditions, 181
 - geospatial applications, 183
 - Lagrangian, 182
 - least squares solution, 183
 - linear programming, 180
 - loss functions, 180
 - objective function, 180
 - operations, 182
 - second order conditions, 181
- Convex function, 777, 1349
- Convex hulls, 926

- Convexity measure, 926
- Convex loss functions, 774
- Convey messages, 1264
- Convolutional filtering, 290–291, 293
- Convolution algorithm, 290
- Convolutional graph neural networks (CGNNs), 270
- Convolutional layer, 263
- Convolutional neural networks (CNNs), 39, 262, 268–269, 1322, 1536
- Convolution operator, 1094
- Convolution principle, 1295
- Convolution theorem, 291, 1301
- Cooley-Tukey algorithm, 372, 376
- Cooling schedule, 1467
- Coordinate representations, 10
- Coordinate systems, 281, 897
- Coordination number, 1458–1460, 1463, 1464, 1468
- Copper mineralization, 207
- Co-prime, 822
- Core/general G-KB, 72
- Core, 1600, 1602
- Coregionalization, 546, 547, 978–980
- Correlation, 196, 197, 624, 1054, 1220
 - analysis, 201
 - distance, 1457, 1458
 - hydrocarbon exploration, 193
 - length, 1459, 1468
 - statistical method, 193
 - stratigraphy, 193, 194
 - structure, 128
 - theory, 1346
- Correlation and scaling (CASC), 196
- Correlation coefficient, 386, 632
 - definition, 201
 - gold and silver, 205
 - hypothesis testing, 205–207
 - machine-learning, 201
 - mineral deposit, 204, 205
 - parametric/nonparametric, 201
 - partial correlation, 205
 - Pearson coefficient, 204
 - population and sample, 202, 203
 - positive correlations, 204
 - random variable, 202
 - rank correlation, 207, 208
 - scatter graphs, 201, 202
 - statistics, 203, 204
- Correlogram, 417, 419, 526
- Correspondence analysis (CorA), 974, 976, 977, 982
- Co-simulation, 560, 1072
- Cosine θ index, 1120
- Cosine similarity measure, 128
- Cost function, 758, 1528
- Council of Mining and Metallurgical Institutions (CMMI), 1130
- Coupled differential equations (ODE and PDE), 1149
- Coupled modeling
 - definition, 208
 - groundwater flow, 209
 - isotropic and homogeneous, 210
- Coupled ocean-atmosphere general circulation model, 583
- Coupling terms, 209
- Courant-Friedrichs-Lewy (CFL) criteria, 1046
- Covariance, 47, 203, 845, 1186, 1187, 1220, 1276, 1605
 - compositional biplot, 140
 - estimator, 1228
 - function, 545, 1346–1348, 1375, 1379, 1380
 - matrix, 338, 1079, 1080
- Covariogram, 1346
 - modeling, 1351
- Cox process, 1059, 1077
- Cracked rock, 1464
- Cracknell, Arthur Phillip, 213, 214
- Cramer-Rao inequality, 1194, 1438
- Ordinary cokriging, 551
- Cressie, Noel A.C., 215
- Cressie-Huang class of models, 1377
- Cretaceous, 1218
 - biozonation, 1161
 - optimum sequence, 1161
- Critical percolation probability, 1469
- Critical phenomena, 25
- Critical porosity, 1087
- Critical systems, 1597
- Critical temperature, 1598
- Critical threshold, 1598
- Critical zone(s), 1148, 1149
 - observatory, 1148
- Cross-correlation, 290, 978
 - function, 292
- Cross-entropy, 777
- Cross-Iddings-Pirsson-Washington (CIPW) norm, 818
- Crossover points, 465
- Crossplot, 1257
- Cross-validation (CV), 486, 507, 511, 953, 1351
 - score, 1394
 - statistics, 670
 - techniques, 1168
- Crowding problem, 1528
- Croxall property, 1117
- Crusta, 1620
- Crustal geotherm, 1030
- Cryological, 1224
- Cryology, 39
- Crystal(s), 589
 - lattice symmetry, 807, 809
- Crystallographic preferred orientation
 - analysis of, 215
 - applications, 219
 - definition, 215
 - pole density function, 217
- Crystallographic symmetry, 216
- Crystallographic systems, 808
- Cubic convolution approach, 1242
- Cubic law, 1232
- Cubic mesh, 1463
- Cubic packing, 1465
- Cubic polynomials, 282
- Cubic-spline interpolation, 1316
- Cuckoo Search, 1048
- Cumulants, 605
 - of WTMM coefficients, 1315
- Cumulative distribution function, 223, 224, 664, 1113
- Cumulative hazard function, 1475
- Cumulative probability distribution function, 1475
- Cumulative probability plot, 223, 224
 - definition, 222–223
 - properties, 223–224
 - relation to probability density and probability mass functions, 223
 - uses in statistics, 224
- Cuneiform, 38
- Cupolas and Oil and Gas Resource Assessment, 1519
- C–upper-approximation, 1601
- Curse of dimensionality, 700

Curvelet transform, 1308
 Curvilinear structures, 1352
 Cutoff, 1638
 Cyber world, 876
 Cybernetic model, 1589, 1590
 Cyber-physical implementation, 878
 Cycloid, 1498
 Cyclostratigraphy, 647

D

2-D acquisitions, 625
 Dangermond, Jack
 education, 225
 professional service, 225
 DAQ devices, 226
 Darboux-type differential equation, 218
 4D archaeological geographical information system, 1622
 Darcy's law, 210, 389, 390, 955, 957, 1091, 1468
 Data, 369, 1168–1169
 acquisition, 226, 227, 229
 analysis, 236–237, 1002
 analytic process, 1322
 archiving, 236
 assimilation, 334
 collection, 235
 compatibility, 966
 compression, 248
 configurations, 547, 548
 consistency, 1362
 discovery, 859
 dissemination, 1020
 duplication, 234
 entry, 497
 framework, 294
 governance, 248
 integration, 248, 382
 integrity, 247
 manipulation, 232
 modelling tools, 232
 owner, 370
 points, 307
 processing, 49, 236
 product, 237
 provenance, 236
 publication, 237
 quality, 247
 reduction techniques, 974
 redundancy, 234
 repository, 236
 resources, 370
 science, 774
 science approaches, 91
 security, 248
 sharing, 237
 sources, 1541
 storage, 248, 249
 structures, 234
 usage license, 371
 visualization, 663
 warehouse, 498
 Databases, 497
 construction, 1536
 system, 231
 Data-based training images, 965
 Database management system (DBMS) catalog, 231
 Data Catalog Vocabulary, 859

Datacite, 235
 Data dictionary(ies), 230–234
 anomaly detection, 233
 application design, 233
 components of, 232
 data integration, 233
 data quality, 233
 data transparency, 233
 in geoscience, 234
 in NLP, 233
 Data-driven, 1166
 approach, 856, 1542, 1543
 techniques, 865
 training image, 965, 966
 Data lakes, 233, 498
 structure and content, 234
 Data life cycle, 1018
 conceptualization, 235
 data analysis, 236–237
 data archiving, 236
 data collection, 235
 data processing, 236
 data product, 237
 data sharing, 237
 definition, 235
 Data management, 858
 and stewardship, 372
 Data Management Center (DMC), 125
 Data management life cycle. *See* Data life cycle
 Data management system
 benefits, 247
 big data management, 249
 definition, 247
 governance, 248
 integration, 248
 metadata management, 247
 quality, 247
 remote sensing data, 249, 250
 security, 248
 storage, 248, 249
 Data mining, 239, 240, 329, 982, 1600
 AC analysis, 239
 CA, 240
 cluster analysis, 240
 definition, 238
 geoscience data, 238–239
 outlier analysis, 240
 RA, 240
 Dataset, 370, 758
 rankit, 764
 Dataset catalog recommendation (DCAT), 89
 David, Michel
 academic career, 250
 education, 250–251
 geostatistics, role and contribution to, 251
 Davies-Bouldin index, 448
 Davis, John Clements, 251, 252
 2D binary image, 594, 595
 1D CNN architecture, 1274
 3D column vectors, 337
 3-D cube, 626
 2-D data matrix, 625
 Dead-leaves model, 566
 Dead reckoning, 342
 Debye's theorem, 1457, 1458
 Decay curve, 52
 Decay rate, 1465

- Decimation by 2, 1307
- Decision rules, 1604
- Decision tree, 257, 261, 465, 783, 1062, 1182
 - definition, 257
 - Eastern Himalaya, 258–260
 - practical implications, 258–261
 - software tools, 259
 - stability in, 258–259
 - Western Himalaya, 260–261
- Decomposition, 622, 1514–1515
- Deconvolution principle, 1295
- Deduction
 - definition, 644
 - vs. induction, 645
- Deep Carbon Observatory (DCO), 90
- Deep CNN applications in geosciences, 266
- Deep convolutional neural network architectures, 262, 265
- Deep crustal seismic (DCS), 676
- Deep learning, 262, 267, 324, 498, 989
 - AE, 270
 - algorithms, 262
 - architecture, 263
 - CNNs, 268–269
 - definition, 267
 - GNNs, 270
 - methods, 1047
 - RNNs, 269–270
 - using meta-heuristics, 1322
- Deep neural network (DNN), 31, 36, 577, 777
 - algorithms, 262
 - models, 1274
 - inverse model, 1262
- Defuzzifier, 450, 455, 462
- Degree of freedom, 206, 1114
- Degree of variability, 1240
- Delaunay tessellations, 1508
- Delay line technique, 290
- Dempster-Laird-Rubin method, 355
- Dendrites, 1264
- Dendritic rivers, 434
- Dendrogram, 977, 1159
 - applications in geosciences, 273
 - compositional data, 141–142
 - definition, 271
 - example of, 272
 - thresholds, 272
- Density, 818
 - logs, 1085
- Density-equalising map projection, 105
- Dependent feature, 1600
- Depletion pattern, 949
- Deposition, 1261
 - environment, 1511
 - facies analysis, 1115
 - history, 1461
- Depth, 257
 - estimation, from gravity and magnetic data, 1094
- Derivative(s), 1177
 - constraints, 1177
- Description logics, 1015, 1016
- Descriptive analysis, 236
- Descriptive approach, 1370
- Descriptive metadata, 657, 658
- Design of the experiments, 1500
- Destructive plate boundaries, 350
- Detailed coefficients, 1307
- Detection, 266
- Determinant of a matrix, 839, 841
- Determined system, 818
- Deterministic chaos, 25
- Deterministic experiments, 1188
- Deterministic methods, 539
- Deterministic nonperiodic flow, 111
- Deterministic signals, 1289
- Deterministic trend, 50
- Detrended fluctuation analysis (DFA), 622, 1215, 1217, 1311–1314
 - See also* Multifractal DFA (MFDFA)
- Detrended Hurst analysis, 1215–1216
- 2D Euclidean object, 1281
- 3D Euclidean object, 1281
- Deutsch, C.V., 1188
- Deviance, 777
- Devonian chronostratigraphic scale, 509
- Devonian spline, 507
- Dewatering techniques, 490, 493
- De Wijs, H.J., 252
- De Wijs data, 252–256
- De Wijs model, 252, 255, 256
 - binomial, 252
 - definition, 252
 - iterative segmentation process, 254
 - multifractal analysis, 255
 - multifractal characteristics, 254
 - in two-dimensional space, 254
- 1D Fourier transform, 1637
- 2D Fourier transform, 1637
- Diagonal averaging, 1329
- Diagonal matrix, 837
- Dice, 1119
- Dictionary, 1014, 1016
- Dictionary of Mathematical Geosciences*, 1219
- Dielectric permittivity, 29, 30
- Difference equation, 293
- Difference of Gaussian (DOG), 287
- Difference of normal (DoN), 737
- Differential effective medium (DEM) approximation, 1464
- Differential entropy, 348
- Differential equations, 274, 275
 - astronomical applications, 276
 - eigenvalues, 276
 - history of, 276
 - notation, 276
 - numerical methods, 276
 - ordinary differential equation, 275
 - partial differential equations, 275, 277
 - physical phenomena, 277
- Differential interferometric SAR (D-InSAR), 580
- Differential solution, 1263
- Differentiation pattern, 1055
- Different Team, Different Experimental Setup. (DTDE), 1210
- Different Team, Same Experimental Setup (DTSE), 1210
- Diffusion, 209, 1464
 - coefficient, 1465
- Diffusion limited aggregation (DLA) model, 26, 434
- Diffusion-type models, 555
- Diffusivity, 320, 1463
- Digital approach, 1009
- Digital cartography, 521
- Digital elevation maps (DEM), 267
- Digital elevation model (DEM), 477, 579, 633, 668, 736, 738, 929, 1135, 1144, 1243
 - applications of photogrammetric adjustment, 282–284
 - finite element method, 288
 - formats, 289

- Digital elevation model (DEM) (*cont.*)
 Gerchberg algorithm, 289
 image matching, 285–288
 photogrammetric adjustment of aerial stereo photo blocks, 284
 polynomial fitting, 288
 RPM, 281–284
 RSM, 280–281, 283
 satellite photogrammetry, for line-scanner imagery, 280–284
 UAV, SfM approach, 284–285
 weighted average, 288
- Digital filters
 definition, 290
 transfer function, 292
 zero-and pole design of, 293
- Digital Fourier transform, 291
- Digital geological map and database, 294
- Digital geological mapping, 294
 process, 294
 software systems, 294
 system, 294
- Digital images, 1203
- Digital model, 296
- Digital number (DN), 300, 308, 1008, 1205, 1514, 1537
- Digital object identifiers (DOIs), 87, 369, 656
- Digital processing, 49
- Digital recorders, 53
- Digital resource, 370, 656
- Digital surface model (DSM), 736
- Digital terrain mapping, 480–481
- Digital terrain model (DTM), 633, 736, 738, 929, 1623
- Digital twin, 875, 876
- Digital Twins of the Earth, 297–298
 cyber-physical model, 296
 definition, 295
 essential concepts and components, 295–296
 significant research and innovation applications, 296–297
- Digitization
 coding, 1296
 quantification, 1296
 sampling-holding, 1296
 steps, 1295
- Dilation, 93, 595, 924, 1515, 1537
 operator, 598, 905, 909
 parameter, 1304
- Dilution model, 566
- Dimensional analysis, 24
- N -Dimensional euclidean space, 18
- Dimensionality, 1221, 1468
- Dimensionality reduction, 35, 308, 338, 1527
- n -Dimensional lattices, 1464
- Dimension factor, 406
- Dimensionless measures
 coordinate system, 299
 CR, 301
 definition, 299
 dimensionality, 299
 earth's surface, 299
 error matrix, 302
 geoscience, 300
 geoscience datasets, 299
 physical quantities, 299
 pixel and pixel value, 300
 scale, 300, 301
 spectral index/band ratio, 302
 unit conversion, 299
- Dimension matrix, 300
- Dipmeter advisor, 38
- Dip-slip events, 330
- Dirac comb function, 1290
- Dirac delta function, 716, 1300
- Dirac pulse function, 1290
- Direct correlation, 978
- Direct determination, 414
- Directed governance, 536
- Direct inversion methods, 681
- Directional derivatives, 1177
- Directionally agonistic function, 774
- Directional structuring elements, 1515
- Directions, 1055
- Direct sampling (DS), 963
- Dirichlet-Thiessen-Voronoi cells, 219
- Disaster management, 1362
- Discontinuity, 405, 1511
- Discontinuous Galerkin (DG) methods, 472
- Discovery, 166, 1054
- Discrete approximation, 22
- Discrete diffusion isofactorial model, 555
- Discrete element method (DEM), 525, 575, 1233
- Discrete Fourier transform (DFT), 291, 1093
- Discrete fracture network (DFN), 1230, 1504
 model, 1231, 1232
- Discrete Gaussian model, 557, 558
- Discrete probability distribution, 1459
- Discrete probability function, 1377
- Discrete prolate spheroidal sequences (DPSSs), 303
 applications in spectral estimation, 305–307
 in geosciences, 307
 properties, 304
 software implementation, 307
- Discrete random variable, 1189
- Discrete scale invariance, 1469–1470
- Discrete smooth interpolation (DSI), 1543
- Discrete-time Fourier transform (DFT), 304
- Discrete-time signal, 290
- Discrete wavelet transform (DWT), 378, 1305, 1307, 1315, 1628
- Discriminant analysis (DA), 114, 130, 307, 975, 977, 982, 1005
- Discriminant function, 977
 analysis, 311
- Discrimination problem, 700
- Discrimination process, 312
- Disjunctive kriging, 554, 556
- Disordered objects, 1004
- Dispersal of organisms, 1219
- “Dispersion on a Sphere”, 388
- Dispersion variance, 1608, 1609
- Disruptive approach, 1505
- Dissector, 1499
- Distance(s), 1055
 cartogram, 105
- Distance-based indices, 1119
- Distortions, 627
- Distribution, 620, 1055, 1276
 function, 722, 1346
 of pixels, 310
- Distributivity, 822
- Disturbed days, 1314
- Divergence, 323
- Divergent plate boundaries, 350
- Divisive hierarchical approach, 982
- DK equations, 554–558
- 3D model, 876
- 3D model from multi-view Images of mock piles, 879

3D Model Reconstruction from Multi-view Images
 (Structure from Motion), 877
DnQm scheme, 1466
 Dolomite, 818
 Domain ontology, 1015
 Domain-relevant community standards, 371
 Dominant frequency, 430
 Don't care condition, 924
 Dot map, 481
 Dotplot, 1257
 Double branching models, 1484
 Double-logarithmic plot, 26
 Double Pareto-lognormal model, 425
 Double trace moment analysis, 443
 Doveton, John Holroyd, 317–318
 Down-sampling, 264, 1308
 scheme, 1631
 Downscaling, 585, 967
 Downward continuation, 1094
 3D point coordinates computation, 283–284
 3D points via triangulation, 877
 Drainage area, 25
 Drainage basin, 26, 633, 1141
 characteristics, 1140
 Drainage migration, 433
 Drainage networks, 431, 433, 1141
 Drift, 545, 1346
 analysis, 495
 Drill and practice programs, 165
 Drilling, 381
 2D singular spectrum analysis (2D-SSA), 1332
 3D structural geology, 1177
DT signal, 1645
 Dual-continuum model (DCM), 580
 Duality property, 909
 Dublin Core Metadata Initiative, 858
 Dunn's index, 448
 Durbin-Watson statistic, 1625
 Durbin-Watson test, 1434
 Dutch-Finnish Ozone Monitoring Instrument
 (OMI), 1383
 3D visualization, 1540
 2D wave-number, 1636–1637
 Dyadic scales, 1308
 Dyadic Wavelet Substitution (DWS), 379
 Dynamical approach, 1371
 Dynamical systems, 754
 Dynamic factor model, 1557
 Dynamic processes, 404
 Dynamic programming, 918–919
 Dynamic range, 53

E

Early computer engagement
 code-based data science, 497
 data analysis, 495
 data processing, 497
 intelligent stage, 497
 mainframe computers, 495
 reproducible science, 495
 R statistical programming language, 495
 RStudio, 495
 supercomputers, 495
 EarthChem, 236
 Earth crust, temperature distribution, 274

Earth geometry, 804
 Earth observation, 628, 629
 Earthquake, 39, 328, 756, 1599
 degree of heterogeneity, 332
 events, 330
 faults, 336
 fractals, 331
 hazard mitigation, 336
 magnitude, 329, 330, 334, 1311
 prediction, 329, 330
 scaling law, 329, 332
 self-organization, 335
 sensitivity, 328
 Earth resources, 1203
 Earth science big data initiatives, 41
Earth Sciences History, 512
 Earth's continents, 330
 Earth's crust, 1261
 Earth shaking, 335
 Earth's landscape, 432–433
 Earth's mantle, 331
 Earth surface, 1029
 Earth surface processes (ESP), 585, 1029, 1135
 bedrock channel evolution, 320–321
 definition, 319
 differential equation, 319–321
 exponential function, 321
 fractals in, 322
 hillslope evolution, 320
 machine learning application in, 323–324
 matrices, 322
 numerical methods, 323
 power-law function, 321–322
 statistical approaches in, 323
 thresholds, 319
 Earth system science
 Anthropocene, 326
 definition, 325
 Earth critical zone, 326
 economic development, 327
 geological diversity, 326
 goal of, 325
 history of, 325
 Horizon 2020 program, 327
 information processing and expression, 327
 infrastructure, 327
 multi-directional relationship, 327
 research tasks, 326
 resource utilization, 327
 tipping elements, 326
 Easy post-processing, 294
 Ecology, 39
 Economic benefit, 1271
 Ecosystem engineering, 536
 Edge-detection, 923
 algorithms, 462
 Edge Gray-Value Rate (EGR), 461
 Edge preservation degree (EPD), 1081
 Edge-weighted graphs, 94
 Effective bulk moduli, 1464
 Effective conductivity, 1463
 Effective electrical conductivity, 1465
 Effective field, 1463
 Effective governance, 536
 Effective medium (EM) approximation, 1463–1464
 Effective physical properties, 1459

- Effective porosity, 389, 1084, 1087, 1091
- Effective shear moduli, 1464
- Effective thermal conductivity, 1467
- Effect of information, 557
- Effect of support, 557
- Efficiency, 1545
- Egocentric globe, 1620
- Eigenfunctions, 1041
- Eigenimages, 686, 687
- Eigenproblems, 276
- Eigenspace, 337
- Eigenvalue(s), 309, 337, 338, 384, 840, 1026, 1027, 1031, 1041, 1121, 1221
 - problems, 1026
- Eigenvalue decomposition, 337, 338
 - of covariance or correlation matrix, 643
- Eigenvector(s), 337, 338, 384, 840, 1026, 1031, 1123, 1220
- Eigenvector-based methods, 35
- Eigenvector-based ordination methods, 312
- Einstein's formula, 1465
- Einstein's rule, 1465
- Elasticity theory, 1460
- Elastic properties, 65, 1464
- Elastic rebound theory, 334, 335, 1477, 1478
- Elastic tensor, 1464
- Electrical resistivity, 675, 677
- Electric conductivity, 53, 1460, 1464
- Electric power generation, 576
- Electric tortuosity, 1458
- Electrofacies, 339, 341
 - definition, 339
 - non-parametric and neural network methods, 340
 - parametric methods, 339–340
 - supervised method, 339
 - unsupervised approach, 339
- Electromagnetic (EM) radiation, 1206, 1207, 1307
- Electron backscatter diffraction (EBSD), 216, 1499
- Electronic engineering, 37
- Elementary arithmetic operations, 276
- Element contour maps, 169
- Elements or entries of matrix, 836
- Elimination, 1637
- Elliptic geometry, 668
- El Niño Southern Oscillation (ENSO), 46
- Elongation, 1496
- Elura Ag-Pb-Zn sulfide deposit, 360
- Embedding dimension, 430
- Emergent behavior, 535
- Emission distribution, 792
- Emissivity, 578
- Empirical laws
 - aftershock frequency, 1474
 - magnitude-frequency relationship, 1472, 1474
 - maximum magnitude, 1474
 - scaling, 1474, 1475
- Empirical mode decomposition (EMD), 1297, 1298
 - advantage, 1315
 - data adaptive, 1315
 - fractal and multifractal studies, 1315
 - IMFs, 1315–1317
- Empirical Monte Carlo, 894
- Empirical power-law function, 1143
- Empirical risk minimization, 774
- Endmembers, 739
- Endurant, 1015
- Energetic lattices, 919–920
- Enriched pattern, 949
- Ensemble, 1277
 - methods, 778
 - of trees, 1182
- Ensemble clustering based t-distributed stochastic neighbor embedding (EC-tSNE), 1530
- Ensemble EMD (EEMD), 1316
- Ensemble Kalman filter (EnKF), 79, 84, 343–344, 893–894
 - definition, 342
 - NS-EnKF, 345
 - with augmented state, 344–345
- Entities, 1536
- Entity-relationship (ER) diagram, 1014
- Entity-relationship model, 1016
- Entropy, 26, 1055, 1462, 1463
 - applications, 349
 - conditional, 347
 - definition, 346
 - differential, 348
 - historical development, 346
- Environmental protection agency, 1441
- Environmental Systems Research Institute (Esri), 225, 497
- Epanechnikov kernels, 1596
- Epidemic-type aftershock sequence (ETAS) model, 1599
 - applications, 1483
 - criticality, 1479
 - earthquake clusters and hypocenter depth, 1483
 - location-dependent parameters, 1483
 - microseismicity, 1480
 - rate-and-state friction law, 1483
 - secondary aftershock clustering, 1479
 - self-exciting Hawkes processes, 1479
 - self-similar, 1484
 - simulation, 1482
 - space-time estimation, 1482
 - stability, 1479
 - temporal conditional intensity, 1479
 - temporal to spatiotemporal, 1481
- Equator Award of the German Rowing Union, 1540
- Equifinality, 26, 27
- Equilibrium problems, 276
- Equipossibility, 645, 650
- Equiprobable, 1196
- Equi-spaced cardinal splines, 1406
- Equivalent barotropic atmosphere model, 584
- Ergodic, 1461
- Ergodicity, 1187
- Erodibility, 321
- Erosion, 93, 595, 1515, 1537
 - in the fluvial system, 1138
 - operator, 909
- Error, 682, 684, 685
 - covariance, 343
 - function, 774
 - matrix, 302
 - vector, 681
- Error-minimization criterion, 40
- Essential zeros, 138
- Estimation, 620, 1277
 - bias, 1427
 - of camera position and orientation, 877
 - error, 1607
 - variance, 549, 1607, 1608
- Estimator, 79–81, 622
 - linear, 81
 - minimum-variance, 80
 - unbiased, 80, 81
 - of the variance, 1607

- Ethernet, 226
 - E-type models, 1197
 - Eucalyptus, 126
 - Euclidean, 1120
 - coordinate system, 1282
 - distance, 128, 288, 695, 1587
 - distance in K-means, 698
 - geometry, 996
 - plane, 821
 - space, 613, 991, 1009
 - dimension, 414, 997
 - space, 934, 935
 - Euler poles
 - absolute, 353
 - and axis of rotation, 351
 - defined, 350
 - path of migrating, 353–354
 - Euler's formula, 1561
 - Euler's rotation theorem, 805
 - European Space Agency (ESA), 405
 - Evaluation metrics, 308, 310–311, 758
 - Even functions, 1347
 - Even polynomial, 1347
 - Event detection, 36
 - Every Earthquake is a Precursor According to Scale (EEPAS), 1484
 - Evidential map, 864, 865
 - Evolution, 1219
 - of landscape topography, 433
 - Evolutionary algorithms, 1022
 - Evolving computational systems
 - development stages, 495
 - exploration, 495
 - formative, 494
 - geostatistics, 495
 - origins, 494
 - petrography, 495
 - Exascale, 498
 - Exchangeability, 1431
 - Exocentric globe, 1620
 - ExoMars Phase 1 Mission, 40
 - ExoMars Phase 2 Mission, 40
 - Expectation-maximization (EM) algorithm, 1273
 - definition, 355
 - remote sensed images, 356–357
 - sklearn package in Python, 357
 - Expected value, 80, 342, 1186, 1276, 1605, 1607
 - Experiment, 1276
 - covariances, 548
 - semivariogram, 548
 - Expert capabilities, 36
 - Explained sum of squares (ESS), 1569
 - Explanatory variables, 381, 385, 386
 - Explicit method, 1046
 - Exploitation, 1021
 - Exploration geochemistry, 864
 - geology using remote sensing data, 457–460, 462
 - Exploration-level drilling, 1169
 - Exploratory analysis, 236
 - Exploratory data analysis (EDA), 91, 364, 1428–1430, 1575
 - data standardization and transformation, 367
 - graphical analysis, 365–366
 - multivariate relationships, 368
 - univariate data, robust classification of, 366–367
 - Explorer, 38
 - Exponential atmosphere, 584
 - Exponential function, 321
 - Exponential loss, 777
 - Exponentially weighted moving average (EWMA), 935
 - Exponential models, 1349
 - Exponential variogram, 1613
 - Expressivity, 1017
 - Extended Kalman filter (EKF), 343, 1050
 - Extended Normal Equation Simulation (ENESIM), 963
 - Exterior orientation parameters, 278
 - Exterior parts, 1562
 - Extraction and matching of feature points in images, 877
 - Extract, transform and load (ETL) technique, 232
 - Extremely randomized forests, 1183
 - Extreme phenomena, 321
 - Extreme values, 1187, 1276
- F**
- Facies, 339
 - Factor analysis (FA), 643, 982, 1337
 - Factor graphs, 38, 797–798
 - Factorial hidden Markov model, 792
 - Factor rotation, 1124–1125, 1224
 - FAIR data principles, 236, 369–372
 - FAIRness, 372
 - Fan-beam image reconstruction, 58
 - filtered backprojection fan-beam algorithm, 59–60
 - reconstruction without rebinning, 59
 - reconstruction with rebinning, 58
 - Fast appraisal system for petroleum resources (FASP), 1518
 - Fast Fourier transform (FFT), 291, 372, 1093, 1296, 1575
 - bore-holes, 375
 - contour map, 376
 - drill-hole data, 373
 - drill-holes, 375
 - exploration, 375
 - fault systems, 372
 - frequency domain, 372
 - harmonic analysis, 376
 - harmonic trend analysis, 372
 - interpolation, 372
 - longitudinal cross-section, 373
 - O notation, 372
 - topographical contour map, 376
 - underground mining data, 376
 - FastPAM, 699
 - Fast wavelet transform
 - classification of hyperspectral images, 378, 379
 - daughter wavelets, 377
 - definition, 376
 - geospatial applications, 378
 - multi-resolution analysis, 377
 - scale components, 377
 - scale parameter, 377
 - spatio-temporal characteristics, 379
 - urban analysis, 379
 - Fat curve, 1351
 - Fat-tailed distribution, 321
 - Fault scarp, 320
 - Fault-trace-oriented singularity mapping technique, 746
 - Favorability analysis, 381, 383–386
 - Favorability estimation, 383
 - Favorability function, 385
 - Favorability maps, 1184
 - FCM algorithm, 447, 448
 - FCM clustering algorithm, 448
 - F-distribution, 376
 - Feature axis, 759
 - Feature extraction, 1064

- Feature learning, 36
- Feedback, 434
 - signal, 774
- Feldkamp's algorithm, 60–61
- F-function, 1058
- Fibre processes, 1503
- Fick's first law of diffusion, 397
- Field data acquisition module, 294
- Field data cleaning module, 294
- Field data collection, 294
- Field duplicates, 1132
- Filter, 902
 - analysis, 292
 - bank, 1307, 1308
- Filtered anomaly, 1639
- Filtered backprojection algorithm, 58
- Filtering, 709, 1295, 1633–1635
- Financial economics, 1517
- Findable, 369–370
- Findable, accessible, interoperable, reusable (FAIR) principles, 86, 88, 89, 239, 656, 1018
- Finite difference (FD), 323
- Finite difference method (FDM), 212, 585, 1045–1046, 1465
- Finite element method (FEM), 212, 288, 525, 711, 1045, 1046, 1233
- Finite-rank time series, 1330
- Finite set, 1329
- Finite volume (FV), 323
- Finite volume method (FVM), 525
- Firefly algorithm, 1048
- First appearance datum (FAD), 193
- First discriminant function, 977
- First order condition, 176
- First-order intensity function, 421
- First-order logic, 1015, 1017
- First-order moment, 1605
- First-order necessary condition (FONC), 176
- First order structure function, 442
- First-order transitions, 1598
- Fisher, R.A.
 - books, 387
 - carriers, 387
 - correlation coefficient formula, 387
 - distribution, 119, 388, 1230, 1487
 - geology, 388
 - mathematical modeling, 388
 - mathematical statistician, 387
 - remanent magnetization, 388
 - scoring algorithm, 691
 - statistical appendix, 388
 - (two-group) linear discriminant analysis, 313
- Fisher-Kolmogorov equations, 703
- Fitness function, 1048
- Fitness value, 466, 1048, 1049
- Fixed points, 1597
- Flat structuring, 600
 - element, 904
- Flowing wells, 652–654
- Flow in porous media, 1031
- Flow path, 1458
- Flow routing algorithms, 933
- Flow simulation, 967
- Fluctuation(s), 623
 - function, 1313
- Fluid flow, 211, 756, 1232
 - pattern, 580
 - simulation, 580
- Fluid sediment, 1261
- Fluorescence Explorer (FLEX), 405
- Fluvial landscape and associate processes and landforms, 1138
- Fluvial systems, 1138, 1139
- Focused ION beam (FIB-SEM), 1499
- Foraminifers, 1219
- Forecasting, 328, 1371
- Foreshock, 330, 334
- Foreshock hypothesis
 - clustering model, 1483
 - induced seismicity, 1483
- Forest cover type dataset, 356
- Forest inventory and analysis (FIA) monitoring, 701
- Forestry, 483
- Formalism, 622
- Formality, 1016
- Formal neuron, 1264
- Formal shape, 1280
- Formation, 1465
 - factor, 1465
- Form compositional biplot, 140
- FORTTRAN compilers, 34
- FORTTRAN IV program, 397
- Forward model, 539, 540, 681
- Forward problem, 63, 672, 681
- Forward projection process, 22–23
- Forward stratigraphic modeling, 394–395
 - geometric models, 396–397
 - hydraulic process-response models, 398–399
 - mixed algorithmic, heuristic and fuzzy-logic models, 399
 - topography-control model, 397–398
- Fossil, 38
- Fourier analysis, 1093
- Fourier domain, 291, 292
- Fourier expansion, 438, 1470
- Fourier series, 1291
- Fourier slice theorem, 57
- Fourier transform (FT), 212, 291, 526, 620, 716, 1043, 1093, 1294, 1295, 1297, 1308, 1626, 1627
 - application, subsurface anomaly parameters, 1298, 1299
 - of density/susceptibility distribution, 1094
 - of gravity/magnetic potential, 1094
 - vs. Hilbert transform, 1302
 - properties, 1298
 - time-varying signal, 1298
 - vs. Z-transform, 1300
- Fourth-order spatial cumulant maps, 611
- Fractal(s), 27, 430, 513, 514, 516, 517, 928, 945, 995–997
 - definition, 430
 - density, 814
 - in ESP, 322
 - methods, 145, 146, 744
- Fractal dimension, 27, 404, 430, 1309
 - estimation, 408–410
- Fractal geometry, 27, 169, 404–406, 587, 997, 1308, 1333, 1598, 1599
 - for SIF-downscaling, 405–406
- Fractal geometry, in geosciences
 - fractal dimension estimation, 408–410
 - fractal/multifractal modeling of point patterns, 419–423
 - fractal perimeter-area relation, 411–412
 - geochemical anomalies, separation of, 410
 - lognormal and Pareto size-frequency distributions for mineral resources, 423–425
 - model of de Wijs, 416–417
 - multifractal spatial correlation, 417–419
 - multifractal spectrum, 412–414
 - singularity analysis, 414–416

- Fractal Models in the Earth Sciences, 704
- Fractal/multi-fractal analysis, 145, 1297, 1298
 - Cartesian reference frame, 1309
 - characteristics, 1308
 - definition, 1308
 - DFA and MFDFA, 1311–1314
 - geometrical shapes, 1310
 - integer and non-integer dimensions, 1310
 - length, 1308
 - observational data, 1310
 - R/S analysis, 1311
 - scaling exponents, 1310, 1311
 - statistical properties, 1310
 - time series analysis, 1310
 - window lengths, 1310
- Fractional Brownian motion (fBm), 620, 621, 1214, 1245, 1576
- Fractional Brownian noise (fBn), 1214
 - of Hurst exponent H , 1217
- Fractional derivative, 320
- Fractional differencing parameter, 1217
- Fractional Gaussian noise, 1217
- Fraction Brownian motion, 431
- Fracture density, 1231
- Fracture distributions, 579
- Fracture length distribution, 1230
- Fracture orientation, 1230
- Fracture propagation, 756, 1233
- Fracture shape, 1233
- Fragments, 25
- Fréchet space, 1517, 1518
- Free energy, 1466
- Free-hand sketch, 785
- Frequency, 404, 845
 - analysis, 1643
 - histogram, 1428
 - localization, 1303
 - representation, 1290, 1291
 - shift, 849
- Frequency distribution
 - definition, 435
 - forms of, 435–436
 - statistics, 436–437
- Frequency-wavenumber analysis
 - definition, 437–438
 - sampling of rocks, 438
 - short-distance nugget effect modeling, 439–441
 - universal multifractals, 441–443
- Frictional force, 335
- Frictional model, 333
- Friction parameters, 329
- Friedel's law, 216
- Frobenius matrix, 1329
- Full-conditional, 563
- Full-dimensional random sets, 1506
- FULL NOrmal Plot (FUNOP), 445, 446
- Full symmetry, 1380
- Fully connected layers, 265
- Fully Constrained Least Squares (FCLS), 178
- Fully coupled approach, 212
- Fully discrete methods, 1046
- Fully separable, 1348
- Functional autoregressive (FAR) models, 1559
- Functional dependencies, 1517
- Function of partition, 623
- Fundamental morphological operators, 595
- Funk-Radon transform, 216, 217
- Fuzzification, 455, 462
 - Fuzzification parameters, 457
 - for the input layers, 459
 - Fuzzifier, 462
 - FUZZIM, 399
 - Fuzzy algebraic product, 455
 - Fuzzy AND operator, 458
 - Fuzzy C-means clustering, 447, 448
 - Fuzzy gamma operator, 454
 - Fuzzy inference system
 - architecture of, 450–452
 - case studies, 453
 - definition, 449
 - examples, 450–452
 - for mineral exploration, 449
 - types, 450
 - Fuzzy logic, 449, 462
 - Fuzzy logic method, 456
 - Fuzzy logic modeling (FLM), 455, 456, 458
 - Fuzzy membership, 454
 - function, 450
 - values, 454
 - Fuzzy operators, 450, 455
 - Fuzzy rules, 450
 - Fuzzy set, 449, 454, 455
 - theory, 447, 455, 462, 865
- G**
- Gadolinium, 819
- Gain control, 49
- Gain ratio, 258
- Gain recovery, 52
- Galois correspondence, 827
- Gaming based CAI program, 166
- Gamma function, 716
- Gamma-gamma (“density”) log, 1216
- Gamma renewal model, 1477
- GARCH process, 1558
- Gas geochemistry, 579
- Gates-Gaudin-Schuhmann (GGS) distribution, 591
- Gaudin-Melloy (GM) distribution, 592
- Gaussian, 486
 - assumption, 960
 - atmospheric noise, 48
 - conditional distribution, 1278
 - distribution, 674, 961, 999, 1230, 1411, 1464
 - disturbances, 720
 - filter, 1638–1639
 - kernel, 1527
 - membership function, 855
 - mixture distributions, 66, 67
 - models, 1349
 - prior distribution, 70
 - process, 553, 1214
 - process regression, 960
 - radial basis kernel function, 993
 - random variable, 48
 - random walk, 1464
 - regressions, 960
 - scaling models, 1248
 - variogram, 1613
 - white noise process, 50
- Gaussian mixture model (GMM), 356
- Gauss-Markov conditions, 1035
- Gauss-Markov theorem, 1433
- Gauss plane, 28
- Gazetteer, 1014

- Gear failure diagnosis, 869
- Geary's index, 898, 899
- GEMAS project, 1225
- Gene expression, 1530
- General AI, 32
- Generalization, 1359
 - error, 1062, 1183
 - operators, 1361
- Generalized autoregressive conditional heteroscedastic (GARCH)
 - model, 1557
- Generalized bell membership function, 855
- Generalized correlation dimension, 432
- Generalized covariance(s), 1347
 - function, 545, 546
- Generalized cross-validation (GCV), 507
- Generalized eigenvalue problem, 337
- Generalized information dimension, 432
- Generalized interpolation, 1177
- Generalized inverse, 672
- Generalized linear models, 691
- Generalized scale invariance, 1251–1253
- Generalized spectrum, 432
- General linear model, 1567, 1574
- General Purpose Interface Bus (GPIB), 226
- Generative adversarial network (GAN), 270, 968
- Generic ontology, 1015
- Generic photogrammetric data adjustment approach, 278–280
- Generic simulation method, 1479, 1480
- Genetic algorithms (GAs), 465, 466, 895, 1048, 1050–1052
 - application, 468
 - in clustering, 466
 - for data clustering, 465
 - implementation steps, 466
 - initial population (randomly selected) and final population, 467
 - mathematical models, 465
 - with maximum fitness value, 468
 - unsupervised classification, 468
- Genetic inverse model, 400
- Genus, 1459
- Geo-anomaly analysis, 744
- Geobiodiversity database (GBDB), 236
- Geobrowsers, 1619
- Geochemical anomalies, 410
- Geochemical data, 366
- Geochemical distributions, 1333
- Geochemical logs, 819
- Geochemical mapping, 14
- Geochemical patterns
 - anomaly, 359
 - characterising and modelling, 359
 - deep learning algorithms, 363
 - design of, 359
 - lateral dispersion, 359
 - in mineral exploration, 361
 - mineralisation via, 359
 - modelling of, 359
 - multivariate methods, 360
 - spatial methods, 362
 - spatial patterns, 359
 - statistical and spatial analysis techniques, 359
 - univariate and bivariate methods, 359–360
- Geochemical reactions, 1031
- Geochemical sampling, 1116
- Geochemistry, 26, 437, 809–811, 1117, 1219, 1530
 - soil Jura, 1582
- Geochronology, 1028, 1029
- Geocomputing
 - applications/challenges, 470, 472
 - computational experiments, 468
 - computational models, 472
 - definition, 468
 - discretisation, 469
 - HPC, 469, 472
 - laboratory experiments, 469
- Geodesic(s), 1172–1173
 - dilation, 93
- Geodesy, 519
- Geodetic models, 804
- Geodiversity, 934
- Geo-environmental engineering, 573
- Geographical and temporal weighted regression (GTWR), 487
- Geographical coordinate system, 475–476
- Geographical information science (GISc), 586
 - accuracy assessment of spatial data, 482
 - applications, 482–483
 - definition, 473
 - digital terrain mapping, 480–481
 - map projections, 476
 - network analysis, 481
 - raster data model, 480
 - research area, 484
 - research challenges, 483–484
 - spatial data analysis, 478
 - spatial data concepts, 473–475
 - spatial data interpolation, 480
 - spatial data model, 476–477
 - spatial decision support system, 482
 - spatial relation, 478
 - statistical analysis, 481
 - vector data model, 478
 - visualization of data analysis, 481
- Geographically and temporally weighted regression (GTWR) model, 1387, 1391
- Geographically weighted regression (GWR), 487, 488, 1390, 1394
 - definition, 485
 - extensions, 487
 - model calibration, 486–487
 - model formulation, 486
 - model output, 487
 - predicted AQI surface, 488
 - principle of, 1391
- Geographic entities, 1343
- Geographic information science, 100
- Geographic information system (GIS), 225, 234, 520, 666, 809, 865, 867, 1241, 1343, 1560
- Geographic knowledge discovery, 1053
- Geographic Resources Analysis Support System (GRASS) GIS, 168
- Geography, 38, 39
- Geohazards, 1622
- Geoid, 1049
- Geoinformatics, 494
 - automation, 500
 - autonomous computing, 500
 - climate change, 499
 - open data, 499
 - private data, 499
 - qualitative, 494
 - quantitative geoscience, 494
 - robotics, 500
 - stratigraphy, 494

- Geological, 1224
 - big data, 1545
 - controls, 1167
 - environment, 208
 - history, 524
 - interfaces, 1179
 - knowledge, 1542
 - literature, 1536
 - map and database, 294
 - observations and measurements, 1542
 - ontology models, 1535
 - processes, 434, 994
 - structure, 687
 - structures and properties, 1542
 - systems modeling, 63
- Geological survey of Canada (GSC), 96, 512, 513, 810
- Geologic interpretation, 293
- Geologic Time Scale 2020 (GTS2020), 503
- Geologic time scale (GTS), 502, 516, 588
 - conventions and standards, 505
 - Cretaceous and Jurassic Periods cubic spline, 510–511
 - definition, 502
 - Devonian spline, 507
 - McKerrow method, 505–506
 - methods and ages, 503
 - Paleozoic “super-spline”, 507–510
 - smoothing splines, 506–507
- Geology, 39
- Geomagnetic quiet, 1314
- Geomagnetic solar quiet day (Sq) variations, 1297
- Geomagnetism, 1219, 1314
- Geomaterials, 573
- Geomathematics (GM)
 - atmosphere, 512
 - compositional data analysis, 514
 - conventional analysis of variance, 515
 - definition, 512
 - 2-D Gaussian distribution, 515
 - fractals, 516, 517
 - geochemistry, 512
 - geophysics, 512
 - geostatistics, 513, 514
 - GTS, 516
 - lognormal and pareto frequency distributions, 517
 - mathematical morphology, 515
 - mathematical statisticians, 512
 - mathematical statistics, 512
 - mineral resource estimation, 514
 - multifractals, 516, 517
 - planar features, 514
 - quantitative stratigraphy, 515, 516
 - sedimentary features, 515
 - statistical inference, 515
 - See also* Mathematical geosciences (MG)
- Geomatics, 512
 - applications, 521–522
 - components, 519–521
 - definition, 519
 - innovations in, 521
- Geomechanics
 - definition, 523
 - financial losses/humans life, 523
 - influence, 523
 - pore pressure, 525
 - research path, 525
 - stress field, 524
 - stress variations, 524
 - subsurface, 523
 - subsurface stresses, 524
 - vertical stress, 525
- Geometrical interpretation, 1034
- Geometrical model, 396–397
- Geometrical shapes, 1308
- Geometrical transitions, 1598
- Geometric anisotropy, 548, 1348, 1379
- Geometric features, 1545
- Geometric length scales, 1465
- Geometric parameter, 1233
- Geometric set theory, 814
- Geometric shape spaces, 1286–1287
- Geometric spreading, 51
- Geometric transformation, 478
- Geometry, 1540
- Geomodel, 339–341
- Geomodel-driven approach, 1542
- Geomorphic agent based processes, 1149
- Geomorphic agents, 324
- Geomorphic connectivity, 1147, 1148
- Geomorphic processes, 633, 1136
- Geomorphic systems, 319, 1135
- Geomorphologic attractors, 1149
- Geomorphologic Instantaneous Unit Hydrograph (GIUH), 1141
- Geomorphology, 25, 430, 432, 1307
- Geomorphometric methods, 929
- Geomorphometry, 634
- Geophysical data, 66, 679
 - inversion, 1051
 - in single dimension, 267
- Geophysical inverse problems, 65, 539, 675
- Geophysical inversion, 539–540, 682
- Geophysical signals
 - linear signal processing, 1298–1302
 - nonlinear signal processing, 1303–1317
 - nonstationary, 1297
 - observational data, 1298
 - stationary/quasi-stationary signals, 1297
 - statistical methods, 1297
 - time/space variation, 1297
- Geophysical survey, 62
- Geophysical systems, 1031
- Geophysical well-log data, 1314
- Geophysics, 39, 65, 1297
- Geo-rectification, 1241
- Geo-registration, 478
- Georgetown experiment, 34
- Geoscience(s), 7, 8, 238, 262, 265, 273, 380, 778, 857, 1500
 - data, 231, 238–239
 - knowledge, 1535
 - literature, 1535–1536
 - maps, 381
- Geosciences digital ecosystem
 - concepts and components, 534–536
 - definition, 534
 - evolutionary aspect, 536–537
 - metasystemic aspect, 537–538
- Geoscience signal extraction, 526, 533
 - Pulacayo Mine and main KTB borehole, 527–528
 - spectral analysis, 528–532
 - unit vector fields, 532
- Geoscientific, 679, 683–685, 688

- Geoscientific (*cont.*)
 analyses, 1561
 data, 678
- GeoSPARQL, 1015
- Geospatial, 499, 1053
 analysis, 240
 coherence, 144, 146, 147
 data, 338, 1387
 interpolation, 1583
 metadata, 858
 science, CAI programs, 168, 169
- Geo-spatiotemporal, 499
- Geo-statistical analysis, 670
- Geostatistical applications, 1001, 1348
- Geostatistical approach, 1373–1374
- Geostatistical data, 1345, 1354–1355
- Geostatistical elastic inversion, 540
- Geostatistical estimation, 3, 871
- GEOstatistical FRACTure simulation method (GEOFRAC), 1232, 1233
- Geostatistical interpolation methods, 1384
- Geostatistical methods, 146, 621, 1346
- Geostatistical process, 1364, 1365, 1368, 1369
- Geostatistical simulation, 871, 1196, 1543
 methods, 1323
- Geostatistical Soft-ware Library GSLib, 1351
- Geostatistical techniques, 1346
- Geostatistics, 50, 96, 145, 250, 513–515, 526, 576–577, 587, 637, 693, 704, 744, 802, 929, 934, 961, 966, 1065, 1345, 1347, 1609
 change of support, 556–559
 data configurations, 547, 548
 Earth Sciences, 543
 economic parameters, 543
 experimental covariances, 548
 generalized approach, 544
 mathematical models, 543
 panels, 544
 parametric models, 961
 positive definite, 961
 random variables, 544
 second-order stationarity, 544
 semivariograms, 548
 simulation, 544
 simulation methods, 961
 spatio-temporal behaviors, 960
 spatio-temporal data, 960
 statistical methods, 543
 statistical techniques, 543
 variogram, 961
- Geosyntax, 568, 570
 essential concepts, 569–570
 philosophical basis, 568
 two-and three-dimensional sedimentary bodies, extension to, 571
- Geotechnics, 573–575
- Geothermal energy, 457
 remote sensing, 577–580
 subsurface temperature modeling, 576–577
- Geothermal field, 577
- Geothermal reservoirs, 580
- Geothermal resource exploration, 457, 458
- GeoTiff format, 289
- Geovisualization, 661, 1056
- Gerchberg algorithm, 289
- German Continental Deep Drilling Program, 935
- G-function, 1058
- Gibbs algorithm, 790
- Gibbs distribution, 797, 798
- Gibbs phenomenon, 292
- Gibbs sampler, 563, 564, 798
- Gibbs sampling, 68, 789, 790
 algorithm, 788
 method, 790
- Gini importance, 1183
- Gini impurity, 1182
- Gini index value, 258
- Glacial erosion, 320
- Glacial landscape and processes, 1143
- Glacial systems, 1142, 1143
- Glaciers, 1142
- Glaciology, 39
- GLADYS expert system, 1103
- Global, 1493
- Global and regional climate models
 adiabatic atmosphere, 584
 ARDC, 584
 atmosphere model, 583, 584
 barotropic atmosphere model, 584
 classification, 583
 climate anomaly, 583
 climate prediction, 583
 climate system, 583
 climate system behavior, 583
 coupled ocean-atmosphere general circulation model, 583
 definition, 582
 equivalent barotropic atmosphere model, 584
 homogeneous atmosphere, 584
 isothermal atmosphere, 584
 vs. local, 585
 model calibration, 583
 standard atmosphere, 584
 thermotropic atmosphere model, 584
- Global Centroid-Moment-Tensor (CMT) Project, 334
- Global change information system (GCIS), 90
- Global change of support, 556
- Global chronostratigraphic scale, 503
- Global geostatistical seismic inversion, 540–542
- Global iterative geostatistical seismic inversion methods, 540
- Global minimum, 1466
- Global Navigation Satellite Systems (GNSS), 520
- Global observation mode (GOM) coverage, 1322
- Global optimization, 675
 nature inspired, 1048
 problems, 1047–1048
- Global outliers, 1055
- Global parameters, 1494
- Global positioning system (GPS), 520, 736
- Global regularity, 624
- Global search behavior, 1021
- Global seismicity 1900–2007, 331
- Global seismicity from 1977 to 2007, 334
- Global tectonics, 331
- Glossary, 1014, 1016
- GMT quantities, 1145
- Gneiting class of models, 1377
- 5G network technology, 35
- Gold deposit, 203
- Gold grade, 203
- Goodchild, Michael Frank, 586, 587
- Google Assistant, 34
- Google dataset search, 235
- Google Earth Engine (GEE), 125

- Google Earth project, 1620
 - GoogLeNet, 265
 - Gower's distance, 1120
 - Graded bedding, 1263
 - Gradient boosting models, 258
 - Gradient descent, 724–726, 1528
 - algorithm, 450
 - Gradstein, F.M., 588
 - Grain(s)
 - characteristics, 589
 - cumulative approach, 593
 - cumulative size distribution function, 589
 - embedded, 589
 - field disturbance effects, 589
 - histogram, 590
 - imaging particle analysis, 592
 - incremental method, 593
 - kernel estimators, 591
 - length parameters, 589
 - light scattering methods, 593
 - loose, 589
 - mass, 589
 - operators, 912
 - orientation, 592
 - pack, 1464, 1465
 - probability density function, 591
 - projected area, 589
 - projections, 592
 - quantity type r , 589
 - quartz, 593
 - sections, 592
 - sedimentation analysis, 593
 - settling behavior, 593
 - shape, 592
 - sieve analysis, 592
 - size analysis, 589
 - size definition, 589
 - spatial images, 592
 - stationary sedimentation rate, 589
 - stereology, 592
 - stokes diameter d_{st} , 589
 - surface, 589
 - tomography, 592
 - volume, 589
 - Grain size, 1458
 - analysis, 809
 - distribution, 25
 - Granite, 593
 - Granular media, 1462–1463
 - Granulometries, 93, 599, 600, 911
 - Granulometries in surficial roughness characterization, 1145
 - Graph based MM operations, 599
 - Graph-based morphological filters, 599
 - Graph cut optimization, 36
 - Graphical analysis, 365–366
 - Graphical procedures, 981
 - Graphic Information System (GIS), 497
 - Graphic processing units (GPU), 498
 - technology, 1619
 - Graphics processing unit (GPU)-based large-scale computations, 697
 - Graph mathematical morphology, 594–597, 599
 - Graph neural networks (GNNs), 270
 - Graph spaces, 596
 - Graph variational autoencoders (GAEs), 270
 - Gravimetric inversion, 1052
 - Gravitational accretion, 330
 - Gravity anomaly, 1299, 1302
 - Gravity data, 1638–1639
 - Grayscale image dilation, 599
 - 1964 Great Alaska Good Friday Earthquake, 334
 - Greedy algorithms, 1180
 - Greenhouse gases (GHGs), 1382
 - Green's function, 1043, 1464
 - method, 1027, 1032
 - Grey-scale images, 600, 601, 904–905, 907, 908
 - alternating sequential filters, 601
 - basic MM operators, 600
 - closing, 600
 - dilation, 600
 - erosion, 600
 - filtering on, 600–601
 - flat structuring, 600
 - geodesic dilation/erosion, 601
 - granulometries, 600
 - non-flat structuring, 600
 - opening, 600
 - segmentation, 601
 - skeletonization, 601
 - thinning/thickening, 601
 - Grid, 1267
 - search method, 685
 - Grid-drilling, 1591
 - Griffiths, John Cedric
 - impact on geology, 602
 - teaching ability, 602
 - Ground control points (GCPs), 1240
 - Ground coordinates, 278
 - of conjugate points, 278
 - estimation of, 281
 - Ground Instantaneous Field Of View (GIFOV), 626
 - Ground Stress Concentration in Underground Mine, 879, 880
 - Groundwater, 489
 - flow model, 491
 - mapping analysis, 1023
 - modeling, 967
 - potential, 1023
 - Grouped decomposition, 1329
 - Gutenberg-Richter empirical law, 334
 - Gutenberg-Richter formula, 756
 - Gutenberg-Richter law, 331, 1472, 1474, 1598, 1599
 - Gypsum, 818
- ## H
- Hack's law, 321
 - Hackathon, 39
 - Hack profile, 321
 - Half absolute relative difference plot, 1133
 - Half-life, 1029
 - Hammersley-Clifford Theorem, 1366
 - Hankel matrix, 1328
 - Hankel transforms, 212
 - Hanning window, 291
 - Harbaugh, John W., 603
 - Hard clustering, 447
 - Hard-core distance, 1074
 - Hard data, 72–74, 559
 - Harff, Jan, 604
 - Harmonic trend surface analysis, 372

- Harnessing the Data Revolution (HDR), 89
- Hat matrix, 1433
- Hausdorff dimension, 996, 997
- Hawkes models, 1482
- Hawkes' mutual excitation model, 1478
- Hazard function, 1475, 1476
- Head mounted displays (HMDs), 1620
- Heat conductivity, 53
- Heat flow boundary value problem, 1467
- Heat flux, 576
- Heat-pipe, 499
- Heat production, 576
- Heat transfer, 580
- Heat-window, 1470
- Heaviside operational calculus, 716
- Hercynian schistosity, 1487
- Hermite polynomials, 556, 558, 562
- Hertz-Mindlin contact theory, 542
- Heterogeneity, 634, 928
- Heterogeneous, 1345, 1541
 - roughness, 432
- Heterotopic, 547, 551
- Heuristic modeling, 219
- Heuristic optimization techniques, 1021
- Hidden layers, 263
- Hidden Markov model, 35, 1273
- Hierarchical clustering, 982
 - algorithms, 271
 - techniques, 977
- Hierarchical image matching (HIM), 286
 - image condensation, 286
 - interest point generation, 286
 - local mapping, 286, 287
- Hierarchical networks, 329
- Hierarchical/non-hierarchical/model-based clustering, 145
- Hierarchy of boundaries, 1004
- High-dimensional data, 127
- Higher-order eigenvalue decomposition (HOEVD), 643
- Higher-order moments, 432, 1276
- Highest confidence first (HCF), 798
- High Hurst exponent, 624
- High-order sequential simulations (*hosim*), 609
- High-order simulation methods, 608, 609
- High-order spatial cumulants, 606–608
- High-order spatial stochastic models, 605
- High-order stochastic simulation, 611
- High-pass filter, 51, 53, 1308
 - coefficients, 1308
- High-performance computing (HPC), 469, 1019
- High-permeability, 1276
- High-resolution earthquake shaking and rupture
 - modelling, 472
- High separability, 310, 311
- Hilbert, D., 615
- Hilbert-Huang transform, 1315
- Hilbert-Schmidt norm, 616
- Hilbert space
 - applications, 616–617
 - definition, 613
 - examples, 615–616
- Hilbert spectral analysis, 1315
- Hilbert transform, 1297, 1316
 - amplitude, 1301
 - cosine and sine waves, 1300
 - definition, 1301, 1302
 - Fourier transform relation, 1302
 - frequency domain representation, 1301
 - phase information, 1300
 - positive and negative frequencies, 1300
 - subsurface anomaly parameters, 1302
 - time-domain, 1301, 1302
- Hill-climbing algorithm, 1321
- Hill-climbing methods, 1320
- Hill-slope, 1029, 1030, 1137, 1138
 - smoothing, 433
- Hinge loss, 778
- Histogram, 435, 1441
 - analysis, 1305
 - method, 414
- History of geology, 1219
- Hit-and-miss operation, 924
- Hit and miss transform, 93, 1537
- Hölder exponent, 414, 432, 620, 1313
- Hole effect, 1349
 - models, 1349
- Holt-Winters multiplicative model, 1548
- Homogeneity, 1003
- Homogeneous atmosphere, 584
- Homogeneous turbulence, 785
- Homogenized thermal conductivity, 1467
- Hooke's law, 210
- Horizontal cylinder, 1302
- Horton, Robert Elmer, 618
- Hortonian overland flow, 618
- Horton's law, 25
- Hosim* algorithm, 610
- Hosim* realization, 610
- Hot spots, 330
- Howarth, Richard J., 619
- Huber estimator, 1435
- Huber loss, 775
 - function, 690
- Huber's method, 734
- Human-device communication, 35
- Human society and sedimentation, 1262
- Human via computer programs, 165
- Hurst, H.E., 1213
 - effect, 1213
 - index, 1631
 - phenomena, 1247
- Hurst exponent, 431, 1213–1215, 1311, 1313, 1314, 1576
 - definition, 620
 - detrended fluctuation analysis, 622
 - multifractional Brownian motion, 620–621
 - spectral density method, 620–621
 - variogram, 621–622
 - wavelet transform modulus maxima lines, 622–623
 - well-logs data, 623–624
- Hybridized metaheuristic optimization algorithms, 1024
- Hybrid learning, 450
- Hybrid PSO-DE method, 1024
- Hydraulic aperture, 1232
- Hydraulic conductivity, 389, 1460
- Hydraulic jet model, 394
- Hydraulic path, 27, 1457
- Hydraulic permeability, 53, 1467
- Hydraulic process-response models, 398–399
- Hydraulic radius, 25, 1468
- Hydraulic tortuosity, 1457, 1458
- Hydrocarbon exploration, 193
- HydroClient database, 239
- Hydroclimatology, 793

- Hydrogeological characterization, 489
- Hydrology, 25, 635, 793
- Hydromechanical coupling, 210
- Hydrothermal alteration minerals, 579
- Hydrothermal fluid, 576
- Hyperbolic cosine, 775
- Hyper parameters, 1183, 1528
- Hyperplane, 778
- Hyperspectral analysis, 740
- Hyperspectral (HS) imaging, 268, 626, 701, 1529
 - airborne/spaceborne, 627
 - applications, 628
 - atmospheric disturbances, 628
 - challenges, 627
 - classification, 628
 - dimensional spaces, 628
 - distortions, 627
 - geometrical disturbances, 627
 - internal sources, 627
 - noise, 627
 - pre-processing, 628
 - technical challenges, 628
- Hyperspectral remote sensing, 177, 864
- Hyperspectral sensors, 579
- Hyperspectral signature, 1207
- Hyponymy, 1013
- Hypotheses, 1053
- Hypothesis test, 205–207, 1625
 - applications in geosciences, 632
 - definition, 630
 - for several samples, 632
 - for single sample, 631
 - for two samples, 631
- Hypsometric analysis, 1140
- Hypsometric curve, 633, 635
 - and integral, 1140
- Hypsometric integral, 633
- Hypsometry
 - application in hydrology, 635
 - definition, 633
 - hypsometric parameters, 633–634
 - relevance in geomorphology, 634–635
- I**
- IBM-PC platforms, 38
- Idempotent operator, 826
- Identically and independently distributed (iid) normal variables, 1214
- Identification, 342
 - process, 313
- Identifier, 369
- Identity matrix, 837
- Identity operator, 902
- IGBADAT, 114
- IGSN 2040 project, 659
- IGSN Central Registry, 657
- Ill-conditioned system, 292
- Ill posedness, 63
- Image analysis, 1499
- Image-based geoscience data analysis, 599
- Image-based modelling and rendering (IBMR), 879
- Image condensation, 286
- Image coordinates, 281
- Image enhancement, 1308
- Image matching, 285–288
- Image processing, 448, 449, 924, 1307
 - applications, 357
- Image resampling, 1241
- Image resolution, 1500
- Image segmentation, 266, 1010
- Imbrie, John, 1119
- Immunology, 1529
- Implicit methods, 1543
- Implicit modelling, 1177
- Imprecise information, 1600
- Imprecision, 447
- Impulse, 1638
 - response of the system, 1295
 - signal, 1645
- Imputation, 637, 1183
 - CoDa, 639–642
 - missing data and nondetects, in geoscientific studies, 637
 - missingness mechanisms, 637–638
 - multivariate data imputation, 638–639
- Inception Net, 266
 - GoogleNet, 265
- Incomplete pole figures, 218
- Inconsistent, 1602
 - information, 1600
 - objects, 1602
- Incorporated research institutions for seismology (IRIS), 125
- Increasing operator, 909
- Incremental/true stochastic gradient descent, 726
- Independent and identically distributed (iid) disturbances, 720
- Independent and identically distributed (iid) observations, 728
- Independent component analysis (ICA), 144, 145, 308, 460, 642–644, 1110
- Independent features, 1600
- Independent, uniform and random (IUR), 1491
- Indexing, 36
- Index of proportional similarity, 1120
- Indicator, 1278
 - approach, 1323–1324
 - co-simulation, 560
 - covariance, 547
 - favorability analysis, 386
 - random field, 1606
 - random function, 1187
 - semivariogram
 - simulation, 348, 349, 560
- Indicator-covariance, 552
- Indicator-semivariogram, 552
- Indiscernibility, 1602
 - relation, 1601
- Individual conditional expectation plot, 1183
- Individual orientation measurements
 - estimation with, 218–219
 - grain modeling with, 219
- Induced seismicity, 1483
- Induction
 - classification, 257
 - vs. deduction, 645
 - definition, 644
- Inelastic absorption, 51
- Inequality, 1177
 - constraints, 1177
- Inference(s), 1439
 - engine, 34, 452
- Infiltration, 618
- Infinite number of looks prediction (INLP) filter, 1081, 1082
- Influence function, 1435

- Information density matrix, 673
- Information engineering and AI, 34
- Information entropy, 842
- Information gain, 1480
- Information synthesis, 386
- Information theory, 293, 842
- Informative prior distribution, 64
- Informatixon system (IS), 1600
- Infrastructure as a Service (IaaS), 124, 498
- Infrastructure for spatial information in the European Community, 1360
 - generalization, 1360–1361
- Ingenious rocks, 1497
- Inhomogeneous Poisson process, 1076
- Inner product, 615
 - space, 614
- Input layer, 1269
- Instantaneous-floating-point (IFP), 53
- Instantaneous frequencies, 1316
- Instantaneous silt formation, 1577
- Instrumental distribution, 563
- Integer programming, 1322
- Integral orientation measurements, 216–218
- Integral transforms, 320, 1042–1043
- Integrated modeling, 1544
- Integrated models, 1377
- Integrated product-sum models, 1377
- Integration, 1545
- Integrity constraints, 1054
- Intelligence testing, 1220
- Intensity, 1074
 - function, 1074
- Interacting processes, 209
- Interactive technique, 165
- Interesting, 1054, 1055
- Interfacial free energy, 1467
- Interfacial tension, 955
- Interior orientation parameters, 278
- Inter-laboratory bias, 1131
- Intermittency
 - codimensions and singularities, 1248–1251
 - spikes, 1248–1249
- Intermittent, 432
- Internal measurement unit (IMU), 736
- Internal node, 257
- International Association for Mathematical Geology (IAMG), 513, 603, 853, 857, 1219, 1234, 1624
- International Association for Mathematical Geosciences (IAMG), 16, 115, 602, 801, 1012, 1066, 1067
- International Association of Mathematical Geologists, 1642
- International Celestial Reference Frame (ICRF), 350
- International Generic Sample Number (IGSN)
 - adoption, 658
 - allocating agents, 657
 - concept, 656
 - end users, 657
 - feature, 656
 - functions, 656
 - history and evolution, 656, 657
 - implementation, 659
 - metadata portals, 657
 - samples, 656
 - syntax and metadata, 657, 658
 - technical architecture, 658
- International Geological Correlation Programme (IGCP), 516
- International Geo Sample Number (IGSN), 657
- International Gravimetric Bureau, 1638
- International Open Data Charter, 1018
- International reporting codes, 1126
- International Science Council (ISC), 801
- International Statistical Institute (ISI), 513
- International symposium for mathematical morphology (ISMM), 1280
- International Union of Geological Sciences (IUGS), 423, 513
- Interoperability, 1360
- Interoperable, 370–371
- Interpolation, 660–663, 865, 1181
 - mathematical modeling, 663
 - techniques, 146
 - type and application, 661
- Interpolation-extrapolation, 1215
- Interpolationsrechnung, 660
- Interpretation element, 1054
- Inter-quartile range (IQR), 366, 664, 666, 1435
- Interval estimate, 776
- Inter-well relationship, 1305
- Intra-field variability, 406
- Intrinsically corregionalized, 547
- Intrinsic correlation, 555
- Intrinsic hypotheses, 1347, 1606
- Intrinsic mode functions (IMFs)
 - degree of subsurface heterogeneity, 1316, 1317
 - estimation, 1315, 1316
 - wavelengths, 1315
- Intrinsic random function, 546
- Intrinsic random function of order (IRF-k), 546, 550
- Intrinsic self-similarities, 1311
- Intrinsic stationarity, 1365, 1411
- Inverse analysis, 1385
- Inverse discrete wavelet transform (IDWT), 1628
- Inverse distance weighted (ICW) method, 480
- Inverse distance weighted (IDW) interpolation method, 487
- Inverse distance weighting (IDW), 1610
 - anisotropy correction, 667
 - applications of, 671
 - definition, 666
 - genesis and historic development of, 666
 - gradient correction, 667
 - mathematical equation, 667
 - modelling, 670
 - neighbourhood size, 667
 - syntax of, 671
 - weighting function, 667
- Inverse Fourier transform, 292, 1298, 1300
- Inverse Laplace transform, 1027, 1042
- Inverse modeling, 344
- Inverse of matrix, 839
- Inverse operator, 683
- Inverse probability principle, 852
- Inverse problems, 63, 81, 83, 84, 322, 678, 679, 681–685, 688, 791
 - nonlinear, 83, 85
 - principal component geostatistical approach, 84
 - successive linearization, 84
 - unknown mean, 83
- Inverse solution, 818
- Inverse stratigraphic modeling, 395–396
 - fitness evaluation, 400
 - investigations, controversies and gaps, 401
 - limitations, 401
 - objective function, 400
 - optimal solution, 401
- Inverse theory, 672

Bayesian inversion, 674–675
 continuous problem, 673–674
 discrete problem, inverse of, 672–673
 joint-inversion, 675
 nonlinear problem, 675
 SVD, 673
 Inverse weighted distance, 865
 Inverse Z-transform, 1300
 Inversion, 321, 342
 method, 1479
 scheme, 400
 Inversion theory, 678–680
 definition, 678
 direct inversion methods, 681
 model-based inversion methods, 681–685
 seismic inversion, 685–686
 SVD gravity inversion, 686–687
 Invertibility, 1347
 Investigate, 624
 Investment decision, 1589
 IRF of order 0 (IRF0), 1347
 IRF of order k (IRF k) theory, 1347
 Iris dataset, 356
 Isarithmic maps, 482
 Islands, 25, 26
 Iso-contours, 1177
 ISODATA clustering algorithm, 447
 Isofactorial models, 554
 Isolines, 433
 Isometric logratio, 144
 coordinates, 137
 representation, 11
 Isometric mapping algorithm (ISOMAP), 991–992
 Isomorphous error, 560
 Isothermal atmosphere, 584
 Isotopic, 547, 551
 isotropic, 548
 process, 421
 Isotropy, 1348
 Iterated conditional modes, 798
 Iteration, 1515
 Iterative algorithm, 1321
 Iterative coupling, 209
 Iterative Fuzzy Edge Detection (IFED) algorithm to blurred remote sensing images, 460, 461
 Iterative geostatistical geophysical inversion, 542
 Iterative geostatistical seismic inversion, 540
 Iterative linear inversion methods, 684–685
 Iterative methods, 337
 Iterative MMMSE (IMMSE), 1079
 Iterative reconstruction techniques, 18
 Iterative weighted least squares (IWLS), 689
 definition, 688
 and robust regression, 689–691
 usages of, 691

J

Jaccard, 1119
 Jackknife method
 bias, 1099, 1100
 regional mineral potential estimation, 1100, 1102
 Jacobian, 84, 343
 matrix, 1033
 Jaynes, Edwin T., 842

J-function, 1059
 Johnson-Mehl-Avrami-Kolmogorov crystal-growth formula, 1503
 Joint approximate diagonalization of eigenmatrices (JADE), 643
 Joint cumulative distribution function, 1113
 Joint-inversion (JI), 675
 Joint probability density function, 843
 Joint probability mass function, 843
 Joint roughness coefficient (JRC), 1232
 Journal, André, 605, 693, 1188, 1277
 Juan Brüggen Prize, 1012
 Jupiter, 40

K

Kaczmarz method, 19
 Kalman filter, 82, 83, 85, 342–343, 345
 dynamic system, 82, 84
 kalman gain, 83
 prediction, 83
 state-space, 82, 83
 unknown inputs, 83
 updating, 83
 white noise, 83
 Kansas Geological Survey (KGS), 251, 857
 Kaolinite, 1468
 Kappa analysis, 303
 Kappa coefficient, 758
 Kardar–Parisi–Zhang (KPZ) equation, 433
 Karst systems, 1565
 Karush–Kuhn–Tucker (KKT) conditions, 176
 Kashmar–Kerman Tectonic Zone (KKTZ), 458
 Keep it simple stupid, 294
 Kendall's coefficient, 632
 Kernel, 320, 1181
 density estimator, 1429
 Kernel functions, 1175, 1521, 1523
 feature space, 1522
 Mercer's theorem, 1522
 re-mapping process, 1522
 Kernel principal component analysis (KPCA) algorithm, 992–993
 Kernel smoothed density function (KSDF), 1058
 k-fold validation, 1168
 Kinematic viscosity, 1466
 King Fahd University of Petroleum and Minerals (KFUPM), 704
 Kirchhoff's laws, 1463
 K-means algorithm, 448, 466
 K-means clustering
 algorithm, 448
 applications, 696–697
 definition, 695
 future scope, 697
 K-means partitioning, 447
 K-medoids, 697–699
 algorithm, 699
 clustering, 699
 K-nearest neighbors (kNN), 639, 662
 applications, 701–702
 definition, 700
 future scope, 702
 Knot theory, 1565
 Knowledge, 1600
 discovery, 1535, 1536
 graph, 1014, 1015
 organization, 1014
 representation, 370, 1015

- Knowledge base, 1601
 - expert systems, 34
 - and graph, 1535
- Knowledge-driven approach, 454
- Koch curve, 996, 997
- Kolmogorov, A.N.
 - book, 703
 - graduation, 703
 - mathematical geoscience, 703
 - mathematics talents, 702
 - publications, 703
 - scaling, 1253
 - sedimentary basins, 703
- Kolmogorov-Smirnov statistic, 1429
- Kolmogorov-Smirnov test, 224, 703, 1441
- Kolumbo volcano, 39
- Korčák's law, 26
- Korvin, Gabor, 703–704
- Kozeny-Carman equation, 1457, 1458, 1468
- Kozeny-Carman relation, 27
- Krige, Daniel Gerhardus
 - education, 704
 - geostatistics, work in, 705
- Kriging, 79, 81, 82, 146, 288, 345, 480, 495, 513, 514, 516, 661, 865, 1181, 1384, 1407
 - application, 712
 - bivariate distributions, 554
 - bivariate Gaussian, 556
 - cokriging, 550–552
 - conditional expectation, 549
 - conditional simulations, 712
 - consistency with data points, 708
 - definition, 705
 - disjunctive, 554, 556
 - DK equations, 554–556
 - estimate, 64, 578
 - global approach, 710
 - history, 705, 706
 - interpolation/estimation, 1610
 - interpolator, 708, 709
 - isofactorial models, 554
 - large data sets, 709–711
 - limitations, 711–712
 - linear, 549, 550
 - linear geostatistics, 706–708
 - local approach and moving neighborhoods, 710
 - MIK, 553
 - non Gaussian models, 712
 - non linear geostatistics and multigaussian, 711, 712
 - non-parametric estimation, 553
 - ordinary, 81
 - predictor, 1365
 - projection, 553
 - properties, 708, 709
 - random variables, 549
 - regionalized variable, 712
 - simple, 82
 - smoothing relationship, 708
 - SPDE, 710, 711
 - standard error, 1365
 - system, 1351
 - universal, 81
 - $Z_v(x_0)$, 552, 553
 - $Z(x_0)$, 552, 553
- Krumbein, William Christian, 714
- Krumbein Medal, 714, 1624
- Krylov subspace methods, 470
- Kullback-Leibler divergence, 347, 767, 1527
 - loss, 778
- Kurtén, Björn, 1219
- Kurtosis, 1313
- L**
 - Labelled dataset, 308
 - Lacunarity index, 432
 - Lagrangian multipliers, 1521
 - Lagrange functional, 843
 - Lagrange multipliers, 81, 1461–1463
 - Lagrange parameter, 1215
 - Lagrangian, 81
 - numerical methods, 603
 - techniques, 525
 - Lagrangian particle dispersion model (LPDM), 1385
 - Lake Barlow-Ojibway, 529
 - Landau's theory, 1469
 - Land cover classification, 177
 - Land cover features extraction, 1023
 - LANDLAB, 323
 - Landscape evolution models (LEMs), 321, 1145–1147
 - Landscape texture, 433
 - Land surveying, 519
 - Land use, 1053
 - Land use land cover (LULC), 300
 - Laplace-Beltrami operator, 717
 - Laplace-de Rham operator, 717
 - Laplace equation, 1465
 - Laplace inverse transform, 715
 - Laplace inversion, 717
 - Laplace operator, 717
 - Laplace's principle of indifference, 842
 - Laplace transform, 212, 715–717, 1027
 - applications, 717
 - change of scale property, 716
 - differentiation and integration, 715
 - history, 716
 - infinite series, 715
 - linearity, 715
 - method, 1042
 - translation property, 715
 - uniform convergence, 715
 - Large Synoptic Survey Telescope (LSST), 237
 - Last Appearance Datum (LAD), 193
 - Last consistent occurrence (LCO), 1153
 - Last occurrence (LO), 1153
 - Latent variables, 798
 - Laterally/diagonally symmetric, 1349
 - Lattice(s), 905, 1463, 1465
 - complete chain, 823
 - of convex sets, 823
 - data, 1355–1356
 - duality by complementation, 825–826
 - duality by inversion, 826
 - extensive and anti-extensive operator, 826
 - idempotent operator, 826
 - of Lipschitz functions, 824
 - of partitions, 824–825
 - percolation, 1464, 1468
 - process, 1366, 1367
 - of $P(E)$ type, 823
 - of regular closed set, 823

- Lattice Boltzmann method (LBM), 1465–1466
- Lattice gas (LG) model, 1458
- Laue group, 217
- Law enforcement bodies, 38
- Law of large numbers in statistics, 787
- Law of permanence of distributions, 556
- Law of proportionate effect, 770
- Law of superposition of strata, 569
- Layer boundaries, 686
- Layered earth model, 681
- Lead (Pb), 1441, 1582
- Leakage, 846
- Learning process, 165
- Least absolute deviation (LAD), 718
 - estimator, 729
 - kriging, 718
 - outlier, 719
 - regression, 718
 - resistant, 718
 - robust, 719
 - simplex method, 718, 719
 - variogram, 718
- Least absolute value (LAV), 720
 - computation of LAV estimate, 721–722
 - corrected errors, LAV regression with, 723
 - definition, 720
 - estimate, 690, 691
 - extensions of, 723–724
 - motivation, 720–721
 - properties of LAV estimator, 722–723
- Least energy dissipation principle, 434
- Least median of squares (LMS), 729–731
 - ADALINE and neural networks, 727, 728
 - cartoon depicting steps, 726
 - definition, 724
 - fitting, 50
 - geophysical applications, 726
 - geophysical data filter, 726, 727
 - gradient descent, 725
 - learning rate, 726
 - linear regression, 725
 - outliers, 726
 - steps, 725
 - transfer function, 727
 - utilization, 726
- Least quantile estimator, 729
- Least squares, 178, 344, 726
 - estimator, 728, 729
 - fitting, 815
 - linear least squares regression, 732–733
 - method, 170, 282, 709
 - non-linear least squares regression, 733–734
 - non-regression least squares methods, 734
 - regression, 1225
 - robust least squares methods, 734–735
 - scaling, 940
- Least squares deviation (LSD), 718
- Least trimmed squares (LTS) estimator, 1227
- Lebesgue measure, 11, 12, 14
- Left-censored data, 637, 638
- Left eigenvector, 337
- Left skewed, 1313
- Legendre polynomial expansion series, 609
- Legendre transformation, 414, 432, 945
- Length scale, 320
- Leningrad, 1624
- Lerchs-Grossmann algorithm, 861
- LERS system, 1600
- Level-1 coefficients, 1307
- Level of significance, 205
- Level set, 1177
- Leverage points, 729, 1435
- Levinson's algorithm, 292, 869
- LibSVM, 1525
- Lidar, 1208
- LiDAR point clouds, 701
- Light detection and ranging (LiDAR), 338, 738, 971
 - advantages, 736
 - airborne LiDAR, 736, 737
 - applications, 736–737
 - definition, 735–736
 - terrestrial LiDAR, 736
 - topographic LiDAR, 736
- Likelihood function, 1476
- Likelihood model, 65
- Likelihood ratio test, 722
- Limitations, 1491
- Limit of detection, 637, 1133
- Limit of quantification, 1133
- Lineament-based fracture characterization, 579–580
- Linear-age calibration levels, 503
- Linear algebra, 322, 336, 613
- Linear approximation, 994
- Linear associations, 203
- Linear cartogram, 105
- Linear coregionalization model (LCM), 978
- Linear discriminant analysis (LDA), 8, 147, 307–309, 311, 982–984, 991
 - Fisher (two-group), 313
 - multiple group, 313–315
 - performance testing and statistical significance, 315–317
- Linear discriminant function, 1520
- Linear dynamical systems (LDS), 643, 1027
- Linear equation, 1203
- Linear functionals on composition, 136
- Linear geostatistics
 - block kriging and area-to-point kriging, 709
 - covariance function/variogram, 706
 - drifts and residuals, 707–708
 - filtering, 709
 - kriging properties, 708, 709
 - OK, 707
 - probability model, 706
 - random function, 706
 - riffs and residuals, 707, 708
 - weights, 706
- Linear interpolation, 663
- Linear inversion methods, 682–684
- Linearity
 - conversion formula, 741
 - definition, 741, 743
 - dependencies, 742
 - example, 742, 743
 - linear regression, 741
 - spatial data, 742
 - visualizing linearity, 742
- Linear kernel function, 1522
- Linear kriging, 549, 550
- Linear least squares regression, 732–733, 1312
- Linear logistic regression analysis, 763
- Linearly independent functionals, 1177
- Linearly separable data, 1520, 1521

- Linear mixture model (LMM), 177
- Linear model of coregionalization (LMC), 547, 871
- Linear models, 782
- Linear operator, 290
- Linear poroelastic theory, 210
- Linear predictors, 1378
- Linear programming, 183
- Linear recurrent relation (LRR), 1330
- Linear regression, 620, 621, 706, 723, 725, 726, 1204, 1205, 1217, 1318, 1433–1436
 - model, 728, 1116, 1225
- Linear scaling, 1247
- Linear signal processing techniques
 - Fourier transform, 1298–1299
 - Hilbert transform, 1300–1302
 - Z-transform, 1299–1300
- Linear state-transition, 342
- Linear surface, 117
- Linear system, 1027, 1175
- Linear time-invariant filter, 1300
- Linear time-invariant (LTI) system, 30, 31
- Linear transformation(s), 337
 - property, 13
- Linear unmixing, 177, 178, 739–740
- Lines of maxima, 1314
- Linguistics and AI, 34
- Linguistic values, 452, 454
- Linked Open Data (LOD), 1019
- Lipschitz functions, 824
- Liquidity index (LI), 574, 575
- LISP machines, 34
- Lithification, 1261
- Lithofacies, 339
- Litho-fluid classification, 69
- Lithogenic elements, 14
- Lithological and geophysical variables, 1576
- Lithology, 430, 434, 1600, 1602
 - classification problem, 1600
 - identification problem, 1604
- Lithosphere, 1045
- L1 loss function, 775
- L2 loss function, 775
- Lloyd-Forgy algorithm, 695
- Loading matrix, 975
- Local change, 585
 - of support, 556
- Local conditional distribution, 555
- Local equilibrium distribution, 1466
- Local linear embedding (LLE) algorithm, 992
- Locally optimal, 1277
- Locally stationary (LS) processes, 51
- Local mapping, 287
- Local minima, 1466
- Local morphometric variables, 930–932
- Local optima, 1528
- Local outliers, 1055
- Local parameters, 1493
- Local singularity, 1315
- Local singularity analysis (LSA), 415, 744–748, 947, 1104, 1105, 1632
 - method, 814
- Local statistical indices, 930
- Log analysis, 318
- Logarithmic, 321
 - averaging procedure, 1313
- Log-binomial distribution, 254, 416
- Log-cosh, 775
- Log-Gaussian prior model, 67
- Logical relations, 1600
- Logistic, 37, 487
 - difference equation, 752
 - loss, 777
 - skew-normal models, 12, 13
- Logistic attractors, 751
 - attractors, 753
 - fracture propagation and growth, 756
 - geological phenomena, 756
 - growth phenomena, 756
 - nonlinear behavior, 756
 - period-doubling, 754
 - period-doubling bifurcations, 753
 - stability, 753
- Logistic map, 107, 112, 751, 754
 - and logistic function, 752
- Logistic regression, 36, 147, 761–762, 777, 1098, 1099, 1167, 1168
 - analysis, 757–759
 - definition, 756, 759
 - evaluation, 758
 - odds, logit transform and logistic function, 761
 - probability estimation, 761
 - stochastic conditional independence, 760
 - stochastic independence, 760
 - and weights of evidence, 760, 764
- LogitBoost algorithm, 777
- Log-likelihood function, 356, 852
- Log-likelihood ratio test
 - asymptotic distribution, 769
 - definition, 766–767
 - multinomial distribution, 767–768
 - normal distribution, 768–769
- Log-log scale, 621
- Lognormal distribution, 26
 - definition, 769
 - estimation of parameters, 771
 - multivariate, 772
 - properties, 770
 - three-parameter, 772
 - truncated, 772
- Log-normal renewal model, 1477
- Logperiodic corrections, 1470
- Log-ratio, 640
 - coordinate representations, 11
 - representation, 5
 - transformation, 872, 1124
- Log transform, 53
- Long memory, 620
- Long-range correlations, 1312, 1314
- Long short-term memory (LSTM), 269, 1536
 - networks, 1274
- Long tail, 1019
- Long-term correlations, 1213
- Long-term mine planning, 861
- Lopatin's time temperature index, 1470
- Lorenz, Edward Norton, 773
- Lorenz equations, 996, 1026
- Loss function, 796, 1435
 - classification, 777–778
 - definition, 774
 - in geoscience, 778
 - regression, 775–777
- Loss ratio, 258
- Lovejoy-Schertzer approach, 443
- Lower approximation, 1600, 1602

Lower dimensional space, 307–309
 Lower quartile, 436
 Low Hurst exponent, 624
 Low-pass filter coefficients, 1308
 L-U decomposition method, 563
 LU simulation method, 99
 Lysenkoism, 1624

M

Machine intelligence, 334
 Machine-interpretability, 1016
 Machine learning, 32, 35–36, 85, 233, 238, 262, 498, 577, 726, 774, 934, 973, 983, 1168, 1385, 1520, 1535, 1536, 1545
 additional problems, 988
 algorithms, 819
 computation time, 988
 datasets, 356
 decision tree, 783
 definition, 781, 987
 in ESP, 323–324
 issues and limitations, 783–784
 linear models, 782
 methods, 1061, 1381
 neural networks, 783
 nonlinear regression, 987
 reinforcement learning, 782
 samples, 988
 semi-supervised learning, 782
 supervised learning, 782
 support vector machine, 782
 unsupervised learning, 782
 Machine-readable FAIR metrics, 89
 Machine translation, 34
 Mach number, 1466
 Ma class of models, 1377
 Macqueen algorithm, 695
 Macroclimate, 585
 Macroscopic directional properties, 219
 Macroscopic distribution, 1462
 Macroscopic fluid density, 1466
 Macroscopic velocity, 1466
 Magmatic nickel-copper deposits, 1098
 Magnetic field, 1030, 1301
 Magnetic permeability, 675
 Magnetotelluric data, 676, 1314
 Magnetotelluric method, 675
 Mahalanobis, 148
 distance, 128, 695, 1120
 Mahalanobis distance-based approach, 1389
 Main shock, 330
 Majority voting, 700
 Mallat, S., 1306
 Mallat's algorithm, 1306, 1308, 1631
 Malthusian exponential growth model for population growth, 752
 Malthusian parameter, 1026
 Mamdani-type FIS, 450
 Mandelbrot, B.B., 784–785
 Mandelbrot's perimeter-area scaling law, 1458
 Mandelbrot's rule, 27
 Manhattan, 1120
 distance, 128, 131
 Manifold learning, 991, 992
 Mann-Whitney's test, 631
 Mantle convection mathematical model, 805
 Manual approach, 1009
 Manual methods, 1495
 Map-making, 647–648
 Mapping and database module, 294
 Mapping contaminants, 1184
 Mapping workflow, 294
 Marching Cubes method, 663
 Marginal cumulative distribution functions, 1517
 Marginal distributions, 1516
 MariaDB, 498
 Mark-correlation functions, 1076
 Marked point pattern (MPP), 1059
 Marked point process, 1074, 1368, 1475, 1502
 Marker-in-cell approach, 399
 Market demand, 1589
 Markov Chain, 788, 961
 applications in geoscience, 793–794
 definition, 791
 geology and geophysics, 793
 hidden Markov models, 792
 hydroclimatology, 793
 hydrology, 793
 Markov Chain Monte Carlo method, 792–793
 process, 788
 remote sensing, 794
 3-state discrete, 791
 Markov Chain Monte Carlo (MCMC), 66, 785–790, 798, 892–893
 algorithms, 68, 788, 789
 method, 1615
 sampling, 794
 technique, 1078
 Markovian decision processes, 37
 Markov mesh models (MMM), 963
 Markov models, 556
 Markov property, 796
 Markov random fields (MRF), 71, 448, 795–797, 799, 962, 1366
 approximation, 710
 factor graphs, 797–798
 solutions, 798–799
 Marquardt parameter, 673
 Mars, 40
 Marsily, G. de
 awards and honours, 800
 biography, 799
 career overview, 799
 scientific contribution, 800
 Mars Sample Return (MSR) Mission, 40
 Martian Moon (Phobos) Sample Return Mission, 40
 Mass exponent function, 432
 Mass-partition function, 414
 Material grouping, 1530
 Material samples, 656
 Mathematical Applications in Geology, 512
 Mathematical geocomplexity theory, 115
 Mathematical geology, 512, 857
 Mathematical geosciences (MG), 1–2, 512–514, 517, 618, 1345
 age of the Earth, 804, 805
 analytical and testing techniques, 803
 big data and artificial intelligence, 814–816
 characteristics, 803
 chemical and physical properties, 803
 definition, 801, 802
 descriptions, 803
 descriptive disciplinary, 802

Mathematical geosciences (MG) (*cont.*)

Earth geometry, 804
 frontiers of, 813
 geochemistry, 809–811
 geometry, 804
 GIS, 809
 history, 801
 IAMG, 801, 802
 interdisciplinary discipline, 803
 mineral crystals, 806–809
 modeling geocomplexity, 813–815
 natural sciences and earth science, 802, 803
 physical detection and survey technologies, 803
 plate tectonics, 805–807
 plate tectonics theory, 802
 probabilistic statistics, 803
 quantitative prediction and assessment of mineral and energy resources, 810–812
 sedimentology, 809
 and text mining, 1535
 Mathematical methods, 864
 Mathematical minerals, 818
 Mathematical modeling, 1264
 Mathematical models, 606
 Mathematical morphology (MM), 513, 515, 587, 594, 820, 835, 901, 906, 907, 929, 1145, 1280, 1514, 1537
 adjunctions, 829–830
 Boolean lattices, 822
 complete lattices, 821–822
 definition, 820
 dilation and erosion, 827–830
 distributivity, 822
 duality principle, 821
 flat structuring elements, 832
 lattice of partitions, 824–825
 lattices of numerical functions, 823–824
 lattices of operators, 825–826
 lattices of sets, 822–823
 monotone convergence and continuity, 822
 Moore families and closings, 826
 Moore families and openings, 827
 operation, 599
 structural theorem of openings, 827
 Mathematical morphology (MM) operators
 on binary images, 93
 greyscale images, 600
 Mathematical statistics, 645, 882
 basaltic lava flows, 513
 geological processes, 512
 mechanisms, 512
 qualitative methods, 512
 quantitative science, 512
 rock types, 512
 scientific methods, 512
 Mathematische Geologie Journal, 1540
 Matheron, G., 835
 Matheron's criterion, 910–911
 Matinenda Formation, 1573
 Matrix, 322, 384, 684, 686
 data, 238
 decomposition, 561, 562
 Matrix algebra
 addition, 837, 840
 column matrix, 837
 definition, 836
 determinant of matrix, 839, 841

 diagonal matrix, 837
 eigenvalues and eigenvectors, 840, 841
 identity matrix, 837
 inverse of matrix, 839
 matrix scalar multiplication, 838
 multiplication, 838, 840
 null or zero matrix, 837
 order of matrix, 836
 power of a matrix, 837
 rank of a matrix, 837, 841
 row matrix, 837
 scalar matrix, 837
 square matrix, 837
 subtraction, 838, 840
 trace of matrix, 839
 transpose of matrix, 839
 Maturation of hydrocarbon, 1470
 MaxEnt probability distribution, 843
 Maximal overlap discrete wavelet transform (MODWT), 379
 Maximum a posteriori (MAP) estimate, 356, 796–798
 Maximum a posteriori probability (MAP), 63
 Maximum entropy, 851
 principle, 349
 Maximum entropy method (MEM), 851, 1459–1460
 applications, 844
 contact force distribution, in granular media, 1462–1463
 definition, 842
 general formulation, 843–844
 notation, 843
 pore shape inversion, 1460–1461
 probability model, 843
 and shale compaction, 1461–1462
 spectral analysis, 844
 Maximum entropy spectral analysis (MESA), 845–848, 850, 851
 advantages, 846
 difficulties, 846–849
 Maximum likelihood, 1580
 classifier, 1056
 method, 12, 516, 691
 Maximum likelihood estimate (MLE), 1436–1438, 1476
 applications, 853
 definition, 851, 852
 overview, 852
 polynomial methods, 853
 Maximum-margin classification, 778
 Maximum noise fraction (MNF), 268
 Maximum partitioning, 913
 Maximum value neighborhood (MVN) heuristic algorithm, 861
 Maxwell's equations, 1030
 McCammon, R.B.
 academic positions, 853
 activities in IAMG, 854
 geology, 853
 IAMG founding member, 853
 publications and technical reports, 854
 research contributions, 854
 McKenzie model, 805, 807
 McKerrow method, 505–506
 Mean, 203, 436, 1054, 1313
 error, 1608
 free path, 1494
 Mean absolute error (MAE), 1204, 1237
 Mean absolute percentage error (MAPE), 775
 Mean-adjusted values, 1215
 Meanders, 24, 25, 27

- Mean percent difference (MPD), 1133
Mean squared error (MSE), 953
Measurement errors, 342
Median, 26, 436
 value, 203
Median absolute deviation (MAD), 367
 estimate, 690
Medium-term mine planning, 861
Medoid, 697
Membership, 449
Membership functions
 characteristics, 854
 definition, 854
 Earth Sciences, 854, 856
 fuzzy sets, 854
 linear form/linearization, 854
 monotonous, 854
 properties, 855
 types, 855
Mercer's theorem, 1522
Mercury intrusion porosimetry, 1085
Mereonymy, 1013
Merging order, 94
Merriam, Daniel F., 1219
Mesoclimate, 585
M-estimators, 1434, 1435
(Meta)data, 231, 232, 369–371, 656–658, 1018
 definition, 858
 geospatial, 858–859
 keywords and controlled vocabularies, 859
 libraries, 858
 management, 247
 semantic web, 859
 standard, 372
 web pages, 858
Meta-heuristic(s), 1320
 algorithms, 1048
 optimization algorithms, 1022
 techniques, 1021
Meta-systemic governance, 535
Meteorological, 1224
Meteorology, 974
Method of characteristics, 321
 for wave equations, 1044
Method of eigenvalue, 1041–1042
Method of integral transforms, 1042–1043
Method of moments, 414
Method of separation of variables, 1040–1041
Metric fractal density, 814
Metric MDS, 939, 943
Metric variance, 14
Metropolis, 68
 algorithm, 786
 Monte Carlo algorithm, 1614
 rule, 1459
Metropolis Hastings, 789, 790
Metropolis-Hastings algorithm (M-H algorithm), 68, 70, 563, 788, 790, 793, 892, 893
Meyer wavelet, 377
Microclimate, 585
Microfauna, 1005
Microprobe modal analysis, 883
Microscopic image, 1458
Microscopic velocities, 1466
Micro X-ray-CT image, 1465
Mid-ocean ridge heat flow diffusion model, 805
Mikowski operations, 827–828
Milankovitch cycle, 526
Milankovitch theory, 647
Military bodies, 38
Miller crystal face index, 807
MIN/MAF autocorrelation factors, 556
Mindat, 236
Mine planning, 860
 definition, 860
 long-term, 861
 medium-term, 861
 short-term, 861
 underground, 861
Mineral crystals, 806–809
Mineral deposits, 38, 208, 381, 383, 650
Mineral detection, 579
Mineral exploration, 380, 449
 activities, 61–63
 decision-making, 1106
 environmental protection strategies, 1105
 stream sediments, 1105
 target selection, 1095
Mineralization, 256, 381, 382, 1103
Mineral mixing, 740
Mineral occurrence database (MINFILE), 148–150
Mineralogy, 39
Mineral potential mapping, 383, 386
Mineral potential maps, 1097–1099, 1102
Mineral prospectivity analysis, 1169
 conceptual modeling, 864
 definition, 862
 evaluation, 866, 867
 evidential map, 864, 865
 geological evidence, 863
 geoscientific data, 863
 predictor map, 865
 regional-scale, 863
 spatial data, 863
Mineral prospectivity maps, 1168
Mineral resource(s), 380
 assessment, 1589
Minerals and organic matter, 1264
Mineral target selection, 380
Minimizing, 1638
Minimum covariance determinant (MCD), 1530
Minimum entropy deconvolution (MED)
 applications, 869–870
 definition, 868
 future scope, 870
Minimum-maximum autocorrelation biplots, 146
Minimum/maximum autocorrelation factors (MAF), 870
 analysis, 144, 145
 application, 872, 874
 biplots, 872
 calculation, 871, 872
 geochemical surveys, 872
Minimum mean square error (MMSE), 1079–1082
 estimator, 64
Minimum noise fraction (MNF) techniques, 460
Minimum screening and acceptance criteria, 1197
Mining, 967
 modelling, 876
Mining Academy Freiberg, 1539
Minkowski distance, 128
Misfit, 681–685
Missing at random (MAR), 638

- Missing completely at random (MCAR), 637
- Missing data, 637, 638, 640, 1169
- Missing not at random (MNAR), 638
- Mitchell-Sulphurets area, 414
- Mixing-models, 114
- mixture rule, 1467
- MM operators on graphs: dilation and erosion, 597–599
- Mobile Net, 266
- Modal analysis, 113, 881, 882
 - Antarctic Meteorite, 882–883
 - binomial distribution, 882
- Modal minerals, 818
- Q*-mode factor analysis (QFA), 1119, 1124
 - choosing and extracting, 1121–1124
 - estimating similarity, 1119–1120
 - factor rotation, 1124–1125
 - linear model, 1120
- R*-mode factor analysis, 1119
 - choosing and extracting, 1221–1224
 - definition, 1219
 - factor rotation, 1224
 - historical background, 1220–1221
- Model(s), 681, 683, 685, 686, 688, 1053
 - assessment, 774
 - based stereology, 1499
 - calibration, 583
 - of de Wijs, 517
 - with edge effect, 555
 - fitting, 1168
 - linearization, 1033
 - parameters, 678, 679, 681, 683–685, 688
 - selection, 774
 - space, 685
 - uncertainty, 1180
 - updating, 1543
 - without edge effect, 555
- Model-based algorithms, 968
- Model-based clustering, 127
- Model-based imputation, 638
- Model-based inversion methods, 681–685
- Modeling of geological systems, 63
- Mode-mixing, 1316
- Moderate resolution imaging spectroradiometer (MODIS), 1546
- Modern biostratigraphy, 1152
- MODFLOW, 391
- Modifiable areal unit problem, 1356
- Modified Akima approach, 1406
- Modified FCLS (MFCLS), 178
- Modified Hay method, 1158
- Modified Omori formula, 1474
- Modified periodograms, 306
- Modified power laws, 26
- Modulation, 1295
 - ratio, 302
- Molecular dynamics model, 1465
- Moment(s), 26
 - and cumulants, 606
 - generating function, 771
- Monofractal, 412, 624, 945
 - behavior, 1310, 1312
 - distribution, 745
- Monomials, 1347
- Monostatic radars, 1207
- Monotone limit, 822
- Monotonic signal, 1316
- Monotonic transformation, 345
- Monte-Carlo, 323
 - inference, 892
 - method, 65, 67, 685, 786, 890, 1190, 1192
 - optimization, 894
 - testing, 891
- Monte Carlo optimization-sampling hybrids, 895
- Monte Carlo sampling, 786
 - drawbacks, 787
 - probability density function, 786
 - random numbers, 786
- Monte Carlo simulation (MCSIM), 712, 815, 1584–1587
 - in optical imaging methods, 786
- MOOC (massive open online course), 165
- Moore families, 826
- Moran index
 - area objects, 898
 - interval data, 898, 899
 - nominal data, 900
 - object types, 899
 - ordinal data, 899, 900
 - spatial auto correlation, 897
- Morlet's method, 1626
- Morphologic age, 320
- Morphological closing, 902
 - definition, 901
 - extensive operator, 902
 - illustrations, 901
 - increasing operator, 902
 - properties, 901
- Morphological dating, 320
- Morphological dilation, 906, 908, 909
 - binary images, 903
 - definition, 902
 - grey-scale images, 904–905
 - properties, 905
- Morphological erosion, 907
 - binary images, 907
 - definition, 906
 - grey-scale images, 907, 908
 - lattice-based definition, 909
 - properties, 908–909
- Morphological filtering, 910
 - alternating sequential filters, 911–912
 - compound segmentation, 916–917
 - connected filters, 914
 - connective segmentation, 914–916
 - dynamic programming, 918–919
 - energetic lattices, 919–920
 - families of minimal cuts, 919–920
 - granulometries, 911
 - hierarchies and semi-groups, 911
 - hierarchies of partitions, 917–918
 - h*-increasingness and partition closing, 919
 - opening and closing, 596
 - optimal cuts, 918–919
 - over and underfilters, iterations of, 911
 - products of filters, 910–911
- Morphological interpolations, 1145
- Morphological opening, 909, 921–923
 - algebraic openings, 922–923
 - definition, 921
 - filtering, 923
 - properties, 922
- Morphological operations, 322
 - on an image, 595
- Morphological shape decompositions, 1145

- Morphological techniques, 599
- Morphology, 928
- Morphometric analysis, 1141, 1224
- Morphometric features, 928
- Morphometric variables, 930
- Morphometry, 929, 930
 - combined morphometric variables, 933
 - definition, 928
 - local morphometric variables, 930–932
 - nonlocal morphometric variables, 932–933
 - two-field specific morphometric variables, 933
- Mosaic cartograms, 104
- Mosaic model, 555
- Mother wavelet, 1304, 1308, 1628
- Motion estimation, 36
- Motion of a particle, 1049
- Mount Albert intrusion, 1569, 1571
- Moving average, 936, 937
 - definition, 934
 - unweighted moving average, 935
 - weighted moving average methods, 935–937
- Moving paper-slip technique, 290
- Mud volcano, 1469
- Multifractional Brownian motion, 620–621
- Multichannel blind deconvolution (MBD) method, 869
- Multichannel filters, 290
- Multi-channel SSA (M-SSA), 1332
- Multiclass problem, 777
- Multiclass strategies, 1522, 1523
- Multicollocated, 548
- Multidimensional scaling (MDS), 144, 145, 744, 982
 - aims, 938
 - application examples, 942–943
 - classical scaling, 939–940
 - definition, 938
 - dissimilarity measure, 939
 - goodness-of-fit and number of dimensions, 941–942, 944
 - map, 938
 - metric, 939, 943
 - non-metric MDS, 940–941
 - types of, 939
- Multifractal behavior, 1310, 1312, 1314
- Multifractal detrending moving average analysis (MFDMA), 254
- Multifractal DFA (MFDFA), 1311
 - applications, 1314
 - forward and backward directions, 1313
 - Hurst exponent, 1313
 - linear least-squares regression, 1313
 - modified least-squares sense, 1312
 - scaling exponents, 1311
 - singularity spectrum, 1313
- Multifractal inverse distance weighted model (MIDW), 946
- Multifractalism, 433
- Multifractal(s), 108, 322, 432, 516, 517, 785, 810, 1576, 1599
 - asymmetry index, 946
 - fractal dimensions, 945
 - measures, 945
 - method of moments, 945
 - model, 814
 - monofractal, 945
 - multifractal spectrum, 945
 - self-similarity, 945
 - singularity exponent, 945
 - singularity spectrum, 1313–1315
 - spatial correlation, 417–419
 - spectrum, 412–414, 432
 - studies, 1314
 - theory, 814
- Multifractional Brownian motion, 1576
- MultiGaussian, 1001, 1276, 1278, 1279
 - approach, 558, 1325
 - kriging, 711, 712
 - random field, 605
 - random function model, 1187
- Multi-group discriminant analysis, 35
- Multilayer networks, 1266
- Multilayer perceptron (MLP), 953
 - elements, 952–953
 - learning and parameter optimization, 953
 - methods, 1063
 - structure and parameters, 952
- Multimedia, 168
- Multinomial distribution, 767–768
- Multi-objective evolutionary method, 1024
- Multi-objective land allocation (MOLA), 1321
- Multiphase flow, 954
 - analytical formulations vs. numerical simulations, 957
 - capillary pressure, 955
 - contact angle, 955
 - drainage, 956
 - imbibition, 956
 - interfacial tension, 955
 - investigations and knowledge gaps, 957
 - mathematical formulation, 954–955
 - relative permeability, 956
 - residual saturation, 956
 - wettability, 955
- Multiple change detection, 629
- Multiple channel network, 1139
- Multiple correlation coefficient, 376, 958
 - closed composition, 958
 - coefficient of determination, 959
 - cross validation, 959
 - goodness of fit, 959
 - multicollinearity, 958
 - principal components, 958
 - variance inflation factor, 958
- Multiple discriminant analysis, 340
- Multiple group linear discriminant analysis, 313–315
- Multiple imputation, 638
- Multiple indicator kriging (MIK), 553
- Multiple-point geostatistics (MPS), 968, 1543
 - algorithms, 961
- Multiple-point pattern (or moment) in MPS, 605
- Multiple point periodicity, 605
- Multiple-point simulation methods, 1325
- Multiple-point statistics (MPS), 605
 - techniques, 1352
- Multiple regression analysis, 1220
- Multiple representations, 1359
- Multiple-source data, 1541
- Multiplicative cascade models, 26
- Multipoint geostatistics, 1188
- Multipoint optimal MED adjusted (MOMEDA) method, 870
- Multi-point simulation, 564, 565
- Multi-point statistics, 559, 565, 566
- Multiquadric RBFs, 1175
- Multi-resolution analysis, 1307, 1308, 1627, 1629–1631
- Multiresolution space-frequency tiling of wavelet analysis, 1628

- Multiscale, 929
 - morphological operations, 1145
 - openings and closings in the modeling of the geomorphologic processes, 1145
- Multiscale covariance maps (MCMs), 268
- Multiscaling
 - aleatoric uncertainty, 971
 - applications, 971, 972
 - clouds, LiDAR, 972, 973
 - definition, 970
 - methods, 972
 - scale-independent entities, 971
 - structures, 971
- Multispectral images, 579
- MultiSpectral Instrument (MSI), 627
- Multispectral (MS) sensors, 627
- Multistage random sampling, 1240
- Multitaper method (MTM), 306
- Multi-user friendly database, 233
- Multivariate analysis, 1117, 1194
 - CA, 977
 - CCA, 975, 976
 - CorA, 976, 977
 - DA, 977
 - definition, 974
 - geosciences, 974
 - geostatistical tools, 977–980
 - non-exhaustive list, 974
 - PCA, 975
- Multivariate analysis of variance (MANOVA), 8
- Multivariate data, 637, 1059–1060
 - analysis, 35, 145, 1119
 - imputation, 638–639
- Multivariate data analysis, in geosciences
 - automating geoscience analysis, 982–983
 - clustering, 981–982
 - CoDa, 983
 - dimensionality reduction, 981
 - factor analysis, 982
 - graphical procedures, 981
 - LDA, 982
 - QDA, 982
 - tree-based modeling, 982
- Multivariate distributions, 1185
- Multivariate geostatistics, 145
- Multivariate interpolation, 660–662, 666
- Multivariate normal distribution, 772
- Multivariate random function (MRF), 978
- Multivariate relationships, 368
- Multivariate skew-normal distribution, 11, 12
- Multivariate space, 127
- Multivariate space-time structure, 974
- Multivariate spatial processes, 1368–1370
- Multivariate spatial structure, 974
- Multivariate spatio-temporal analysis, 1380
- Multivariate statistical methods, 143
- Multivariate statistical techniques, 974, 975
- Multivariate techniques, 381, 638
- Multivariate time series, 1557–1558
- muPETROL, 38
- Muscovite, 1464
- Mutation, 465
- Mutual information, 347
- MySQL, 498
- N**
 - Nanotechnology, 37
 - Narrow AI, 32
 - NASA, 87
 - NASA-PSTAR, 40
 - NASA World Wind, 1622
 - National Bureau of Standards (NBS), 114
 - National Oil Company of Chile (ENAP), 1012
 - National Science Foundation (NSF), 656
 - Natural disasters, 1261
 - Natural genetics, 468
 - Natural hazards, 929
 - and earthquakes, 335
 - Natural language, 1536
 - Natural language processing (NLP), 34–35, 233, 791, 1535
 - Natural processes, 274, 966
 - Natural resources, 1261
 - Nature-based algorithms, 1022
 - Nature-inspired algorithms, 1022
 - Navier-Stokes equation, 389, 1030, 1139
 - Nearest neighbor, 448
 - approach, 1242
 - Negative binomial distribution, 416
 - Negative matrix factorization (NMF), 308
 - Negative phasor, 1301
 - Negative skewness, 437
 - Neighbor geometry, 668
 - Neighborhood, 1266
 - algorithm, 895
 - policy, 666
 - size, 667
 - Neighbouring, 621
 - Němec, Václav, 985
 - Neptunism, 649, 1096
 - Nereid Under Ice robot, 39
 - Nested pit method, 861
 - Nested structures, 1350
 - Network analysis, 1344
 - Network-based data, 239
 - Network cartograms, 105, 106
 - Network topology, 951
 - Neural network, 145, 147, 340, 727, 728, 783, 1046, 1061
 - additional problems, 990
 - applications, 989
 - cases, machine learning, 987
 - definition, 986, 990
 - equations, 986, 987
 - limitations, 987
 - neurons, 989
 - science/engineering, 986
 - tools, 989
 - Neuro-fuzzy framework, 453
 - Neutron, 818
 - capture spectroscopy, 819
 - Newton's second law, 525
 - Newton's cooling problem, 1025
 - Node, 1463, 1468
 - Noise, 679, 682, 688, 1288
 - Nominal random variable, 562
 - Non-causal signals, 1297
 - Non-conditional simulation, 560
 - Non-contiguous cartograms, 104, 105
 - Nondetects, 637, 638, 640–642
 - Non-differentiable curves, 1308

- Non-flat structuring, 600
 - Non-flat structuring element, 904
 - Non-Gaussian, 345, 1277
 - distribution, 432
 - Nonhomogeneous hidden Markov model, 792, 793
 - Nonlinear, 319, 683
 - advection, 433
 - equation, 1026, 1203
 - filters, 290, 1470
 - geostatistics, 711, 712
 - least squares regression, 733–734
 - mono vs. multifractal signals, 1312
 - multivariable maximum likelihood estimation, 1436–1438
 - multivariable minimization, 1438
 - physical problems, 434
 - problem, 675
 - processes, 145
 - signal analysis techniques, 1298
 - signals, 1303, 1310
 - state-transition, 342
 - test signal, 1317
 - time series model, 1556
 - Nonlinearity, 994, 997, 998
 - chaos, 995–996
 - fractal, 995–997
 - Non-linearly separable data, 1521
 - Nonlinear mapping algorithm, 993
 - ISOMAP, 992
 - KPCA, 992–993
 - LLE algorithm, 992
 - manifold learning and basic algorithms, 991
 - Nonlinear signal processing
 - data adaptive techniques, 1303
 - fractal and multifractal analyses, 1308–1317
 - mathematical transformations, 1303
 - wavelet analysis, 1303–1308
 - Nonlocal effects, 320
 - Nonlocal morphometric variables, 932–933
 - Non-metric MDS, 940–941
 - Non-negative Least Squares (NNLS) algorithm, 178
 - Non-parametric estimation, 553
 - Non-regression least squares methods, 734
 - Non-relational databases, 498
 - Non-renewable natural resources, 1589
 - Non-slip, 1466
 - Non-stationary random functions, 545, 546
 - Non-stationary signals, 376, 1297, 1310
 - Nonstationary source separation (NSS), 1558
 - Non-supervised classification, 1005
 - Non-unique, 679
 - Norm, 615, 681, 682, 686, 1348
 - Normal distributions, 445, 768–769, 1440
 - approximate normality, 1000
 - central limit theorem, 999
 - determine exact normality, 1000
 - natural, 999
 - related distributions, 1001
 - social sciences, 999
 - standard, 999
 - statistical parametric mapping, 1001
 - Normal equations, 1278
 - Normal/Gaussian distribution, 1190
 - Normalization constant, 62
 - Normalization constraint, 843
 - Normalized power spectrum, 440
 - Normal matrix, 1033
 - Normal random variables, 47
 - Normal-score(s), 1279
 - transform, 345, 556
 - Normal-score EnKF (NS-EnKF), 345
 - Normed space, 614
 - North American plate, 332
 - Norwegian Sea Foraminifera (NCF), 1162
 - Not-a-Knot, 1404
 - Not hierarchical clustering, 977
 - Nowcasting, 334, 336
 - Nuclear logs, 1085
 - Nuclear magnetic relaxation (NMR), 1464
 - Nugget anisotropy, 1348
 - Nugget effect, 1350, 1611
 - Null and alternative hypotheses, 205
 - Null hypothesis, 206, 1439, 1440
 - Null or zero matrix, 837
 - Number of objects, 1494
 - Number of points, 1494
 - Number-size multifractal model (N-S), 946
 - Numerical methods, 209, 320, 323, 1045–1047
 - Numerical model, 580
 - Numerical simulations, 580
 - Numerical solution, 1027, 1028
 - Numerical stratigraphic inverse modelling, 396
 - n-variate distributions, 1185
 - Nyquist frequency, 844
 - Nyquist's criteria, 1308
 - Nyqvist theroerm, 660
- O**
- Oak Ridge Moraine (ORM), 653
 - Object based image analysis (OBIA), 1009–1011
 - boundary based approach, 1010
 - definition, 1008
 - image segmentation, 1010
 - input image, 1010
 - non-parametric classification, 1010
 - output classified image, 1010
 - parametric classification, 1010
 - processing steps, 1010
 - region based approach, 1010
 - remote sensing, 1011
 - spatial and spectral characteristics of pixels, 1010
 - Object-based models, 566
 - Object-based simulation, 566
 - Object detection, 267
 - Objective/error function, 395
 - Objective function, 1438
 - Object recognition, 36
 - Oblique axis rotation, 1124
 - Observation matrix, 342
 - Observed data, 682–685
 - Ocean effect, 1314
 - Oceanography, 38, 1224, 1352
 - Off-surface orientation, 1177
 - Oil and gas reservoirs exploration, 1023, 1024
 - Oil volume random variable, 1191
 - Olea, Ricardo A.
 - educational and career achievements, 1012
 - personality, 1012
 - role in IAMG, 1012

- Omori's Law, 331
- Omori-Utsu formula, 1474
- One-dimensional forward modeling, in direct current (DC) resistivity, 1052
- One-hot encoding, 777
- One-versus-one (OVO), 1523
- One-versus-rest (OVR), 1523
- One-way coupling, 209, 212
- Online/incremental algorithm, 695
- Ontological engineering, 1013
- Ontology, 371
 - definition, 1013
 - origin, 1013
 - types of, 1014–1017
- Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), 657
- Open data
 - definition, 1017
 - in geosciences, 1019–1020
 - LOD, 1019
 - pre-requisites of, 1017–1018
 - principles, 1018
 - State of Open Data, 1019
- Open fracture, 1232
- Open license, 1017
- OpenNebula, 126
- Open pit mine dewatering, 490
- Open-pit mine hydrogeology, 492
- Open-pit mining operations, 490
- Open source software (OSS), 1622
- OpenStack, 126
- OpenStreetMap, 499
- Operationalism, 74
- Optimal estimate, 343
- Optimal filtering, 292
- Optimal unbiased linear combination, 1215
- Optimization, 386, 679, 681, 688, 778, 1021, 1181, 1466
 - algorithms, 1022
 - based methods, 964
 - functions, 774
 - methods, 1320, 1321
- Orbital constraints, 280
- Orbital parameters, 282–283
- Orbital system, 281
- Orbiting Carbon Observatory-2 (OCO-2), 1382
- Order, 1054
 - of matrix, 836
 - parameter, 1469, 1598
 - statistics, 665, 1429
- Ordered objects, 1003
- Ordinal data, 900
- Ordinary cokriging of indicators, 552
- Ordinary differential equation, 275, 1040
 - carbon cycle, 1031
 - crustal geotherm, 1030
 - definition, 1024
 - earth's evolution and processes, 1025
 - energy balance, 1029
 - evolution of mantle temperature, 1029
 - first order ODEs, 1025
 - geochemical reactions, 1031
 - geochronology, 1028, 1029
 - geophysical systems, 1031
 - geosciences problems, 1025
 - Green's function, 1027
 - groundwater mound, 1031
 - hill-slope, 1029, 1030
 - initial value problem, 1024
 - linear systems, 1027
 - magnetic field, 1030
 - nonlinear equations, 1026
 - numerical solution, 1027, 1028
 - orogens, 1030
 - qualitative methods, 1028
 - second order ODEs, 1025, 1026
 - special functions, 1028
 - temporal evolution, 1025
 - transform methods, 1027
 - two-point boundary value problem, 1024, 1026
- Ordinary indicator kriging, 553
- Ordinary kriging (OK), 549, 707, 1365, 1378, 1608
- Ordinary least squares (OLS), 177, 485, 487, 1032
 - empirical considerations, 1032
 - fundamental problem, 1032
 - Gauss-Markov conditions, 1034
 - geometrical interpretation, 1034
 - maximum-likelihood estimation, 1036
 - model-based prediction, 1033, 1034
 - model linearization, 1033
 - model parameters, 1033
 - multiple determination, 1036, 1037
 - nonlinear system, 1032
 - regression, 720, 723, 1199–1202
 - residual analysis, 1037
 - residuals, 1034
 - residual sum of squares, 1036
 - statistical optimality, 1035
- Ordination spaces, 1122
- Ordovician-Silurian (OS), 508
- Orebody, 61
- Ore generation models, 1166
- Ore grade estimation, 1177
- Ore targets, 1166
- Organic chemistry, 34
- Organic maturation, 1263
- Orientation, 1514
 - density function, 217
 - parameters, 279
 - statistics, 219
- Ormsby filter, 292
- Ormsby-type bandpass filters, 291
- Orr-Sommerfeld equation, 1576
- Orthogonal arrays, 1194
- Orthogonal indicator residual model, 555
- Orthogonality, in Hilbert space, 616
- Orthogonal matrix, 337
- Orthogonal polynomials, 556
- Orthogonal wavelet, 1307
- Orthometric heights, 1051
- Orthonormal coordinates, 136
- Orthonormal logratio (olr) representation, 11, 12
- Orthopyroxene, 1569
- Osmotic computing, 536
- Ostracods, 1219
- Otsuka, 1119
- Outlier, 208, 1055
 - removal, 1530
- Outlier analysis (OA), 240
- Out-of-bag (OOB) sample, 1182
- Output layer, 1269
- Overall accuracy (OA), 1584–1586
- Over-determined system, 20, 818
 - of equations, 1460
- Overestimation, 776

Over-fitting, 783, 1406
 Overlay analysis, of GIS, 1344
 Ozone monitoring instrument (OMI), 1387

P

Pacific Plate, 330, 333
 Packing, 1461
 Pair correlation function, 1074
 Pairwise distance functions, 700
 Palaeoecology, 1219
 Palaeontology, 1219
 Paleobiology Database, 38
 PaleoDeepDive project, 38
 Paleomagnetism, 388, 1486, 1625
 Paleontology, 39
 “Paleoseismic” record, 335
 Paleoseismology, 333
 Paleostreams, 27
 Paleozoic quartz phyllites, 1487
 Paleozoic sandstone, 1083
 PA-MAE (PARallel k-Medoids clustering with high Accuracy and Efficiency), 699
 Panchromatic (PAN) and multispectral (MS) resolution, 230
 Parallel-beam image reconstruction
 filtered backprojection algorithm, 58
 Fourier slice theorem, 57
 radon transform, 56–57
 Parallel tempering, 895
 Parameter resolution matrix, 673
 Parametric errors, 866
 Parametric splines, 1406
 Parametrization, 1524, 1525
 Pareto distribution, 26, 407
 Pareto frequency distributions, 425
 Pareto-lognormal distribution, 424, 425
 Pareto-lognormal probability density function, 425
 Partial correlation, 205
 coefficient, 958
 Partial dependence plots, 1183
 Partial differential equation (PDE), 209, 275, 276, 319, 469, 1025, 1031
 definition, 1039
 general form, 1039
 Partially heterotopic, 547
 Partially observable Markovian decision process (POMDP) approach, 37
 Partially Ordered Markov Model (POMM), 1370
 Partially ordered set, 821, 905, 909
 Partially separable, 1348
 Partial overall accuracy (POA), 1584, 1587
 Particle angularity indices, 1283
 Particle size classifications, 574
 Particle size distribution (PSD) Analysis, 878, 879
 Particle swarm optimization (PSO), 701, 1048
 applications in geosciences, 1050–1052
 basic algorithm of, 1049
 definition, 1047
 for GNSS network design, 1050
 inversion of residual gravity anomalies, 1051
 local geometric geoid model, 1051
 optimum well type and location, 1051
 variants of, 1050
 Partitional clustering, 977
 Partition function, 432, 843, 1315
 Partitioning around Medoids (PAM), 697, 698
D-part random composition, 13
C-part random subcomposition, 13
 Parzen window, 291
 Passband, 1637
 Passive data dictionary, 232
 Passive dewatering, 490
 Passive sensors, 1207
 Patch-based multi-view stereo (PMVS), 877, 878
 Patch based sequential methods, 963
 Pattern, 756
 formation, 334, 335
 in geoscience, 1055–1056
 types of, 1054–1055
 Pattern analysis, 1057
 in geosciences, 1057–1058
 in multivariate data, 1059–1060
 in univariate data, 1058–1059
 Pattern classification
 definition, 1061
 supervised learning, 1061
 Pattern recognition, 145, 166
 applications in geosciences, 1065
 definition, 1063
 geometric/structural approach, 1064
 learning paradigms, 1064
 neural network approach, 1064
 statistical approach, 1064
 template matching, 1064
 workflow, 1064
 Pawlak, Zdzislaw, 1600
 Pawlak rough set model, 1600
 Pawlowsky-Glahn, Vera
 career, 1066
 compositional modeling, 1066
 education, 1066
 PCI, 226
 P class, 1021
 PC raster system, 101
 Peak Signal-to-Noise Ratio (PSNR), 461
 Pearce element ratio (PER) diagrams, 361
 Pearson, Karl, 1119
 Pearson coefficient, 204
 Pearson correlation coefficient, 129, 208, 387, 807, 1518
 Pedotransfer functions (PTF), 971
 Penalised complexity (PC), 1372
 Pengda, Zhao, 1067
 Percentages, 1119
 Percentiles, 436, 1055
 Percolation, 1598, 1599
 model, 1458
 probabilities, 1468
 theory, 972, 1231, 1463, 1467–1468
 threshold, 1463, 1467, 1468
 Perdurant, 1015
 Periodic boundary conditions, 1465
 Periodic component, 1349
 Periodicity, 430
 Periodogram, 1217
 estimator, 305
 method, 1093
 regression, 1217
 Permeability, 27, 1089–1091, 1232, 1457, 1458
 Permian carbonate reservoirs, 818
 Permutation hypothesis tests, 1430
 Permutation importance, 1183
 Permutation invariance, 135
 Permutation tests, 849, 1431, 1432
 Perplexity, 1528
 Persistence, 432, 1213, 1215, 1217

- Persistent, 1311
 - identifiers, 656
 - random walk, 620
- Persistent unique identifiers (PID), 657
- Perturbation, 10, 13
- Petrographic modal analysis, 882, 883
- Petrography, 882
- Petroleum, 1512
 - geology, 61
- Petrological statistician, 1575
- Petrology, 39
- Petrophysical inversion problems, 69
- Petrophysical properties, 65
- Petrophysics, 318
- Pexider's functional equation, 1470
- Phanerozoic, 502, 1125
- Phase spectrum, 1298
- Phase transitions, 1467–1470, 1597
 - universality in, 1598
- Phenomenalism, 74
- Phi scale, 714
- Phonemes, 35
- Photogrammetric adjustment
 - aerial stereo photo blocks, 284
 - attitude behavior, reconstruction of, 282
 - corrections, 283
 - 3D point coordinates computation, 283–284
 - orbital parameters, 282–283
- Photogrammetry, 36, 300, 519, 929
- Physical annealing, 1320
 - process, 1321
- Physical Bayesian conditionalization rule, 74
- Physical properties, 678, 679, 686, 687
- Physical samples, 656
- Physical weathering, 1136
- Physicochemical properties, 1545
- Physics-based forward modelling, 470
- Piecewise linear interpolation, 661
- Pitchfork bifurcation, 1028
- Pivot coordinates, 144
- Pixel, 758
 - based sequential methods, 963
 - and patch based unilateral methods, 964
 - vectors/patterns, 1203
- Pixel based image analysis (PBIA), 1009, 1011
- Pixel-based methods, 967
- Planar orientation, 1177
- Planck's law, 577
- Planetary Science and Technology from Analog Research (PSTAR)
 - program, 40
- Plasma turbulence, 1216
- "Plastic" and other dissipative "non-Newtonian" systems, 334
- Plasticity index (PI), 574
- Plate kinematic models, 351
- Plate motion calculator, 351
- Plate tectonics, 330, 645
 - theory, 805–807
- Platform as a Service (PaaS), 124, 498
- Plurality, 371
- Plurigaussian, 564
- Plurigaussian simulations, 562, 563
 - applicability of, 1070
 - concepts, 1068
 - conventional, 1070
 - cosimulation, 1072
 - definition, 1067
 - development of, 1070
 - Gaussian random field(s), 1068
 - generalization of, 1070
 - importance for modeling, 1067
 - investigations of, 1070
 - joint simulation, 1072
 - multi-Gaussian simulation, 1070
 - rationale, 1068
 - standard method of, 1070
 - theoretical background of, 1068
 - truncated and, 1069
 - truncation thresholds, 1070
 - types of contacts, 1068
- Point-block models, 557
- Point cloud analysis, 737–738
- Point-counting, 113
- Point estimate, 776
- Point fraction, 1494
- Point patterns, 419, 421, 423, 514, 517, 1073, 1356–1358
- Pointplot, 1257
- Point process(es), 1073, 1074
 - statistics, 1073
- Point-support, 556
- Poisson, 487
 - distribution, 52, 514, 602, 883
 - model, 1475, 1480
 - point process, 1367, 1502
 - probability function, 1409
 - process, 1059, 1076
- Poisson-Boolean model, 1502
- Polarimetric covariance matrices (PCM), 338
- Polarimetric SAR (Pol-SAR), 338, 1079–1082
- Pole density function, 217
- Pole figures, 216
- Polynomial, 1175
 - factors, 555
 - fitting, 288
 - kernel function, 993
 - regression, 1204
 - trends, 1314
- Population, 202, 203, 437, 1054
 - correlation, 205
- Pore(s), 589, 1460, 1468
 - bodies, 1459
 - channels, 1459
 - network, 1459
 - pressure, 525
 - space, 1466
 - spectra, 1461
- Pore-shape distribution, 1460
- Pore-size, 25
 - distribution, 1085
- Pore structure
 - definition, 1083
 - fluid flow path, 1083
 - porosity vs. physical properties, 1085
 - quantitative evaluation, 1084, 1085
 - rock types, 1083
 - visual observation, 1083, 1084
- Pore-surface, 27

- Pore-throats, 27
- Porosity, 25, 27, 574, 818, 1090, 1091, 1216, 1261, 1263, 1458, 1461, 1464, 1468
 - definition, 1086
 - determination, 1090
 - engineering
 - classification, 1089
 - parameters, 1089
 - permeability, 1089, 1090
 - geological classification, 1088, 1089
 - types of, 1086, 1087
- Porous media
 - climate change, 392
 - flow and biogeochemical reactions, 392
 - flow model properties and parameters, scale dependency of, 391
 - principles, 388–390
 - quantifying flow, 389–391
 - uncertainty analysis, 392
- Porous medium
 - characteristics, 1091
 - definition, 1090
 - geological porous media, 1091
 - heterogeneity, 1091, 1092
- Porous rock, 1457
- Portable equipment, 294
- Pose estimation, 36
- Positive definite, 1175, 1347
- Positive measure, 1347
- Positive phasor, 1301
- Positive skewness, 437, 771
- Posterior distribution, 66
- Posterior PDF formulations, 75
- Posterior probability, 146, 796
 - distribution, 66, 796
 - of existence, 62
- Post-hoc steps, 106
- Post-processing, 923
- Potential energy, 1461
- Potential evapotranspiration, 258, 259
- Potential information and knowledge, 1343
- Power function, 432, 1463, 1470
- Powering, 10
- Power law, 24, 433, 1230
 - behavior, 1298, 1311
 - distribution, 26, 404–406
 - function, 321–322
 - relation, 431
- Power model, 1349
- Power of a matrix, 837
- Power spectral density (PSD), 1397
 - application, 1093–1094
 - definition, 1092
- Power spectrum, 372, 376, 432, 845, 847, 848, 850, 1216, 1298
 - estimates, 846
- Power spectrum-volume (P-V) fractal model, 1401–1402
- P-p plot, 1429, 1430
- Preamplifiers, 53
- Precambrian stratigraphy, 502
- Precipitation, 1056
 - seasonality, 259, 260
- Precision, 1–4, 758, 1127–1128
- Precision-recall curves (PR), 764
- Predator-prey equations, 1026
- Predictable, 1469
- Prediction, 774, 1215, 1270, 1347, 1378
 - of stationary time series, 292
- Prediction-rate plot, 866, 867
- Predictive analysis, 236
- Predictive deconvolution, 292
- Predictive geologic mapping
 - bias and Jackknife method, 1099, 1100
 - definition, 1095
 - discrepancies, 1095
 - Gejiu mineral district and Northwestern Zhejiang Province, 1105, 1106
 - geophysical and geochemical methods, 1095
 - igneous and metamorphic rocks, 1095
 - Jackknife applications, 1100, 1102
 - local singularity analysis, 1104, 1105
 - mineral potential maps, 1097–1099, 1102
 - plate tectonics, 1095
 - sound geoscientific theory, 1095
 - usage, 1097
 - weights-of-evidence modeling, 1103
- Predictive mapping, 1184
- Predictive models, 1167
- Predictor(s), 1433
 - map, 865
- Preferential orientation, 1231
- Preservation of mass, 344
- Presorting, 1158
- Pressey's multiple-choice machine, 165
- Pressure, 1460
- Pretrained model, 265, 266
- Principal component, 309, 310, 385
 - biplots, 146
 - method, 384
- Principal component analysis (PCA), 35, 144, 145, 268, 308, 338, 340, 379, 641–643, 870, 939, 974, 975, 981, 982, 991–993, 1060, 1119, 1220–1224, 1228, 1334
 - definition, 1108
 - dimensionality reduction, 1110
 - example of, 1111
 - extensions of, 1110
 - geometrical interpretation, 1109–1110
 - image processing techniques, 458
 - as optimization problem, 1109
- Principal coordinate loadings, 1122
- Principal coordinates analysis (PCoA), 940, 1119
- Principal stress, 525
- Principle component analysis (PCA), 953
- Principle of analogy, 383
- Principle of maximum entropy, 842, 843
- Principle of working on coordinates, 10
- Principles of Geology, 388, 512
- Principles of Mathematical Geology, 1624
- Prior distribution, 798
- Prior model, 65, 67
- Prior probability, 796
 - of existence, 62
- PRISM, 499
- PRISMA mission, 626
- Probabilistic assessments, 1516
- Probabilistic distribution, 795
- Probabilistic finance, 1516
- Probabilistic inference, 785

- Probabilistic planning, 37
- Probability(ies), 432, 757, 934, 1606
 - field simulation, 565
 - function, 796
 - gain, 1480
 - mass function, 223, 346, 1189
 - measure, 796
 - of mineralization, 1166
 - space, 843
 - and statistics, 202
- Probability density, 1440
 - definition, 1112
 - distribution, 65, 223, 435, 633, 778, 1230, 1278, 1310, 1527
 - function, 205, 223, 608, 674, 688, 1113, 1190, 1460, 1462, 1467, 1475
 - geoscience problems, 1115
 - measurement, 1112, 1113
 - random variables and vectors, 1113, 1114
 - statistical and computational essays, 1114, 1115
- Probable errors, 852
- Problem solving, 166
- Process-based models, 967
- Process modelling, 1624
- Process validation, 144, 147, 148, 152–155, 159, 160
- Procrustes distance, 1287
- Procrustes shape coordinates, 1286
- Profile sections, 1004
- Programmable Universal Manipulation Arm (PUMA) system, 37
- Programmed gain control (PGC), 49, 51
- Programming languages, 226
- Projection, 549, 553, 1490, 1491
 - in Hilbert space, 616
- Prolate spheroidal wave functions, 304
- Prompt Assessment of Global Earthquake Response (PAGER), 499
- Propagation effects, 52
- Proportions, 1119
- Proposal distribution, 790
- Prospectivity mapping, 1117
- Prospectivity modelling, 453
- PROSPECTOR, 38
- PROSPECTOR II, 38
- Provenance, 371
- Provenance ontology (PROV-O), 89
- Proximity regression, 1116, 1117
 - applications, 1117
 - definition, 1116
- Pruning, 257
- PSD analysis and blasting optimization, 880
- Pseudo-random number generation process, 224
- PSO-RBF estimation model, 1051
- Psychology, 1119, 1220
- Public health, 38
 - authority, 1440
- Public policies, 231
- Pulacayo Zinc Deposit, 438
- Pulp duplicates, 1133
- Pulse sequence, 1465
- Pushbroom scanners, 625
- p-value, 1431
- P waves, 329
- Pyramid algorithm, 1629
- Pyroclastic deposits, 1500
- Pythagorean theorem, 708
- Python, 498
- Python-based landscape evolution model, 323
- Q**
 - Q–Q graphs, 1441, 1443
 - Q–Q plots, 1333, 1334, 1429, 1430
 - QR decomposition, 1434
 - QSCreator, 1154
 - Quadrant symmetric, 1348
 - Quadratic discriminant analysis (QDA), 147, 982
 - Quadratic function, 726
 - Quadratic optimization, 1524, 1525
 - Quadratic surface fitting method, 1051
 - Quadrature mirror filters (QMFs), 378
 - Qualified references, 371
 - Qualitative categories, 202
 - Qualitative feature, 1600
 - Qualitative methods, 1028
 - Qualitative variables, 2
 - Quality assurance (QA), 1130
 - accuracy, standards and control limits, 1127
 - challenges in programs, 1128
 - definition, 1126
 - precision and duplicates, 1127–1128
 - QAQC procedure, 1126–1127
 - and quality control, 1126
 - Quality Assurance and Control Program, 1130
 - Quality control (QC), 1134
 - analysis of QC duplicates, 1133–1134
 - analytical duplicates/pulp duplicates, 1133
 - blanks, 1130–1131
 - challenges in QC programs, 1134
 - check/umpire assays, 1131
 - coarse duplicates/prep duplicates, 1132
 - definition, 1130
 - field duplicates/check samples, 1132
 - measure and monitor accuracy and bias, 1130–1131
 - measure and monitor precision, 1132–1133
 - standards, 1130
 - Quality management, 1130
 - Quantification, 1296
 - Quantile(s), 436, 1055, 1429
 - loss function, 776
 - regression trees, 1183
 - Quantitative analysis, 857
 - Quantitative characterization, 1491
 - Quantitative feature, 1600
 - Quantitative fractography, 1498
 - Quantitative geological measurements, 202
 - Quantitative geomorphology, 1135, 1136
 - mathematical morphology, 1144, 1145
 - Quantitative model, 1167
 - Quantitative stratigraphy, 16, 193, 513, 515, 516, 1165
 - correlation, 1163–1165
 - data selection, run parameters and unique events, 1153–1154
 - definition, 1152
 - dendrogram, 1159
 - QSCreator, 1154
 - ranked optimum sequence, 1156–1158
 - RASC, 1155–1156
 - scaled optimum sequence, 1158–1159
 - testing stratigraphic fidelity, 1159–1161
 - Quantitative target selection, 1166–1168
 - Quantitative techniques for resource evaluation, 1259
 - Quantitative terrain analysis, 635
 - Quantitative variables, 1, 381
 - Quantization noise, 1296
 - Quartiles, 1055
 - Quartimax, 1124

- Quartimin, 1124
- Quasi-linearization approach, 675, 676
- Quasi-stationarity, 545, 1297
- Quaternions
 - definition, 1169
 - geodesics, 1172–1173
 - real, 1171
- R**
- Radar equation, 1207
- Radar scattering coefficient, 1207
- Radial basis functions (RBFs), 1181, 1407, 1522
 - approximation methods, 1180
 - CPD, 1175, 1176
 - definition, 1175
 - derivative and inequality constraints, 1177–1179
 - discontinuities, 1179–1180
 - implicit modelling, 1177
 - limitations, 1181
 - multiquadric, 1175
 - selection, 1180
 - software implementations, 1181
 - SPD, 1176
- Radial power, 1180
- Radioactive decay of heavy isotopes, 330
- Radioactive isotopes, 576
- Radiometric resolution, 626, 1207, 1307
- Radon-222, 579
- Radon transform, 56–57
- Ramp filter, 60
- Ramp function, 1290
- Random, 851
 - composition, 13
 - events in time, 1219
 - experiments, 1188
 - field, 544, 1457
 - fractals, 430
 - geometry, 1458–1459
 - granular assemblies, 1462
 - mixture, 1463
 - network, 1463
 - packing, 1465
 - permutation, 1278
 - process, 47, 544, 546
 - realization, 1189
 - resistors, 1463
 - sample, 203, 1190
 - sets, 1367, 1368
 - signals, 1289
 - tessellations, 1507
 - vector, 546, 1516, 1517
 - walk, 50, 68, 792, 1464–1465
- Random forest (RF), 145, 147, 150, 153, 258, 265, 465, 639, 1056, 1061, 1062, 1273
- Random forest algorithm
 - definition, 1182
 - earth science applications, 1184
 - model interpretation, 1183
 - variations and software, 1183–1184
- Random forest classifier (RFC), 1322
- Random function, 63, 544, 845, 1186, 1276, 1346, 1376, 1410, 1456
 - algorithmically-defined random functions, 1187–1188
 - covariance, 1186
 - definition, 1185
 - ergodicity, 1187
 - expected value, 1186
 - indicator random function, 1187
 - multiGaussian random function model, 1187
 - realization vs. random variable, 1185
 - stationarity, 1186–1187
 - variance, 1186
 - variogram, 1186
- Randomly distributed points, 1495
- Random variable, 203, 342, 549, 607, 608, 778, 1276, 1517, 1605, 1607
 - analytical and numerical, 1190
 - combination of, 1191–1192
 - definition, 1189
 - discrete and continuous, 1189–1190
- Range, 545, 1214, 1311
 - anisotropy, 1348
 - of correlation, 1074
- Rank, 1004
 - coefficient, 207
 - correlation, 207, 208
 - of a matrix, 837, 841
- Ranked optimum sequence, 1156–1158
- Ranking, 194
- Ranking and Scaling (RASC), 194, 196–199, 588, 1152–1156, 1158–1163, 1165
- Rao, Calyampudi Radhakrishna, 1194
 - awards, 1194
 - orthogonal arrays, 1194
 - Rao-Blackwellization, 1194
 - Score and Lagrangian multiplier tests, 1194
- Rao, S. V. L. N.
 - education, 1195
 - instrumental role, 1195
- Rao-Blackwellization, 1194
- Rare-earth elements, 39
- Rare event, 147
- Raster, 930
 - data analysis, 479
- Rate-and-state friction law, 1483
- Ratio-correlation, 114
- Rational polynomial coefficient (RPC), 281, 283
- Rational polynomial model (RPM), 281–284
- Rational quadratic models, 1349
- Ray method, in wave propagation, 1044
- Ray parameter, 676
- Re3data, 235
- Real-data time series, 1330
- Realization, 621, 1185, 1276–1279
 - definition, 1196
 - equiprobable, 1196
 - generation methods, 1196
 - for mechanistic model uncertainty, 1197–1198
 - ordering and ranking, 1197
- Real quaternions, 1171
 - skew field of, 1171–1172
- Real valued covariance, 1352
- Reasenber-Jones model, 1478
- Recall, 758
- Recharge, 344
- Recognition, 1053
- ReCon, 1322
- Reconstructed porous media, 1464
- Reconstruction, 1078
 - algorithm, 1459
- Recovery functions, 557, 558
- Recrystallization, 1263
- Rectangular cartogram, 103, 104

- Rectangular function, 1290
- Recurrent connection networks, 1266
- Recurrent graph neural networks (RGNNs), 270
- Recurrent neural networks (RNNs), 269–270, 1274
- Recursive filtering, 290, 293
- RedHat Linux, 499
- Reduced major axis (RMA) regression, 1131
 - AERONET measurements, 1200
 - AOD, 1201
 - definition, 1199, 1200
 - dependent and independent variables, 1200, 1201
 - evaluation method, 1201
 - OLS, 1200–1202
 - satellite-based and ground-based instruments, 1200
 - utility, 1200
- Reduction, 1602
- Redundant Array of Independent Disks (RAID), 125
- Reed-Jorgensen model, 425
- Reference ontology, 1015
- Reference shape, 1280
- Referent vector, 1267, 1269
- Reflectance absorption, 579
- Reflection/axial symmetry, 1380
- Reflection coefficients, 52
- Reflection destination, 876
- Reflection losses, 51
- Regional anomaly, 1639
- Regionalization, 544, 978
- Regionalized favorability analysis, 386
- Regionalized variables, 1607
- Regional mineral potential estimation, 1100, 1102
- Regional observation mode (ROM) revisit time, 1322
- Region model, 94
- Region of support, 1306
- Registration metadata, 657, 658
- Regression, 323, 718, 958, 959, 1523, 1524
 - estimators, 729
 - loss function, 775–777
 - problem, 1182
- Regression analysis (RA), 240
 - data classification, 1203
 - definition, 1203
 - digital images, 1203
 - evaluation, 1204
 - example, 1204
 - linear and non-linear approaches, 1203
 - methods, 1203
 - metrics, 1205
 - remote sensing, 1203, 1204
 - themes, 1203
 - weather parameters, 1203
- Regular grid format, 289
- Regularity condition, 1304
- Regularization, 585, 679, 684, 795
- Regular point processes, 1077
- Reid's elastic rebound theory, 1475
- Reinforced learning, 1267
- Reinforcement learning, 36, 782
- Rejection algorithm, 892
- Related distributions, 1001
- Relating, 622
- Relational databases, 498
- Relative error, 1495
- Relatively robust representation (RRR), 337
- Relative permeability, 956
- Relative plate motions, 350
- Relative size, 775
- Relative standard deviation (RSD), 1128
- Relaxation, 1464
 - time, 1466
- Relaxivity, 1465
- Releasing code, 1212
- Releasing data, 1211–1212
- Remotely sensed imagery, 1240
- Remote sensing (RS), 213, 214, 227, 273, 300, 308, 323, 404, 520, 577, 864, 929, 1011, 1095, 1203, 1204, 1307
 - advantages, 1513
 - applications, 1208–1209
 - continuous mapping of topographic change, D-InSAR, 580
 - data pre-processing, products and analysis, 1208
 - definition, 1206
 - and geoscience applications, 594
 - hydrothermally altered minerals, detection and mapping of, 579
 - image analysis, 1514, 1537
 - lineament-based fracture characterization, 579–580
 - physical basis and signatures, 1206–1207
 - satellites, 228–230
 - sensors, platforms and orbits, 1207–1208
 - surface temperature estimation, 577–578
 - trace gases, 1382–1383
- Remote sensing data, 227, 249, 250, 475
 - acquisition, 227
 - sources, 230
- Remote sensing technology (RS), 804
- Renewal density, 1477
- Renewal/recurrence models, 1477–1478
- Renormalization group, 1597
 - models, 109, 1469
- Repositories, 370
- Representative elementary volume (REV) scale, 957
- Representative volume elements (RVE), 1507
- Reproducibility, 1209
 - vs. replicability, 1209–1210
 - and replicability in science, 237
- Resampling, 1240–1243
 - methods, 1430–1433
 - in remote sensing applications, 1243
- Rescaled range, 1213, 1214
- Rescaled-range analysis (R/S analysis), 1311, 1314
 - Anis-Lloyd corrected R/S Hurst exponent, 1216, 1217
 - definition, 1213
 - detrended Hurst analysis, 1215–1216
 - Hurst exponent H as indicator of long-term behavior, 1215
 - modified scaling law, 1216
 - periodogram regression, 1216–1217
 - power spectrum, 1216
 - practical issues, 1217
 - random processes, with Hurst exponent $H \neq 1/2$, 1214
 - structure function method, 1216
 - traditional R/S analysis, 1215
- Reserve estimation, 27
- Reservoir extension, 27
- Reservoir petrophysical characteristics, 1600
- Reservoir zones, 1305
- Residual(s), 1034, 1441, 1442
 - analysis, 1037, 1476
 - anisotropy, 1574
 - anomaly, 1639
 - copper concentration values, 1571
 - 2-D autocorrelation function, 1574
 - exploratory boreholes, 1571
 - histograms, 1571
 - kriging method, 1574

- saturation, 956
 - statistical inference, 1572
 - structural deformation, 1573
 - trend+signal+noise model, 1573
 - uranium mineralization, 1573
 - Whalesback copper deposit, 1571
 - Residual learning networks (ResNets), 266, 663
 - Residual sum of squares (RSS), 1036
 - Resilience, 298
 - Resistivity, 581
 - Resolution, 227
 - Resource description framework (RDF), 371, 498
 - Resource model, 862
 - Resource-oriented approach, 298
 - Response function, 292
 - Response *N*-vector, 1433
 - Response time, 321
 - Response variables, 1033
 - Reusable, 371–372
 - Reverse-jump MCMC, 794
 - Reversible-jump Monte Carlo, 893
 - Reyment, Richard Arthur
 - education, 1218
 - publications, 1219
 - research, 1219
 - scientific career, 1218
 - Reynolds number, 955, 1576, 1577
 - RGB images, 263
 - Richardson-Mandelbrot plot, 26
 - Richardson number, 1577
 - Rich metadata, 369
 - Riemann manifold, 711
 - Right eigenvector, 337
 - Right skewed, 1313
 - Rigorous sensor model (RSM), 280, 283
 - coordinate systems, 281
 - ground coordinates, estimation of, 281
 - image coordinates, estimation of, 281
 - Riley Composite Standard (SRC), 196
 - Ring of fire, 332
 - Ripley's estimator, 422
 - Ripple, 1638
 - River(s), 25, 1054
 - long profile, 321, 1140
 - morphology, 1139
 - R language, 1184
 - RMS amplitude, 52, 53
 - RMS stacking velocity, 52
 - Robot autonomous systems (RAS), 40
 - Robot task management, 37
 - Robust estimation, 446
 - Robust least squares methods, 734–735
 - Robustness, 778
 - Robust regression, 689–691
 - Robust statistics
 - and data analysis, 1225
 - definition, 1225
 - least-squares estimator, 1226
 - M estimator, 1227
 - regression model, 1225–1226
 - Rock damage, 1469
 - Rock failure, 1469
 - Rock fracture pattern, 1229
 - aperture, 1231–1232
 - connectivity of fracture, 1231
 - density, 1231
 - fracture modeling, 1232
 - length distribution, 1230
 - orientation distribution, 1230
 - representative elements for characterizing, 1230
 - Rock fragmentation, 25
 - Rock hydration, 1530
 - Rock lithology, 1184
 - Rock mechanics, 61, 333
 - Rock physics models, 542
 - Rock properties, 688
 - Rock type, 207
 - Rock weathering, 1136, 1137
 - Rodionov, D.A., 1003, 1005, 1008, 1234–1235
 - Rodrigues' formula, 352, 1172
 - Rodriguez-Iturbe, Ignacio, 1235
 - Root mean square deviation (RMSD), 1236
 - Root mean square error (RMSE), 292, 676, 736, 775, 1204, 1236–1238, 1241
 - Root node, 257
 - Rosin-Rammler-Sperling-Bennett (RRSB) distribution, 591
 - Rotating drum memory module, 37
 - Rotation(s)
 - concatenation of, 1172
 - matrix, 282
 - in two-dimensional and three-dimensional Euclidean, 1170
 - Rotational transformation, 1301
 - Rotation, in three-dimensional Euclidean, 351–352
 - Roughness, 431, 929
 - Rough set theory (RST), 1600
 - approximation accuracy, 1603
 - approximations in, 1601–1602
 - classification using, 1604
 - reduction of features, 1602
 - rules, 1604
 - Rounded zeros and zero count, 138
 - Row matrix, 837
 - R-package compositions, 142
 - R*-program, 506
 - Ruff, 236
 - R square, 1204
 - R statistical programming language, 495
 - Rug plots, 1594
 - Rules generation, 1602
- S**
- Same Team, Same Experimental Setup (STSE), 1210
 - Sample, 202, 203
 - covariance, 203
 - design, 147
 - mean, 203
 - size, 436, 1240
 - space, 10, 1185, 1189
 - Sampled values, 1346
 - Sampling, 1239
 - error, 1240
 - in geospatial data analysis, 1240
 - interval, 52
 - with replacement, 1431
 - strategies, 1497, 1500
 - without replacement, 1431
 - Sampling-holding, 1296
 - San Francisco Earthquake, 333, 334
 - Sandstone, 27, 1216, 1465, 1468
 - Santorini Island, 39
 - SAR (Synthetic Aperture Radar) Land applications, 790
 - Satellite data, 967
 - Satellite sensor, 227

- Saturation, 574, 1467
- Saturn, 40
- Scalar function, 1177
- Scalar product, 290
- Scalar product operation, 19
- Scalar quantity, 1352
- Scale, 1168
 - dependency, 634
 - shift factor, 1306
 - independent, 404
 - invariance, 135, 430, 1470
 - invariant, 406
 - parameter, 1304
- Scale-dependent measure, 775
- Scaled invariance feature transform (SIFT), 877
- Scaled optimum sequence, 1158–1159
- Scale-frequency relation, 1304
- Scale-invariant decomposition, 1308
- Scale invariant feature transform (SIFT), 285, 287
 - keypoint descriptor, 288
 - keypoint localization, 287–288
 - orientation assignment, 288
 - scale-space extrema detection, 287
- Scale-space extrema detection, 287
- Scaling, 27, 194, 196, 197, 1458
 - definition, 1244
 - and dimension, 405
 - exponent, 24, 1310, 1463
 - factor, 406
 - function, 1307
 - functions, fluctuations and the fluctuation exponent, 1244–1246
 - in higher dimensional space, 1251
 - laws, 24
 - noises, 1093
 - sets, 1244
 - stratification in earth and atmosphere, 1251–1253
 - structure, 927
 - structure functions and spectra, 1246–1248
 - symmetry, 26
- Scalogram, 1305, 1306, 1315
- Scanning electric microscope, 1184
- Scattered data approximation, 1175
- Scattergram, 1257–1258
- Scatter graphs, 201, 202
- Scatterplots, 1257, 1258
- Scene reconstruction, 36
- Schema.org, 858
- Schewart Control Chart, 1127
- Schmitt net, 1486–1488
- Schuenemeyer, J.H., 1259
- Schwarzacher, Walther, 1260
- Science-fiction, 38
- Science of team science studies, 91
- Science of Where, 587
- Scientific workflows, 1211–1212
- Scikit-Learn package, 1184
- Score matrix, 975
- Score test, 723
- Scoring procedure, 1480
- Screen effect, 609
- Scree test, 1221
- Sea ice, 39
- Searchable catalog, 370
- Searchable resource, 370
- Search neighborhood, 1278
- Seasonal autoregressive integrated moving average (SARIMA), 1546
- Sea surface temperature (SST), 644
- Second-order intensity function, 421
- Second-order mass exponent, 418
- Second-order moments, 1605
- Second-order phase transitions, 1598
- Second order source separation (SOS) model, 1558
- Second-order spatial statistics, 605
- Second-order stationarity, 544, 545, 1365, 1411, 1606
- Second-order stationary process, 546, 559, 560
- Second-order stationary random field, 1607
- Second principle of geostatistics, 557
- Second Shannon coding theorem, 346
- Section, 1490, 1491
- Sediment(s), 1139, 1261, 1511
 - transport, 1143
- Sedimentary conditions, 1262
- Sedimentary rocks, 25, 818, 1261, 1458, 1497, 1511, 1513
- Sedimentary sequence, 1005
- Sedimentary structures in rocks, 1263
- Sedimentation and diagenesis, 1263, 1264
- Sedimentology, 39, 1219
- Sedimentology-Mathematical Index of Sediment Grain Size Classification, 809
- SEDSIM, 603
- Segmented images, 405
- Seismic activity, 336
- Seismic amplitude, 51, 66
- Seismic data analysis, 1292
- Seismic data processing, 51, 292
- Seismic events, 330
- Seismic inversion, 539
 - problem, 65, 69
- Seismic moment, 1397
- Seismic profiles, 51
- Seismic reflection surveys, 539, 1397
- Seismic refraction method, 1051
- Seismic signal analysis, 869
- Seismic tomographic models, 470
- Seismic tomography, 1499
- Seismic trace, 52
- Seismic velocity, 675–677
- Seismometrics, 1472
- Seismostatistics, 1472
- Seismology, 696
- Selectivity of distributions, 557
- Self-adaptive map, 1267
- Self-affine, 404, 433, 1214
 - fractals, 431–432, 1308, 1310
- Self-affinity, 1251
- Self-consistency
 - and effective media, 1463–1464
 - equation, 1460
- Self-information, 346
- Self learning, 36
- Self-/mutually exciting models, 1478, 1479
- Self-organized criticality, 25, 109, 321, 434, 1599
- Self-organizing feature map (SOFM) classifier, 701
- Self-organizing maps (SOM), 36, 145, 240, 1061, 1267, 1530
- Self-similar, 404
 - fractals, 430–431, 1308
 - process, 1311
- Self-similarity, 24–26, 322, 620
- Semantic links, 1536
- Semantic relationship, 1013
- Semantic Web, 88, 498, 859, 1019
- Semantic word sequences, 1536

- Semicircular vector means (SVM), 116
- Semi-discrete collocation method, 1045
- Semi-discrete method, 1045
- Semi-lunar tidal influences, 1314
- Semisolid state, 574
- Semi-supervised learning, 36, 782
- Semivariogram, 50, 418, 419, 545, 548, 1215, 1346, 1349, 1365
 - matrix, 547, 870
 - modeling, 548
- Sensed data, 3
- Sensitivity, 323
- Sensors, 226, 227
- Sensor technology, 465
- Sentiment analysis, 233
- Sentinel multispectral imager (MSI), 758
- Separability, 1348
- Separation of variables, 320
- Separation vector, 1346
- Sequence classification
 - definition, 1273
 - in Geosciences, 1273–1274
 - using deep learning algorithms, 1274
- Sequence-to-label classification, 1274
- Sequence-to-sequence classification, 1274
- Sequential approach, 1505
- Sequential binary partition (SBP), 138
- Sequential Gaussian simulation (SGSIM), 560, 610, 1188, 1278, 1543, 1585
 - estimation vs. simulation, 1277
 - random function, 1276
 - sequential simulation, 1277–1278
- Sequential indicator simulation (SISIM), 561, 1324, 1543
- Sequential minimal optimization (SMO), 1525
- Sequential patterns, 1053
- Sequential simulation, 891, 1277–1278, 1325
- Serpentinization, 1569
- Serra, Jean
 - early life, 1280
 - multimedia application, 1280
 - Texture Analyser, 1280
- Service oriented architecture (SOA), 1620
- Settling velocity versus sphere diameter of quartz particles, 1262
- Severe weather forecasts, 1184
- Sexual dimorphism, 1219
- SGeMS, 1351
- Shale, 1461, 1465
 - barriers, 1276
- Shannon, Claude, 842
- Shannon entropy, 346, 842, 843, 1460, 1461
- Shape, 1221, 1280, 1492, 1496, 1513–1515
 - analogistic method, 1280
 - deformation grids, 1283–1285
 - factor, 1458, 1496
 - formal, 1280
 - form ratios, form factors and particle shape spaces, 1282–1284
 - geometric shape spaces, 1286–1287
 - mathematical aspects, 1281–1282
 - parameters, 1176
 - reference, 1280
 - shape deformation grids, 1283–1285
- Shapiro–Wilk test, 1441
- Share-alike, 1018
- Shared dictionaries, 234
- Shepard diagram, 941
- Shepard-Kruskal iterative algorithm, 941
- Shepard's method, 662
- Shewhart control charts, 1130
- Shimbel index, 1148
- Short-distance nugget effect modeling, 439–441
- Shorth estimate, 729
- Short-range correlations, 1312
- Short-term mine planning, 861
- Short time Fourier transform (STFT), 1303, 1627
- Short time series, 622
- Short wave infrared (SWIR), 627
- Sibson's interpolation, 662
- Side looking airborne radar (SLAR), 1207
- Sieve analysis, 592
- Sifting operations, 1315
- Sigmoidal membership function, 855
- Sigmoid function, 993
- Signal
 - comparison, 1303
 - definition, 1288, 1297
 - extraction, 36
 - processing, 1297
 - types, 1289
- Signal analysis
 - classification, 1289
 - convolution and deconvolution principle, 1295
 - definition, 1288
 - digitization, 1295, 1296
 - FFT, 1296
 - filtering, 1295
 - Fourier series, 1291
 - Fourier transform, 1294, 1295
 - frequency representation, 1290, 1291
 - functions, 1290, 1291
 - geophysics example, 1296
 - modulation, 1295
 - noise, 1288
 - processing systems, 1288
 - s/N* ratio, 1289
- Signal-plus-noise method, 438
- Signal plus noise model, 1329, 1575
- Signal-to-noise ratio (SNR), 849, 1289
- Signature matrix (SME), 178
- Sign function, 1290
- Sill, 545
 - anisotropy, 1348
- Similarity, 541, 621, 977
 - and dissimilarity indices, 1119
 - solutions, 1044–1045
- Simple cokriging, 550
 - of indicators, 552
- Simple features, 1015
- Simple kriging, 549, 707, 708, 1278
 - of indicators, 552
 - system, 1378
- Simple random sampling, 1239
- Simplex method, 10, 718
- Simpson, 1119
- Simulated annealing (SA), 566, 798, 894–895, 964, 1050, 1052, 1320–1322, 1459, 1466–1467, 1614
 - applications, 1618
 - implementation, 1618
 - initialization, 1618
 - methodology, 1615
 - VFSR (*see* Very fast simulated reannealing)
- Simulation, 166, 168, 544, 1430, 1466
 - Boolean models, 566
 - Boolean random function model, 566

Simulation (*cont.*)

- categorical variables, 1325
 - conditional, 562
 - continuous variables, 1325
 - dead-leaves model, 566
 - definition, 1322
 - dilution model, 566
 - Gibbs Sampler, 563, 564
 - hard-data, 559
 - indicator approach, 1323–1324
 - matrix decomposition, 561, 562
 - multiGaussian approach, 1325
 - multiple-point, 1325
 - multi-point simulation, 564, 565
 - multi-point statistics, 559
 - object-based models, 566
 - object-based simulation techniques, 559, 566
 - plurigaussian simulation, 562, 563
 - probability field simulation, 565
 - realizations, 1326–1327
 - scale of observation, 559
 - scales, 559
 - second-order stationary process, 559, 560
 - sequential, 1325
 - Sequential Gaussian Simulation, 560
 - Sequential Indicator Simulation, 561
 - simulated annealing, 566
 - smoothing, 559
 - smooth patterns, 559
 - soft-data, 559
 - substitution random fields, 566
 - techniques, 559
 - Truncated Gaussian simulation, 562, 563
 - Turning Bands algorithm, 565
 - two-point statistics, 559
- Simulation methods, 961
- algorithm-based methods, 963
 - definitions, 961
 - Gaussian simulation, 963
 - large-scale continuity, 962
 - model-based methods, 962
 - multiple-point statistics, 962
 - nodes, 961
 - optimization-based approaches, 964
 - pixel and patch based unilateral methods, 964
 - pixel-based methods, 963
 - pixel based sequential methods, 963
 - spatio-temporal information, 962
- Simultaneous algebraic reconstructive technique (SART), 22–23
- Simultaneous autoregressive (SAR) model, 1366
- Simultaneous diagonalization, 555
- Simultaneous iterative reconstructive technique (SIRT), 21
- Simultaneous localization and mapping (SLAM), 289
- Simultaneous third-order tensor diagonalization (STOTD), 643
- Single-channel convolutional filters, 293
- Single imputation, 638
- Single matrix equation, 279
- Single normal equation, 565
- Singleton membership function, 855
- Singularity(ies), 620, 1305, 1314
- index, 745–746
 - manifold, 1315
 - spectrum, 322
- Singularity analysis, 414–416
- definition, 1332
 - geochemical atlas of Cyprus, 1336
 - methodology, 1334–1336
- Singular spectrum analysis (SSA), 1328, 1330, 1331
- decomposition, 1328
 - definition, 1328
 - embedding, 1328
 - extensions and relations, 1332
 - grouping, 1329
 - prediction, 1330
 - reconstruction, 1329
 - separability, 1329–1330
- Singular value decomposition (SVD), 140, 337, 643, 673, 684, 686–687
- Singular values, 687
- Sinusosity, 25
- Sinusoid basis functions, 1308
- Siri*, 34
- Site-specific/specificatory S-KB, 73
- Size, 589, 1221, 1492, 1513–1515
- Size-distribution, 25, 26
- Skeleton, 1459
- Skeletonization, 322, 923
- Skewness, 1313
- Skew-normal distribution, 1114
- simplex, 12
- Sklearn package in Python, 357
- Slepian sequences, 303
- Slice sampling, 793
- Sliding windows, 52, 53
- Slip orientation identification, 329
- Slope, 622
- anisotropy, 1348
- Sloss' geometric space-filling model, 395
- Smart geological survey, 294
- Smile effect, 627
- Smoothing, 559
- Smoothing, 1371
- methods, 1548
 - splines, 506–507, 1406
- Smoothing factor (SF), 506
- Smoothing filter
- central notion, 1339
 - definition, 1339, 1342
 - linear filters, 1341
 - mean filters, 1342
 - statistical, 1342
- Snedecor F-distribution, 631, 632
- SNESIM (Single Normal Equation Simulation), 565, 605, 606
- Socioeconomic factors, 1589
- Soft clustering, 447
- Soft data, 72–74, 559
- Software as a Service (SaaS), 124, 498
- Soil, 1467
- mass, 573
 - science, 25, 61
- Solar/planetary systems, 1026
- Solid deformation, 211
- Solution, 679, 681–686, 688
- Sonic log, 818
- Source rock(s), 1470
- maturity, 1470
- Sources of uncertainty, 862
- Space/time series, 1093–1094
- Fourier series analysis, 1094
- Space, 1055
- Spaceborne sensor, 1207
- Space-time diffusion clustering models, 1481

- Space-time domain, 1374
- Space-time variogram function, 1375
- Spanning-tree progression analysis of density-normalized events (SPADE), 1530
- SPARQL, 1015
- Sparse matrix inversion algorithms, 710
- Sparse-matrix methods, 1367
- Spatial analysis, 1363, 1560
 - correlations and knowledge, 1345
 - definition, 1342
 - derivative information, 1343
 - geospatial data, 1343
- Spatial association, 453
- Spatial autocorrelation, 870, 897, 900
 - characteristics, 897, 1350
 - geostatistical analysis, 1351
 - hole effect, 1351
 - linear behavior, 1350
 - nested structures, 1350
 - parabolic behavior, 1350
 - periodic behavior, 1351
 - qualitative variables, 897
 - sill/range, 1350
 - similarity among attributes, 897
 - similarity of location, 897
 - structural analysis, 1351
 - tolerance region, 1351
 - trend, 1351
- Spatial best linear unbiased predictor (BLUP), 705
- Spatial bootstrap
 - procedure, 97, 99
 - resampling, 98–99
- Spatial connectivity, 605
- Spatial correlation, 628, 1233
- Spatial cumulants, 606, 607
- Spatial data, 520, 756, 1361
 - analysis, 183, 1053
 - characteristics, 1353, 1354
 - classification, 1354
 - definition, 1353
 - geostatistical data, 1354–1355
 - lattice data, 1355–1357
 - mining, 1053
 - models, 476–477
 - point patterns, 1356–1358
- Spatial data infrastructures (SDI), 534, 1015, 1358
- Spatial decision support systems (SDSS), 482, 1241
- Spatial-dependence measures, 1368
- Spatial direction, 1348
- Spatial discretization, 1364, 1369–1371
- Spatial distribution, 410
- Spatial entities, 1055
- Spatial frequency, 1636
- Spatial heterogeneities, 485, 486
- Spatial interaction models, 1560
- Spatial interpolation, 666, 671, 1344
- Spatially referenced, 1055
- Spatial measurement, 1344
- Spatial misregistration, 625
- Spatial network, 481
- Spatial object, 474
- Spatial orthogonality, 555
- Spatial phenomena, 1342
- Spatial point pattern (SPP), 1058, 1059
- Spatial point process, 1367, 1368
- Spatial process models, 1363, 1364
- Spatial proximity, 897
- Spatial query, 1344
- Spatial random field, 1605
- Spatial regression methods, 1117
- Spatial resolution, 626, 1207, 1307
- Spatial RF (SRF), 1346
- Spatial scanners, 625, 626
- Spatial statistics, 587, 605, 1344, 1345, 1605
 - Bayesian hierarchical model, 1364
 - Bayes' Rule, 1364
 - conditionally independent, 1363
 - definition, 1362
 - empirical hierarchical model, 1364
 - geostatistical process, 1363–1365
 - hierarchical statistical model, 1363
 - joint probability model, 1363
 - lattice process, 1364, 1366, 1367
 - marginal distribution, 1363
 - marks, 1363
 - measurement uncertainty, 1363
 - models, 605, 1371
 - multivariate spatial processes, 1368, 1369
 - point process, 1364
 - predictive distribution, 1364
 - random sets, 1367, 1368
 - scientific uncertainty, 1363
 - simple kriging predictor, 1364
 - spatial discretization, 1369, 1370
 - spatial-location information, 1363
 - spatial point process, 1367, 1368
 - spatiotemporal processes, 1370, 1371
 - uncertainty, 1363
- Spatial-temporal random variables, 1376
- Spatial uncertainty, 348–349, 1585–1587
- Spatial variability, 1185
 - of fluvial processes, 1140
- Spatial variables, 382
- Spatio-temporal
 - anisotropy in space-time, 1379
 - basic theoretical framework, 1375–1376
 - computational aspects, 1378–1379
 - data, 438, 961
 - definition, 1373
 - domain, 1353
 - geostatistical approach, 1373–1374
 - kernel, 1392–1394
 - models, 960
 - models for variogram/covariance function, 1377
 - model validation, 1377–1378
 - prediction, 1378
 - problems in space-time, 1374–1375
 - processes, 1366, 1370, 1371
 - sample variogram and covariogram, 1376
 - sampling, 979
 - scales, 319
 - separability, 1379–1380
 - signals, 1297
 - space-time domain, 1374
 - structural analysis, 1376
 - symmetry, 1380
- Spatio-temporal analysis, 1384–1385
 - validation of datasets, 1383–1384
 - visualization, 1383–1384
- Spatiotemporal modeling, 1387
 - applications of, 1387
 - definition, 1386

- Spatiotemporal modeling (*cont.*)
 - outlier detection algorithms, 1389
 - outlier detection method, 1388
- Spatio-temporal point pattern (STPP), 1058, 1059
- Spatiotemporal random field theory (STRF), 73–74
- Spatiotemporal weighted regression (STWR), 487, 1390, 1391, 1394, 1395
- Separating interior, 1562
- Spearman rank correlation, 130
- Spearman's coefficient of rank correlation, 207
- Specification, 1057
- Specific genus, 1459
- Specific length, 1494
- Specific surface, 27, 1457, 1458
 - area, 1494
- Specific thermal conductivities, 1466, 1467
- Spectral, 929
 - amplitudes, 1301
 - clustering, 1529
 - distribution function, 1352
 - estimator, 845
 - exponent, 432
 - index, 302
 - radius, 337
 - ratio, 302
 - resolution, 1207
 - signature, 625
 - theory, 337
 - unmixing, 629
- Spectral analysis, 526, 528–532
 - applications within geosciences, 1397–1398
 - definition, 1396
 - non-parametric methods, 1397
 - parametric methods, 1397
 - power spectral density, 1397
- Spectral density, 845, 1217
 - function, 1352
 - method, 620–621
- Spectral element method (SEM), 470
- Spectral fitting method (SFM), 405
- Spectra of amplitude, 620
- Spectrogram analysis, 846
- Spectrum, 1313, 1637
 - analysis, 1399
 - of exponents, 623
 - of singularities, 623
- Spectrum-area (S-A) fractal filtering method, 812
- Spectrum-area (S-A) fractal model, 1398, 1399
 - methodology, 1399–1400
 - in Sweden, 1400–1401
- Spectrum-area (S-A) model, 1334
- Spectrum-area multifractal model (S-A), 947
- Spherical crystallographic X-ray transform, 218
- Spherical divergence, 49
- Spherical model, 1349
- Spherical neighbourhood geometry, 669
- Spherical pores, 1460, 1465
- Spherical variogram, 1611
- Spin magnetization, 1464
- τ - s plane, 1306
- Spline approximation, 52
- B -splines, 1404–1406
- Splines, 516
 - cardinal, 1405
 - cubic, 1404
 - end conditions, 1403, 1404
 - equi-spaced nodes, 1404
 - least square, 1406
 - linear, 1403
 - natural, 1403
 - non-uniformity spaced grids, 1405, 1406
 - parametric, 1406
 - smoothing, 1180, 1406
- Splitting, 257
- Spreadsheet, 232
- Spurious correlation, 135
- Spurs, 924, 926
- Squared Euclidean, 1120
- Square integrable, 1304
- Square matrix, 837
- Squares estimators of generalized space-time autoregressive (STAR), 1000
- Stability, 1054, 1269
- Stable distributions, 1430
- Stable frequency distributions, 408
- Standard(s), 1130
 - atmosphere, 584
 - bootstrap confidence intervals, 1431
 - statistical techniques, 544
- Standard deviation, 203, 436, 1190, 1215, 1311, 1638
 - continuous distributions, parameter space, densities, expected values and, 1408
 - definition, 1408
- Standard error (SE), 1237, 1440, 1441, 1443
- Standardization, 1015
 - techniques, 131
- Standardized communications protocol, 370
- Standardized Gaussian, 345
 - histogram, 1279
- Standard reference materials (SRM), 1127, 1130, 1131
- Standards-compliance, 370
- Stanford Center for Reservoir Forecasting (SCRF), 693
- StanfordCoreNLP, 38
- Stanford Heuristic Programming Project, 34
- State, 1265
- States of knowledge, 64
- State-space model (SSM), 1546
- Stationarity, 50, 146, 545, 964, 1073, 1346, 1606
 - definition, 1410
 - diagnostic tools, 1413
 - formulation, 1411–1413
 - hypotheses, 1606
 - intrinsic, 1411
 - limitations, 1413
 - second-order, 1412
 - strict, 1411
- Stationary, 1186, 1187
 - random function, 1187
 - signal, 49
 - time-series, 292
- Statistical approaches, 323
- Statistical aspects, 1034
- Statistical bias
 - basic statistics, global fluctuations for, 1417–1419
 - conditional bias, 1416–1417, 1422–1427
 - estimate and estimator, 1415
 - estimation bias, 1427
 - estimator quality, 1415
 - spatial correlation, 1420
 - spatial estimate and spatial estimator, 1415
 - spatial estimation, local fluctuations for, 1419–1420
 - transfer function bias, 1416

- Statistical computing
 - definition, 1428
 - exploratory data analysis, 1428–1430
 - linear regression, 1433–1436
 - nonlinear multivariable maximum likelihood estimation, 1436–1438
 - resampling methods, 1430–1433
- Statistical estimation, 845
- Statistical formulation
 - applications, 544
 - assumptions of stationarity, 545
 - coregionalization, 546, 547
 - covariance and variogram functions, 545
 - generalized covariance function, 545, 546
 - non-stationary random functions, 545, 546
 - prediction problems, 545
 - probabilistic analysis, 544
 - random process, 544
 - random variables, 544
 - regionalization, 544
 - spatio-temporal, 544
 - standard statistical techniques, 544
 - stationarity, 545
- Statistical fractals, 430
- Statistical hypothesis testing. *See* Hypothesis test
- Statistical inferential testing
 - definition, 1439
 - normal distribution, 1440
 - normality tests, 1441
 - plausible models, 1443
 - principles, 1443
 - quantitative footing, 1439
 - rock formations and sediments, 1439
 - F* test, 1440, 1441
 - z* tests, 1440
 - uncertainties, 1439
 - P* value, 1442, 1443
- Statistical maps, 481
- Statistical methods, 603
- Statistical moments, 1312, 1315, 1346
- Statistical parametric mapping (SPM), 1001
- Statistical physics, 1461
 - of granular materials, 1470
- Statistical process control (SPC), 1452
 - in grade control in mining, 1454
- Statistical properties, 1297
- Statistical quality control (SQC)
 - acceptance sampling, 1451
 - definition, 1451
 - SPC, 1452
- Statistical rock physics (SRP)
 - computational phase transitions, 1467–1470
 - maximum entropy method, 1459–1463
 - self-consistency and effective media, 1463–1464
 - stochastic geometry, 1456–1459
 - thermodynamic algorithms, 1464–1467
- Statistical seismology
 - definition, 1472
 - forecasting and testing, 1480–1484
 - history, 1472
 - observation technologies, 1485
 - probability gain framework, 1484
 - simple individual empirical laws, 1472–1475
 - statistical hypothesis testing, 1484
 - time-dependent forecasting models, 1475–1480
- Statistical self-similarity, 431
- Statistical significance, 207, 1439, 1442, 1443
- statistical technique, 1579
- Statistical testing, 1624
- Statistical time-series analysis, 50
- Statistics, 436, 446
 - for spatial data, 214
- Steeper, 1637
- Steepness, 322
- Step function, 1290
- Stepwise multidimensional scaling, 106
- Stereographic projections, 119
 - definition, 1486
 - graph types, 1486
 - remote sensing and geomatics, 1488
 - Schmitt net, 1487, 1488
 - two-dimensional, 1488
 - Wulff net, 1487
- Stereological parameters, 1493
- Stereology, 1490, 1509
 - basic principles and methods, 1493–1496
 - in computer era, 1499–1500
 - definition, 1491
 - design based stereology, 1499
 - model based stereology, 1499
 - sampling strategy and testing objects, 1497–1498
 - scope and limitations, 1491–1492
- Stimuli, 1267
- Stirling's formula, 1459
- Stochastic, 323
 - approach, 862
 - declustering, 1480–1482
 - logic, in geostatistics, 76
 - methods, 64
 - models, 1323
 - Poisson process, 1059
 - seismic inversion methods, 540
 - sequential simulation, 540
 - simulation, 1279
 - uncertainty, 1583
- Stochastic geometry
 - ACF, 1457–1458
 - applications, 1501
 - definition, 1501
 - 3D pore space, 2D sections, 1459
 - models of, 1502–1505
 - random functions, random fields, Kozeny-Carman equation, 1456–1459
 - random geometry, of tortuous flow paths, 1458–1459
- Stochastic partial differential equations (SPDE), 710, 711
- Stochastic process, 50, 544, 845, 1021, 1189
 - See also* Random function
- Stochastic Simulation with Patterns (SIMPAT), 963
- Stoichiometry, 144
- Stokes law, 1142, 1261
- Storage assessment units (SAUs), 1518, 1519
- Storage devices, 248
- STOTD and JADE algorithms, 643
- Stoyan, Dietrich, 1510
- Strange attractor, 434
- Strangeness-based outlier detection algorithm (StrOUD), 1389
- Stratified sampling, 1239
- Stratified/zonal anisotropy, 1379
- Stratigraphic sequence, 1511
 - BIF, 1513
 - coal, 1512
 - definition, 1511
 - petroleum, 1512

- Stratigraphy, 39, 1095, 1219
 - sequence, 1513
 - Streaming, 1466
 - Stream Power Incision Model (SPIM), 321
 - Stream power index, 933
 - Stress distributions, 329
 - Stress field, 524
 - Stress release model, 1478, 1480, 1481
 - Stress tensor, 955
 - Stretched exponential function, 1458
 - Stretching exponent, 1458
 - Strictly collocated, 548
 - Strictly positive definite (SPD), 1175, 1176, 1347
 - Strict stationarity, 50, 545, 1411
 - Strike-slip events, 330
 - Strong stationarity, 50
 - Structural analysis, 1376
 - Structural equation modeling (SEM), 982
 - Structural field observations, 1178
 - Structural geology, 1486, 1487
 - Structured Query Language (SQL), 498
 - Structure from motion (SfM), 284, 285, 877, 878
 - Structure function, 1216, 1217
 - Structuring element (SE), 322, 906–909, 924, 1514, 1537
 - characteristic information, 1514
 - decomposition, 1514–1515
 - directional, 1515
 - property of iteration, 1515
 - shape, 1514
 - size, 1515
 - Student's T distribution, 631
 - Subcomposition, 5, 11, 14
 - coherence, 135
 - Subduction, 330
 - Sub-geological regions, 1530
 - Subsequences, 1054
 - Subspace, 1348
 - Substitution random fields, 566
 - Substructures, 1054
 - Subsurface, 678, 679, 681
 - heterogeneity, 1316
 - rock properties models, 539
 - stresses, 524
 - Subsurface anomaly parameters
 - Fourier transform, 1298, 1299
 - Hilbert transform, 1302
 - Subsurface temperature modeling
 - geostatistics, 576–577
 - machine learning, 577
 - Successive residuals, 562
 - Sumatra-Andaman earthquake and tsunami of December 26, 1994, 330
 - Summarise, 1055
 - Sum-to-one constrained least squares (SCLS), 178
 - Sun-induced chlorophyll fluorescence (SIF), 405–406
 - Supercomputers, 495
 - Superconducting transition, 1598
 - Supervised, 929
 - learning, 36, 774, 782, 951, 983, 1064, 1182, 1267
 - Support, 556
 - Support vector machine (SVM), 147, 240, 258, 265, 465, 782, 989, 1056, 1061, 1062, 1273
 - classifier, 644
 - definition, 1520
 - kernel functions, 1522
 - linearly separable data, 1520, 1521
 - method, 1065
 - multiclass strategies, 1522, 1523
 - non-linearly separable data, 1521
 - quadratic optimization, parametrization and computational cost, 1524, 1525
 - regression, 1523, 1524
 - Surface, 928
 - impedance, 675
 - modelling, 1180
 - processes, 1504
 - roughness, 930
 - temperature estimation, 577–578
 - texture, 929, 930
 - Surface Reconstruction from Point Cloud, 878
 - Surrogate models, 400
 - Survey measurements, 519
 - Survival function, 1475
 - Susceptibility, 1598
 - Suspended sediment load, 1139
 - Sustainability, 298
 - SVR machine learning parameters, 1023
 - Swarm-intelligence, 1022
 - S waves, 329
 - Symmetric and positive semi-definite, 1607
 - Symmetric matrix, 722
 - Symmetry, 437, 1348, 1380
 - Synapses, 1265
 - Synclines, 372
 - Synthetic aperture radar (SAR), 338, 579, 1079, 1080, 1082, 1208
 - images, 696
 - Systematic(s), 38
 - sampling, 1239
 - uncertainty, 1583
 - System catalog, 231
- T**
- Takagi-Sugeno-type, 450
 - Tangent constraints, 1179
 - Tapering, 291
 - Target prediction, 382
 - Target variables, 381, 384, 385
 - Taxonomic descriptions, 38
 - Taxonomy, 1014
 - t-distributed Stochastic Neighbour Embedding (t-SNE), 144, 145
 - algorithm, 1528
 - box plots, 1533
 - chemical distribution, 1530
 - cluster delineation application, 1529, 1530
 - 2D coordinates, 1529
 - 2D plot, 1532
 - definition, 1527
 - exploration datasets, 1529
 - geological domains, 1530
 - high dimension data, 1528, 1531
 - implementation, 1527, 1528
 - low dimensional space, 1530
 - multidimensional data, 1529
 - optimization methods, 1528
 - results, 1534
 - stochastic variation, 1530
 - subpopulations, 1530
 - two-dimensional visualization, 1530
 - visualization, 1531
 - TEC data, 1297
 - Tectonic plate motion

- finite rotations, 352
- infinitesimal rotations, 352–353
- kinematic modeling of, 352–353
- Tectonic plates, geometric modeling of, 351
- Tectonics, 430
- Tectonic sketch maps, 645, 646
- Television Infrared Observation Satellite, 1206
- Temperature, 1056, 1459, 1461, 1466, 1467, 1470
 - distribution, 576
 - seasonality, 259–261
- Template matching, 1064
- Temporal covariance function, 1380
- Temporal information, 475
- Temporal point processes, 1479, 1480
- TensorFlow2, 498
- Tensor Processing Units (TPU), 498
- Terminal node, 257
- Terra ASTER sensors, 579
- Terrain, 1054
- Terrestrial laser scanning (TLS), 736, 737
- Tesseract, 38
- T*-test, 631
- Z*-test, 632
- Text mining
 - definition, 1535
 - for geoscience literature, 1535–1536
 - and mathematical geosciences, 1535
- Texture(s), 589
 - analysis, 216, 1308
 - goniometer, 216
- Thematic map, 481
- Theory-based data-driven approach. *See* Bayesian maximum entropy (BME)
- Theory of breakage., 770
- Theory of fuzzy sets to cluster analysis, 447
- Thermal, hydrous, mechanical, and chemical (THMC), 972
- Thermal conductivity, 1085
- Thermodynamic algorithms
 - LBM, flow in porous media, 1465–1466
 - random walk, 1464–1465
 - SA, 1466–1467
- Thermodynamics, 1598
- Thermo-hydromechanical coupling, 211
- Thermomechanical coupling, 211
- Thermotropic atmosphere model, 584
- Thesauri*, 1014
- Thickening, mathematical morphology, 1537, 1538
- Thiele, Thorvald, 660
- Thiergärtner, Hannes
 - early life, 1539
 - extraordinary professor, 1540
 - personal life, 1540
- Thinning, 924, 1459, 1538
 - method, 1479
- Thin plate spline (TPS), 1176, 1285, 1286
- Thin-sections, 1459
- Thompson-Howarth analysis, 1134
- Thompson-Howarth long method plot, 1128
- Thompson's deformation-grid approach, 1284
- Three-dimensional geologic modeling (3D geomodeling), 1540
 - data-driven and geomodel-driven, 1542
 - data sources, 1541
 - definition, 1540
 - explicit and implicit methods, 1543
 - geological structures and properties, 1541
 - geostatistics-based 3D simulation and modeling, 1543–1544
- Three-dimensional structure, 1490
- Three great evolutionary faunas, 1125
- Three-parameter lognormal distribution, 772
- Threshold, 319, 434
- Threshold autoregressive (TAR) model, 1556
- Thresholding of the images, 1458
- Throats, 1459, 1460, 1468
- Tidal wave, 330
- Tikhonov regularization, 686
- Time-dependent forecasting models
 - basic formulation, 1475
 - conditional intensity, 1475–1477
 - ETAS model, 1479
 - functional and spectrum analysis, 1475
 - generic simulation method, 1479, 1480
 - point process, 1475
 - probability forecasts and performance evaluation, 1480
 - Reasenbergs-Jones model, 1478
 - renewal/recurrence models, 1477–1478
 - self-and mutually exciting models, 1478, 1479
 - stress release model, 1478
- Time distance decay weighting strategy, 1392
- Time-frequency analysis, 1631–1634
- Time-frequency localization, 1303
- Time-frequency plane, 1304
- Time-invariant, 290
- Time-limited DPSSs, 303
- Time localization, 1303
- Time series, 47–49, 1004, 1347
 - data, 238
- Time series analysis (TSA), 1546, 1574
 - autoregressive moving average model, 1553
 - data, 1551
 - definition, 1551
 - descriptive tools, 1552–1553
 - ETS model, 1548
 - forecasting, 1554–1555
 - functional, 1558–1559
 - with long memory, 1556
 - multivariate, 1557–1558
 - objectives, 1551
 - SARIMA process, 1548–1551
 - smoothing methods, 1548
 - software, 1559
 - state-space model, 1546
 - stationarity, 1546
- Time-variant scaling, 52
- Time-varying filters, 290, 292
- Time-varying frequency components, 1303
- Time-varying gain, 49
- Tobler, Waldo, 1559–1560
- Tobler's optimisation algorithm, 105
- Toeplitz matrix, 292, 869
- Tomographic image reconstruction, 55
- Tool-oriented approach, 298
- Top-level ontology, 1015
- Topographic change, 580
- Topographic index, 933
- Topographic map, 481
- Topography, 928
 - control model, 397–398
- Topology, 912, 1181, 1268
 - connectivity, 1561
 - constraints, 1361
 - data model, 809
 - equivalence, 1469

- Topology (*cont.*)
 euler's formula, 1561
 fracture networks, 1563, 1565
 geometry, 1561
 homeomorphic, 1561
 relations, 478
 structural models, 1562, 1563
 Torch, 498
 Tortuosity, 27, 1089, 1091, 1457, 1458, 1468
 Total alkali-silica (TAS) diagram, 1566–1567
 Total Column Observing Network (TCCON), 1384
 Total electron content (TEC), 1297
 ionospheric, 1314, 1315
 multifractal behavior, 1314
 spatio temporal behavior, 1314
 Total or absolute porosity, 1091
 Total porosity, 388
 Total Sun-induced chlorophyll fluorescence (SIF_{TOT}), 405, 406
 Total variance, 140
 TOUGH2, 580
 Trace-by-trace stochastic seismic inversion, 540
 Trace gases, 1382
 remote sensing-based measurement, 1382–1383
 Trace of a matrix, 839
 Trace operator, 843
 Tracking, 36
 Traditional bootstrap, 96
 applications of, 97
 resampling, 97–98
 Traditional geological mapping, 294
 Traditional R/S analysis, 1215
 Training, 953
 data, 36, 265
 dataset, 725
 set, 866
 Training images, 1352
 construction, 966
 data-based, 965
 features, 965
 requirement, 964
 stationarity, 964
 validation, 967
 variograms, 965
 Trajectory matrix, 1328, 1330
 Transfer, 1637
 characteristics, 290
 function, 292, 293, 952, 1645
 function bias, 1416
 learning, 36, 265
 Transformation, 84
 Transformed time sequence, 1476
 Transform methods, 1027
 Transient, 320
 Transition function, 1266
 Transitive, 545
 Translation-invariance property, 1305
 Translation parameter, 1304
 Translation property, 715
 Transparency, Responsibility, User focus, Sustainability
 and Technology (TRUST Principles), 1018
 Transportation, 483
 of water, 1139
 Transport laws, 319
 Transpose of a matrix, 839
 Trapezoidal membership function, 855
 Traveling salesman problem (TSP), 1321
 Traveltime channel networks, 926
 Traverses, 1004
 Tree-based modeling, 982
 Tree-based predictive models, 1183
 Tree complexity, 1183
 Tree-ring data, 1216
 Trends, 1054
 Trend surface analysis, 117, 481
 computer-based methods, 1567
 definition, 1567
 examples, 1568
 inverse distance weighting methods, 1567
 method, 1567
 mineral composition, 1567
 Mount Albert example, 1569, 1571
 pyroxene crystals, 1574
 residuals, 1571, 1573, 1574
 Trial and error simulations, 37
 Triangular membership function, 855
 Triangular space, 1282
 Triangulated irregular network (TIN), 289, 738
 model, 477, 480
 Trigonometric functions, 29
 Trilobite, 1222
 Tropospheric Monitoring Instrument (TROPOMI), 1383
 True 3D, 930
 Truncated Gaussian, 564
 simulation, 562, 563, 1325
 Truncated lognormal distribution, 772
 Truncating, 291
 Trust region algorithm, 1438
 t-test, 205
 Tukey, J.W.
 geology, 1575
 “Geostatistics” symposium, 1575
 honours, 1575
 mathematical geoscience, 1575
 paper presentation, 1576
 Tukey biweight function, 690
 Turbidity currents, 1262, 1577
 Turbidity Current with Roof (TCR) model, 1577
 Turbulence
 definition, 1576
 turbidity currents, 1577
 Turbulent energy dissipation, 785
 Turcotte, D. L., 1578
 Turing's Bayes factors, 760
 Turing test, 34
 Turning, A., 32
 Turning Bands algorithm, 565
 Two-point or second-order statistics, 605
 Two-point statistics, 559
 Two-sample t test, 1432
 Two-sided equal tail permutation p-value, 1432
- U**
 UAV-mounted sensors, 627
 UML class diagram, 1014
 Unbalanced design, 1579, 1580
 Unbiased estimation, 1497
 Unbiasedness condition, 1608
 Uncertain rules, 1603

- Uncertainty(ies), 323, 381, 681, 685, 1196, 1543, 1583, 1586
 - analysis, 392
 - in geochemical anomaly classification models, 1584
 - MCSIM, 1584–1586
 - quantification, 856
 - spatial uncertainty, 1585–1587
 - stochastic, 1583
 - systematic, 1583
 - Unconstrained Least Squares (UCLS), 178
 - Underdetermined system, 818
 - Underestimation, 776
 - Underground mine planning, 861
 - Undirected road network, 596
 - UNEXMIN, 39
 - Unified metadata, 232
 - Uniform conditioning, 558
 - Uniform convergence, 715
 - Uniform Manifold Approximation and Projection (UMAP), 144
 - Uniform Resource Identifier (URI), 1019
 - Uniform resource locators (URLs), 90
 - Unimate industrial robot, 37, 38
 - Union, 924
 - Unit circle, 830–831
 - Unit delay operator, 1299
 - United, 1055
 - Unit eigenvector, 385
 - Unit impulse function, 716
 - Unit regional production value
 - characteristics, 1589
 - definition, 1589
 - global strategy, 1589
 - log dollar value, 1590
 - mineral resources, 1589
 - nonrenewable resources, 1589
 - standardized values, 1589
 - substantial production, 1589
 - Unit regional value, 1589, 1591, 1592
 - definition, 1591
 - geological information, incorporation of, 1591–1592
 - original regional applications, 1592
 - Unit vector field(s), 532
 - analysis, 1487
 - Univariate anomalies, 359
 - Univariate data, 366–367, 1058–1059
 - Univariate distributions, 1593, 1595
 - Univariate interpolation, 660–662
 - Univariate random function, 547
 - Univariate random variable, 223, 1190
 - Univariate statistics, 1593–1595
 - Univariate time series method, 1557
 - Universal gravitational constant, 1299
 - Universality, 433, 1598, 1599
 - class, 1597
 - definition, 1597
 - percolation, fractals and multifractals, 1598–1599
 - in phase transitions, 1598
 - self-organized criticality and earthquakes, 1599
 - Universal kriging (UK), 550, 707, 1347, 1365, 1567
 - Universal multifractals, 441–443
 - Universal Serial Bus (USB), 226
 - Universal transverse mercator (UTM) projection, 476
 - Unmanned aerial vehicles (UAV), 284, 736, 1207
 - un-sampled spatial locations, 1607
 - Unstructured, 1535
 - Unsupervised, 929
 - approach, 339
 - data analysis approach, 1203
 - dimensionality reduction, 308
 - learning, 36, 774, 782, 983, 1267
 - machine learning, 1527
 - Unweighted moving average, 935
 - Updated probability, 62
 - Upper approximation, 1600, 1602
 - Upper limit of detection (ULD), 144
 - Upper quartile, 436
 - Uppsala University, 1218
 - Upward continuation, 1094
 - Uranium mineral system, 453
 - Urban, 1054
 - U.S. Environmental Protection Agency, 693
 - U.S. Geological Survey (USGS), 391, 810
 - USGS 2000 World Petroleum Assessment, 1518
 - USGS map of principal faults, 332
 - USGS Oil and Gas Resource Projections, 1518
 - USGS probabilistic assessment of CO₂ storage capacity, 1518–1519
 - US National Aeronautics and Space Administration, 40
 - UX-I*, 39, 40
- V**
- Validation, 928, 1168
 - dataset, 36
 - set, 866
 - Value-by-area map, 103
 - Vanishing moments, 622, 1314
 - Varanasi data set, 139
 - Variability, 1055
 - Variable importance, 1183
 - Variance, 203, 714, 1054, 1186, 1278, 1313, 1346
 - a priori variance, 1605
 - covariance, 1605, 1606
 - definition, 1605
 - dispersion variance, 1608, 1609
 - estimation variance, 1607, 1608
 - first and second-order moments, 1605
 - indicator random field, 1606
 - inflation factor, 958
 - properties, 1606, 1607
 - ratio, 310
 - sample variance, 1607
 - spatial statistics, 1605
 - stationarity hypotheses, 1606
 - variogram, 1605, 1606
 - Variance-covariance matrix, 486, 871, 1607
 - Variational autoencoders (VAEs), 270
 - Variation matrix, 140
 - Varimax orthogonal axis-rotation, 1124
 - Variogram, 146, 251, 621–622, 963, 1186, 1346, 1349, 1375, 1376, 1543, 1581, 1605, 1606
 - definition, 1609
 - function, 545
 - interpolation methods, 1610
 - main principles, 1610–1611
 - model, 4, 540, 541
 - samples and datasets, 1609
 - types of, 1611–1613
 - Varves, 1577
 - Vector autoregressive moving average (VARMA)
 - model, 1557
 - Vector data analysis, 478

Vector data visualization, 481
 Vectorial data, 1352
 Vectorial quantities, 276
 Vector space, 613
 structure, 553
 Vector wave-number, 1636
 Venn diagram, with types of spaces, 613
 Vertebrate brain, 1267
 Vertical outliers, 1226
 Vertical sections, 1497
 Vertical stress, 525
 Very fast simulated reannealing
 acceptance probability, 1616
 annealing schedules, 1617
 definition, 1614
 generation of proposal states, 1616
 reannealing, 1616
 termination criteria, 1618
 Very high-resolution (VHR) remote sensing images, 448
 VGGNet, 265
 Virtual and augmented reality, 166
 Virtual environment, 876
 Virtual geographic environments (VGEs), 1622
 Virtual globe
 applications, 1620–1622
 browser-based visualization tool, 1619
 definition, 1619
 future scope of, 1622
 history of, 1619
 Virtual governance, 536
 Virtual machine technique, 124
 Visible and Near Infrared (VNIR), 627
 Vistelius, Andrey Borisovich, 1219, 1624
 early childhood, 1623
 education, 1624
 scientific achievements, 1624
 Visual diagnostics, 1226
 Visualization, 1383–1384, 1545
 of multidimensional data, 1268
 tool, 1527
 Vitritine reflection, 1470
 Vocabularies, 371
 Void ratio, 574, 1089
 Void space, 573
 Volcanic-hosted massive sulphide (VHMS), 1063
 Volcanic massive sulfide (VMS) Cu geochemical anomaly map, 171
 Volcanic rocks, 1566
 Volcanogenic massive sulphide deposits, 1098
 Volterra-Lotka equations, 1026
 Volume fraction, 1494
 Volume-variance effect, 1
 Volunteered geographic information (VGI), 483, 587, 1359
 Voronoi cell, 323
 Voronoi tessellation (VT), 696, 1507
 Voxels, 1459, 1464–1467

W

Waiting time, 1475–1477, 1479
 Wald test, 723
 Wald-Wolfowitz runs test, 1434
 Walther's Law, 568, 569
 Ward method, 129
 Warning systems, 329
 Washburn equation, 1085

Water flow and sediment scale processes, 1139
 Watershed lines, 916
 Watson, Geoffrey S., 1625–1626
 Wave conversion, 51
 Wave equation, 1044
 Wavefront spreading, 51
 Wavelength, 1636, 1640
 dependent magnification, 628
 Wavelet(s)
 applications in geoscience, 1628–1635
 definition, 1626
 filtering, 1633–1635
 fractal analysis, 1629–1632
 function, 1307
 historic analysis, 1626
 multiresolution analysis, 1629–1631
 neural networks, 1635
 time-frequency analysis, 1631–1634
 Wavelet analysis, 1297, 1303–1308
 basic theory, 1304, 1305
 CWT, 1305
 DWT, 1305, 1307
 EMD, 1315–1317
 vs. Fourier transform, 1304
 remote sensing, multiresolution, 1307, 1308
 scales, 1304
 signal characterization, 1303, 1304
 temporal behavior, 1303
 time-localization, 1303
 WTMM, 1314–1315
 Wavelet coherence analysis (WCA), 1632
 Wavelet transform modulus lines (WTMM), 622, 623
 Wavelet transform modulus maxima (WTMM), 1311, 1314–1315
 Wave-number, 432
 calculation, 1637
 definition, 1636
 2D wave-number, 1636–1637
 Wave-number filtering, 1637
 Butterworth filter, 1637
 Chebyshev filters, 1637–1638
 Gaussian filter, 1638–1639
 2-Way travel-time, 52
 Weak stationarity, 50
 Weather forecasting using numerical weather prediction models, 328
 Weather Research and Forecasting model, 1385
 WebGIS, 484
 Web Ontology Language (OWL), 498
 Web pages, 858
 Web world, 233
 Weibull distribution, 591
 Weibull renewal model, 1477
 Weight, 1265
 Weighted average, 288, 666, 667
 Weighted least squares (WLS), 486
 problem, 1435
 Weighted moving average (WMA) interpolation techniques, 1610
 Weighted moving average methods, 935–937
 Weight function, 292
 Weights of evidence, 654, 762–764, 1059, 1576
 Weights-of-evidence modeling, 514, 1103
 Weiss crystal face index, 807
 Welch's method, 306
 Well-logging data, 577
 Well-logs data, 623–624, 1269, 1297
 Wendland's functions, 1176

Wettability, 955
Wetting film, 1465
Whalesback copper deposit, 373
Whiskbroom scanner, 625
White noise, 1093, 1214, 1311, 1316
 process, 50, 1553
White top-hat transform, 923
Whitten, Eric Harold Timothy
 career, 1642
 education, 1642
Wiener process, 1412
Wiggin's entropy function, 869
Wilcoxon test, 631, 632
William Christian Krumbein Medal, 96, 1012, 1235
Window, 291
 length, 1328
Windowed Fourier transform, 1303
WinGSLIB, 1351
Winning neuron, 1268
Wireless communication links, 38
Window-based Fourier transform, 1626
Woodbury matrix, 710
WordNet, 1014
Workflow(s), 1210–1211
 management, 37
Working Group on California Earthquake Probabilities (WGCEP), 1477
WorldClim, 499
World Meteorological Society, 329
World-wide geographic maps, 1619
World-wide network protocols, 370
Writing code, 1212
Wulff net, 1487

X

Xie-Beni (XB) index, 448
XML, 499
XML markup language, 38
X-ray computed tomography (CT), 55
X-ray diffraction analysis, 818
xy-plot, 1257

Y

Young's equation, 1467

Z

Zadeh, Lotfi A., 1643, 1644
zCompositions, 139
Zero-pole diagram, 28
Zero-pole plot, 28
Zicxac Inline Pin (ZIP), 248
Zipf's law distribution, 810
Zoeppritz equations, 69
Zonal anisotropy, 548, 1348
Z-plane, 28
Z-score, 999
Z-transform, 290, 291, 716, 1298
 definition, 1644
 DT signal, 1645
 Fourier transform relation, 1300
 in geosciences, 1645
 of impulse signal, 1645
 non-causal signals, 1299
 polynomials, 30
 unit delay operator, 1299